NATIONAL DEFENCE UNIVERSITY “CAROL I”

REGIONAL DEPARTMENT OF DEFENSE RESOURCES
MANAGEMENT STUDIES

INFORMATION SECURITY MANAGEMENT – EVOLUTIONS

Workshop unfolded during the postgraduate course in Information Security Management

14- 15.06.2010, Brasov

Coordinator:
LTC Prof. eng. Daniel Sora, PhD

National Defense University „Carol I” Publishing House
BUCHAREST 2011

Scientific board:

LTC Professor eng. Daniel Sora, PhD
LTC Senior Lecturer Cezar Vasilescu, PhD
Junior lecturer Aura Codreanu, PhD

ISBN: 978-973-663-952-4

CONTENT

INFORMATION AND PHYSICAL SECURITY

LTC Marius Daniel CĂLINOIU

INTRODUCTION

The practice of information security has become much more complicated and the need for qualified information security professionals has become critical.

Physical security involves the provision of a safe environment for information processing activities with a focus on preventing unauthorized physical access to computing equipment. Three categories include: **(1) threats and facility requirements**, **(2) personnel physical access control**, and **(3) computer physical security**.

As technologies evolve, the protection of resources becomes increasingly more complex. Nevertheless, information security is predominantly an organizational issue, and as such, establishing and enforcing policies and standards is critical to the successful administration of the Information Security Program.

Physical security is often a discounted discipline, yet attention to safeguarding the physical environment can yield a satisfactory level of protection. This document offers a comprehensive look at implementing a physical security program, which begins with a risk assessment so that the appropriate most cost-effective controls are implemented. Additionally, the paper illustrates the multiple biometric technologies and defines each in terms of rejection and acceptance rates. Ultimately, maintains that a good physical security program is an organization’s first line of defense.

A **company's physical and logical information networks and user interfaces have been completely separate** for years.

Building access, or physical security systems are typically put in place by either the owner of the building or, in the case of larger businesses, by the corporation's security department. Network and data security, or logical security, systems are the domain of the IT department. Each developed separately within the organization. Corporate security departments developed to protect physical assets through locks, surveillance and alarm systems and are typically staffed by people with backgrounds in law enforcement, not technology. In contrast, protecting a company's information and knowledge assets has been one of the main tasks of IT since day one. This role has evolved into protecting both company and employee data since the dawn of the Internet age.

All corporate assets - from office equipment to employee belongings - need to be protected and hackers, industrial saboteurs and terrorists must be prevented from wreaking havoc with networks, applications and databases. However, because physical and logical security systems have traditionally been handled separately with little or no cross over, few companies realize how much a converged system could help.

In many ways, building access security systems have always acted as the first line of defense against unauthorized access to any company assets, physical or logical. If an intruder could not gain entry to a company's offices, that person could therefore not gain access to corporate applications and sensitive data. However, with the advances in technology, this is no longer the case as telecommuting and remote access become more prevalent every day. A company's IT assets and critical data can no longer be protected by physical security systems alone.

There have been other, more conventional, attempts made at solving the issue of unauthorized access to company information, but they all stop short of true integration. Some of these have included:

- Multifunction cards using either proximity capabilities or a traditional magnetic strip combined with a digital certificate or other credentials to identify users when they enter buildings or access their computer. However, there is no way to correlate access policy across systems or revoke all the various credentials contained on the card simultaneously.
- Identity management solutions can enable provisioning for new users, streamlining the creation of directory accounts and required user applications, as well as physical access privileges and web-application access control. However, they are costly and time-consuming to implement and are not a realistic solution for small to mid-size businesses.
- Consolidation is closest to an integrated physical and logical approach, as it gathers logs from application, network and physical access systems and generates consolidated reports by user. The problem with this approach is that it is time consuming to set up and still only lets administrators see what has already happened; it does not control access or prevent a transgression from happening in real time.

Physical and logical security concerns continue to mount, bringing the problems with the above solutions and issues such as inadequate security policy and lax enforcement to the forefront. Today, more and more organizations are realizing that a combination of their physical and logical security systems will help strengthen their security and better protect their company, employee and customer data.

I. Threats and Facility Requirements

The injury to the national interest or to private/ non-national interests increases with the sensitivity of the disclosed information. Injury may include damage to the defense and maintenance of the economic, social or political stability of a country, compromise of other governments' interests, breach of privacy, liability or financial loss, loss of confidence in the government, or decrease of government efficiency.

Unauthorized disclosure of protected or classified information can occur:

a. accidentally through loss or negligence by employees who were granted access to the information;
b. intentionally by individuals who have authorized (i.e., have been properly security screened and have a need to know) access to the information; and
c. intentionally by individuals who gain unauthorized access to information by whatever means, e.g., targeting of protected and classified information by criminal, terrorist or foreign intelligence elements.

Unauthorized disclosure of Secret or Protected C information will create more injury than unauthorized disclosure of Protected A or B information. In addition, some classified or protected information may be more attractive than other information in the same security classification and may, therefore, require safeguarding above the baseline delineated for this level of information.

The approach to physical security complements other aspects of the Security Policy and is based on the theory that the external and internal environments of facilities can be designed and managed to create conditions that, together with specific physical security safeguards, will protect against unauthorized access, detect attempted or actual unauthorized access and activate an effective response.

I.1 Physical Security

I.1.1 The human factor

Recent FBI statistics indicate that 72% of all thefts, fraud, sabotage, and accidents are caused by a company's own employees. Another 15 to 20% comes from contractors and consultants who are given access to buildings, systems, and information. Only about 5 to 8% is done by external people, yet the press and management focus mostly on them. The typical computer criminal is a non-technical authorized user of the system who has been around long enough to locate the control deficiencies.

When implementing control devices, make certain that the controls meet the organization's needs. Include a review of internal access, and be certain that employees meet the standards of due care imposed on external sources. "Intruders" can include anybody who is not authorized to enter a building, system, or data.

The first defense against intruders is to keep them out of the building or computer room. However, because of cost-cutting measures in the past two decades, very few computer facilities are guarded anymore. With computers everywhere, determining where to install locks is a significant problem.

To gain access to any business environment, everybody should have to pass an authentication and/ or authorization test. The three ways of authenticating users involve something:

- That the user knows (a password).
- That the user has (a badge, key, card, or token).
- Of their physiognomy (fingerprint, retinal image, voice).

I.1.2 Locks

In addition to securing the campus, it may be necessary to secure the computers, networks, disk drives, and electronic media. One method of securing a workstation is with an anchor pad, a metal pad with locking rods secured to the surface of the workstation. The mechanism is installed to the shell of the computer. These are available from many vendors.

Many organizations use cables and locks. Security cables are multistrand, aircraft-type steel cables affixed to the workstation with a permanently attached plate that anchors the security cable to the desk or other fixture.

Disk locks are another way to secure the workstation. These small devices are quickly inserted into the diskette slot and lock out any other diskette from the unit. They can prevent unauthorized booting from diskettes and infection from viruses.

Cryptographic locks also prevent unauthorized access by rendering information unreadable to unauthorized personnel. Encryption software does not impact day-to-day operations while ensuring the confidentiality of sensitive business information. Cryptographic locks are cost-effective and easily available.

I.1.3 Level of security

Before any controls can be implemented into the workplace, it is necessary to assess the current level of security. This can be accomplished in a number of ways. The easiest one is a “walk-about.” After hours, walk through the facility and check for five key controls:

a. Office doors are locked.

b. Desks and cabinets are locked.

c. Workstations are secured.

d. Diskettes are secured.

e. Company information is secured.

Checking for these five key control elements will give you a basic understanding of the level of controls already in place and a benchmark for measuring improvements once a security control system is implemented. Typically, this review will nearly show a 90% control deficiency rate. A second review is recommended six to nine months after the new security controls are in place.

This chapter examines two key elements of basic computer security: **physical** security and **biometrics**.

Physical security protects your organization's physical computer facilities. It includes **access to the building**, **to the computer room(s), to the computers** (mainframe, mini, and micros), **to the magnetic media**, and **to other media**.

Biometrics devices record physical traits (i.e., fingerprint, palm print, facial features, etc.) or behavioral traits (signature, typing habits, etc.).

In the beginning of the computer age, it was easy to protect the systems; they were locked away in a lab and only a select few "wizards" were granted access. Today, computers are cheaper, smaller, and more accessible to almost everyone.

During the mid-twentieth century, the worldwide market for mainframe computer systems exploded. As the third-generation systems became available in the 1960s, companies began to understand their dependence on these systems. By the mid to late 1970s, the security industry began to catch up: with Halon fire suppression systems, card access, and RACF and ACF2. In the final quarter of the century, mainframe-centered computing was at its zenith.

By 1983, the affordable portable computer began to change the working landscape for information security professionals. An exodus from the mainframe to the desktop began. The controls that had been so hard won in the previous two decades were now considered the cause of much bureaucracy.

Physical security is now needed in desktops. For years, conventional thinking was that a computer is a computer is a computer is a computer.

Controls are even more important in the desktop or workstation environment than in the mainframe environment.

The computing environment is now moving from the desktop to the user. With the acceptance of telecommuting, the next challenge will be to apply physical security solutions to the user-centered computing environment.

With computers on every desk connected via networks to other local and remote systems, physical security needs must be reviewed and upgraded wherever necessary. Advances in computer and communications security are not enough; physical security remains a vitally important component of an overall information security plan.

I.2 Physical Security Assessment

In today's environment, analysis of the physical security of facilities and properties has become an even more critical aspect of an organization's information security and business continuity planning.

During an onsite assessment, consultants perform physical inspections of facilities and operations. They begin each physical security review by gaining an understanding of the resources being protected and the perceived threat environment. Through interviews and limited reviews of local policies and procedures covering physical security operations, they gains an understanding of the level of protection desired and needed in a given location.

Armed with this understanding, the consultant's team conducts the review of the facility. Key areas assessed include:

- Facility Security
    - Entry points
    - Data center
    - User and sensitive environments
    - Access control and monitoring devices
    - Guard personnel
    - Wiring closets
- Internal Company Personnel
    - Control and accountability
    - Use of equipment
    - Security procedure compliance
    - Awareness
    - Use of break areas and entry points
- External Visitor and Contractor Personnel
    - Control and accountability
    - Use of equipment
    - Security procedure compliance
    - Use of break areas and entry points
- Computer Systems and Equipment
    - Workstations
    - Servers
    - Backup media
    - PDAs
    - Modems and physical access points (visual ID only)

- Sensitive Information and Data
    - Control
    - Storage
    - Destruction

They can expand overt assessment process through the use of covert red-team assessment techniques. These efforts include tactics such as social engineering, pretext entry, security systems bypass, device/Trojan planting, long range surveillance and other methods. Covert assessment is a secondary add-on package.

Physical security reviews are performed and analyzed in the context of your organization's overall risk management strategy. The criticality of assets within the environment and the perceived threat environment directly affect the level of exposure that is classified as acceptable. By analyzing the combined factors of assets, threat, and exposure, physical security review provides much more than a list of actionable security recommendations. Experts prioritize exposures and make recommendations to align physical security with your overall risk management strategy. This holistic view enables you to protect the right assets with the right level of security.

II. Personnel Physical Access Control

In the past few years, the corporate world's image of the personnel function has undergone a significant change. An organization's employees are now considered a corporate resource and asset, requiring constant care and management. Changing legal conditions affecting personnel practices have underscored the need for clearly defined and well-publicized policies on a variety of issues.

The corporation and the employee have specific legal and ethical responsibilities to each other, both during and after the period of employment. Hiring and termination criteria, trade secrets, and non-competition clauses are all issues that can cause serious legal problems for a corporation and its employees.

II. 1 Information Security and Personnel Practices

This chapter addresses personnel issues as they relate to information systems security, particularly hiring and termination procedures. Methods to protect both the corporation and the employee from unnecessary legal problems are discussed, and problems regarding trade secrets and non-competition clauses are reviewed.

II.1.1 The professional environment

The information systems and information security professions are in a vibrant and exciting industry that has always operated under a unique set of conditions. The industry relies on the unquestioned need for absolute confidentiality, security, and personal ethics. An organization and its reputation can be destroyed if its information security procedures are perceived as being inadequate or unsatisfactory. Yet, misuse or outright theft of software and confidential information can be relatively easy to accomplish, is profitable, and is often difficult to detect. Innovations can be easily transferred when an employee leaves the corporation, and information systems personnel have always been particularly mobile, moving among competitors on a regular basis.

These factors are extremely important as they relate to the corporation and its personnel practices. A newly hired programmer or security analyst, whose ethical outlook is largely unknown to management, may quickly have access to extremely sensitive and confidential information and trade secrets. Unauthorized release of this information could destroy the corporation's reputation or damage it financially. An employee who has just accepted a position with a major competitor may have access to trade secrets that are the foundation of the corporation's success.

II.1.2 Hiring practices

Corporations must take special care during the interview to determine each candidate's level of personal and professional integrity. The sensitive nature and value of the equipment and data that employees will be handling require an in-depth screening process. At a minimum, this should include a series of comprehensive interviews that emphasize integrity as well as technical qualifications. References from former employers should be examined and verified.

The best way to verify information from an employment application is to conduct a thorough reference check with former supervisors, co-workers, teachers, and friends listed by the applicant on the application. Former employers are usually in the best position to rate the applicant accurately, providing a candid assessment of strengths and weaknesses, personal ethics, and past earnings, among other information.

Many employers have become increasingly cautious about releasing information or making objective statements that rate former personnel. Such employees have successfully sued corporations and supervisors for making derogatory statements to prospective employers. Many employers will furnish written information only about the applicant's dates of employment, positions held, and salaries earned, choosing to ignore more revealing questions. Often, an informal telephone check may reveal more information than would be obtained by a written request. If two large employers regularly hire each others' employees, it would be worthwhile for their personnel managers to develop a confidential personal relationship.

Use of a reference authorization and hold-harmless agreement can help raise the comfort level of the former employer and get more complete information from a job applicant's previous employer. In such an agreement, the applicant authorizes the disclosure of past employment information and releases both the prospective employer and the previous employer from all claims and liabilities arising from the release of such information. An employer who uses such an agreement should require every job applicant to sign one as a condition of applying for employment. A copy of the agreement is then included with the request for references sent to the previous employer.

When sending or responding to a reference request that includes a reference authorization waiver and hold-harmless agreement, it is important for employers to make sure that the form:

- Is signed by the job applicant.
- Releases the employer requesting the information as well as the previous employer from liability.
- Clearly specifies the type of information that may be divulged.

A responding employer should exercise extreme caution before releasing any written information about a former employee, even if the former employee has signed a reference authorization waiver. Only information specifications permitted by the waiver should be released. If there is any ambiguity, the former employer should refuse to release the requested information. The former employer is safest if only the date of hire, job title, and date of termination are released.

II. 2 Controlling Restricted-Access Areas

Departments have several choices available to control access to restricted-access areas:

- personal recognition,
- access badges,
- mechanical measures,
- electronic control of access, etc.

The appropriate choice will depend on the location of building, number of employees, threat and risk assessment etc.

II.2.1 Identification Cards

All employees **must be** issued an identification (ID) card, which as a minimum includes, the name of the department, the bearer's name and photo, a unique card number and an expiry date. A signature is recommended.

II.2.2 Access Badges

Access badges indicate authorized employees and visitors. Where personal recognition or escorts are not feasible, a temporary access badge must be issued to all visitors (including non authorized employees, contractors, service personnel) which clearly identifies them as a non employee.

II.2.3 Electronic Access Control

An electronic access control is a safeguard that will assist in controlling access to a facility. A threat and risk assessment will assist in determining the need and cost effectiveness of such a system. Sometimes when a department chooses to implement an electronic access control system, the requirements for an ID card and an access badge are combined in one electronic access control card.

II.2.4 Closed Circuit Video Equipment (CCVE)

Closed circuit video surveillance/ assessment equipment may assist a department in providing appropriate monitoring of access to their facility. A threat and risk assessment will assist in determining the need for CCVE.

II.2.5 Security Control Centre

A security control centre, whether proprietary or off site, is a focal point for monitoring the various systems such as an electronic access control system, an electronic intrusion detection system and closed circuit video equipment. This centre will typically include other personal or life safety equipment such as the fire alarm panel. A control centre of this nature would typically only be used in the larger facilities.

II.2.6 Sensitive discussion areas

A sensitive discussion area (SDA) is an area that is specially designed and managed to prevent the overhearing of Protected and Classified information at various levels of sound attenuation. Owing to the cost of building and operating an SDA, departments should carefully assess the need, the risk and cost-effectiveness of options.

II.2.7 Secure rooms

Secure rooms are rooms constructed according to technical standards for the storage of Protected and Classified information and assets.

Classified and Protected information stored in the appropriate type of secure room need not be further protected by storage in additional security containers, unless the application of the need-to-access principle is still a concern. A records office where protected and classified information is stored on open shelves must be constructed as a secure room.

III. Computer Physical Security

Today's portable computing environment can take on a variety of forms: from remote connectivity to the home office to remote computing on a standalone microcomputer with desktop capabilities and storage. Both of these portable computing methods have environment-specific threats as well as common threats that require specific protective measures. Remote connectivity can be as simple as standard dial-up access to a host mainframe or as sophisticated as remote node connectivity in which the remote user has all the functions of a workstation locally connected to the organization's local area network (LAN). Remote computing in a standalone mode also presents very specific security concerns, often not realized by most remote computing users.

III. 1 Portable computing threats

Portable computing is inherently risky. Just the fact that company data or remote access is being used outside the normal physical protections of the office introduces the risk of exposure, loss, theft, or data destruction more readily than if the data or access methods were always used in the office environment.

III.1.1 Data Disclosure

Such simple techniques as observing a user's remote access to the home office (referred to as shoulder surfing) can disclose a company's dial-up access phone number, user account, password, or log-on procedures; this can create a significant threat to any organization that allows remote dial-up access to its networks or systems from off-site. Even if this data or access method isn't disclosed through shoulder surfing, there is still the intermediate threat of data disclosure over the vast amount of remote-site to central-site communication lines or methods (e.g., the public phone network). Dial-up access is becoming more vulnerable to data disclosure because remote users can now use cellular communications to perform dial-up access from laptop computers.

Also emerging in the remote access arena is a growing number of private metropolitan wireless networks, which present a similar, if not greater, threat of data disclosure. Most private wireless networks don't use any method of encryption during the free-space transmission of a user's remote access to the host computer or transmission of company data. Wireless networks can range in size from a single office space serving a few users to multiple clusters of wireless user groups with wireless transmissions linking them to different buildings. The concern in a wireless data communication link is the threat of unauthorized data interception, especially if the wireless connection is the user's sole method of communication to the organization's computing resources.

All of these remote connectivity methods introduce the threat of data exposure. An even greater concern is the threat of exposing a company's host access controls (i.e., a user's log-on account and static password), which when compromised may go undetected as the unauthorized user accesses a system under a valid user account and password.

III.1.2 Data Loss and Destruction

Security controls must also provide protection against the loss and destruction of data. Such loss can result from user error (e.g., laptop computers may be forgotten in a cab or restaurant) or other cause (e.g., lost baggage). This type of data loss can be devastating, given today's heavy reliance on the portable computer and the large amount of data a portable computer can contain. For this reason alone some security practitioners would prohibit use of portable computers, though increased popularity of portable computing makes this a losing proposition in most organizations.

Other forms of data loss include outright theft of disks, copying of hard disk data, or loss of the entire unit. In today's competitive business world, it is not uncommon to hear of rival businesses or governments using intelligence-gathering techniques to gain an edge over their rivals. More surreptitious methods of theft can take the form of copying a user's diskette from a computer left in a hotel room or at a conference booth during a break. This method is less likely to be noticed, so the data owner or company would probably not take any measures to recover from the theft.

III.1.3 Threats to Data Integrity

Data integrity in a portable computing environment can be affected by direct or indirect threats, such as virus attacks. Direct attacks can occur from an unauthorized user changing data while outside the main facility on a portable user's system or disk. Data corruption or destruction due to a virus is far more likely in a portable environment because the user is operating outside the physical protection of the office. Any security-conscious organization should already have some form of virus control for on-site computing; however, less control is usually exercised on user-owned computers and laptops. While at a vendor site, the mobile user may use his or her data disk on a customer's computer, which exposes it to the level of virus control implemented by this customer's security measures and which may not be consistent with the user's company's policy.

IV. Natural disasters and controls

The focus of physical security has often been on human-made disasters, such as sabotage, hacking, and human error. Don't forget that the same kinds of threats can also occur from natural disasters.

Fire - A conflagration affects information systems through heat, smoke, or suppression agent (e.g., fire extinguishers and water) damage. This threat category can be minor, major, or catastrophic. **Controls:** install smoke detectors near equipment; keep fire extinguishers near equipment and train employees in their proper use; conduct regular fire evacuation exercises.

Environmental failure - This type of disaster includes any interruption in the supply of controlled environmental support provided to the operations center. Environmental controls include clean air, air conditioning, humidity, and water. **Controls:** since humans and computers don't coexist well, try to keep them separate. Many companies are establishing command centers for employees and a "lights-out" environment for the machines. Keep all rooms containing computers at reasonable temperatures (60 to 75ºF or 10 to 25ºC). Keep humidity levels at 20 to 70% and monitor environmental settings.

Earthquake - A violent ground motion results from stresses and movements of the earth's surface. **Controls:** keep computer systems away from glass and elevated surfaces; in high-risk areas secure the computers with anti-vibration devices.

Liquid Leakage - A liquid inundation includes burst or leaking pipes and accidental discharge of sprinklers. **Controls:** keep liquid-proof covers near the equipment and install water detectors on the structural floor near the computer systems.

Lightning - An electrical charge of air can cause either direct lightning strikes to the facility or surges due to strikes to electrical power transmission lines, transformers, and substations. **Controls:** install surge suppressors, store backups in grounded storage media, install and test Uninterruptible Power Supply (UPS) and diesel generators.

Electrical Interruption - A disruption in the electrical power supply, usually lasting longer than one-half hour, can have serious business impact. **Controls:** install and test UPS, install line filters to control voltage spikes, and install antistatic carpeting.

CONCLUSIONS

Physical security involves the use of safeguards that are **interdependent** within a **system** that can **protect, detect** and **respond** to an unwanted event. In order to understand how to create an effective physical security system, it is important to understand the **elements**, **methods** and **applications** that can be used to design an effective and safe environment for the targets or assets.

It could be argued that, when having to deal with a skilled and determined adversary, the first safeguards to be considered when designing physical security should provide detection. This evolves from the idea that, given enough time, any physical barrier can be overcome. This position is further supported by the fact that if detection is early it facilitates the arrival of a response before the delay provided by the protection safeguard elapses, thus preventing the compromise.

However, since the elements of protection, detection and response are interdependent and must be considered relative to one another, it does not really matter with which element the design starts as long as all elements are looked after.

Companies where employees hold open the door for others to walk through may need to review their level of security awareness. The first step in implementing a physical security program is determining the level of need and the current level of awareness. To implement a cost-effective security program (1) analyze the problems, (2) design or procure controls, (3) implement those controls, (4) test and exercise those controls, and (5) monitor the controls. Implement only controls needed to meet the current needs, but make sure that additional control can be added later if required.

Physical security is an organization’s first line of defense against theft, sabotage, and natural disasters.

REFERENCES

1. Publications

- Effective Security, 2nd ed., Lawrence Fennelly, Butterworth-Heinemann, 1997
- Introduction to Security, Robert J. Fischer, Gion Green, Butterworth-Heinemann, 1998
- Russell, D. and Gangemi, G.T., *Computer Security Basics,* O' Reilly & Associates, Inc., Sebastopol, CA, 1991.
- Jackson, K. and Hruska, J., *Computer Security Reference Book,* CRC Press, Inc., Boca Raton, FL, 1992.

2. Internet

1. **http://www.rcmp-grc.gc.ca, accessed on June, 2nd, 2010**
2. **http://www.infosectoday.com, accessed on June, 2nd, 2010**
3. **http://www.ccert.edu.cn, accessed on June, 6th, 2010**
4. **http://www.foundstone.com, accessed on June, 7th, 2010**
5. **www.tbs-sct.gc.ca, accessed on June, 9th, 2010**
6. **www.authorstream.com, accessed on June, 10th, 2010**
7. **http://www.servepath.com, accessed on June, 13th, 2010**
8. **http://searchsecurity.techtarget.com , accessed on June, 13th, 2010**
9. **http://www2.fpm.wisc.edu , accessed on June, 13th, 2010**
10. **http://www.reliablefire.com , accessed on June, 13th, 2010**
11. **http://www.sans.org , accessed on June, 13th, 2010**
12. **http://en.wikipedia.org , accessed on June, 13th, 2010**
13. **http://security.uchicago.eu , accessed on June, 13th, 2010**
14. **http://www.cccure.org , accessed on June, 13th, 2010**
15. **http://www.securityfocus.com , accessed on June, 13th, 2010**

SECURING DNS

CPT Lucian CROITORU

INTRODUCTION

DNS is powerful, ubiquitous and largely ignored. That's a very dangerous combination.

Virtually All Applications Rely on DNS.

- Email
- The world wide web
- Peer to peer applications
- Instant messaging
- Voice over IP, etc., etc., etc.

Virtually ALL applications are built on top of DNS, and rely on DNS to function. This puts DNS in a radically different role than an application such as FTP – if FTP doesn't work, everything else will continue to function, but that's not true of DNS! If DNS is down, everything else also tends to come to a screeching halt.

DNS is the foundation technology (or at least DNS is one of just a handful of particularly key foundation technologies – I'll certainly concede that BGP is equally as important as DNS, for example).

I. WHY WORRY ABOUT DNS?

“If I Can Control Their DNS… … I can control their world.”

Going to eBay? Doing some online banking? Sending important email? Maybe, maybe not, depending on what sort of DNS resolution occurs. If a bad guy controls your DNS, he can send you to a convincing alternative site under his control…

"But, but… even if the bad guys hijack my DNS, the fake website they might have set up won't have the right SSL certificate!"

SSL certificate issues are not enough to flag DNS misdirection as an issue -- users just don't get the whole certificate thing, and will just blindly accept any self-signed certificate they've been handed for a "secure" site.

Just as most non-technical users don't "get" subtle SSL certificate related issues, most non-technical users also don't "get" DNS.

Because DNS is, or can be, complex, and because non-technical users generally don't need to understand DNS to use the Internet (at least when everything is working the way it is supposed to), many people never bother to learn anything about DNS -- it just works, and they blindly and trustingly rely on it.

Unfortunately, because DNS usually "just works," users are not sensitized to the ways that DNS can be perverted or corrupted by a miscreant, and DNS-related areas are not the focus of most consumer-grade system security review tools.

This increases the need for technically-oriented security professionals to pay attention to DNS on behalf of our non-technical users.

The Bad Guys and Gals Are Interested in DNS & Do Understand DNS-Related Vulnerabilities.

Miscreants can (and have!) attacked the trustworthiness of DNS data on a variety of levels, including:

- doing cache poisoning, where misleading results are seeded into the DNS data that many DNS servers save locally, eventually getting provided to local users even though it's inaccurate;
- releasing malware that tweaks host file entries and/or DNS registry entries on the PC, so the bad guys send users directly to the wrong web site rather than the web site they intended.

Some hacker/crackers also view DNS as a convenient mechanism whereby they can limit user access to key resources, such as antivirus updates needed for the remediation of infections.

The bad guys also recognized DNS is a key enabling technology for botnet command and control survivability.

Sometimes security guys are accused of sowing fear, uncertainty and doubt (FUD), but truly, DNS is potentially an incredibly potent "death ray." Why?

- There are **millions** of DNS servers deployed on the Internet.
- **DNS uses UDP**. Because of that, **DNS has issues when it comes to accepting and responding to spoofed query sources**.
- **Because DNS accepts a tiny query as input, and (potentially) generates a huge response as output, DNS operates as a high-gain online traffic amplifier**.

There's also the simple reality: DNS servers are used to conduct some of the largest DDoS attacks in history.

Speaking of DDOS, DNS Servers Are A Prime Target for DDoS, Too…

Name servers aren't just a tool for conducting distributed denial of service attacks, customer-facing recursive **DNS servers are also a target for distributed denial of service attacks**: if someone can kill the DNS servers our customers are using, we are off the network even if our transit links aren't flooded with traffic.

DNS Services Have Been Broadly Neglected.

DNS has traditionally not been a focus of institutional love and investment. When it comes to DNS, lots of people are running:

- old code,
- on old gear,
- with crude operational tools,
- a low level of redundancy,
- poor service monitoring and
- part time or student (rather than fulltime) DNS administrators.

DNS isn't "cool."

Doing DNS for a organisation is not a particularly glamorous or high prestige job (few novices aspire to some day become a DNS administrator – they all want to work in Marketing, instead!

There are no routinely scheduled reoccurring conferences devoted exclusively to DNS-related research or operational praxis, with the exception of ISC's OARC meetings (see https://www.dns-oarc.net/).

An effort by ICANN staff to create a DNS-CERT has not exactly been enthusiastically embraced (see http://www.icann.org/en/public-comment/#dns-cert).

DNS is thus simultaneously operationally critical **and** managerially insignificant to the point of often being obscure/unknown.

DNS Is No Longer Just for Translating Domain Names to IP Addresses.

DNS has become a general-purpose distributed database.

DNS block lists, as used to block spam, are one example of nontraditional data distributed via DNS, and RouteViews IP-to-ASN data is another, and ENUM data (see www.enum.org) is a third.

A comment from Eric A. Hall, ca. April 16, 2001: *"The current DNS will only keep working if it is restrained to lookups, the very function that it was designed to serve. It will not keep working if the protocol, service, tables and caches are overloaded with excessive amounts of data which doesn't benefit from the lookup architecture."*

http://www.ops.ietf.org/lists/namedroppers/namedroppers.2001/msg00247.html

That comment notwithstanding, people are now doing wild stuff.

Our DNS (or, more precisely, our rDNS) may determine how some people decide to treat our email and other network traffic.

For example, some ISPs check that rDNS exists for a host that is attempting to send mail. **No rDNS?** For a growing number of sites that means, "Sorry, we won't be able to accept email from that dotted quad…" For instance, see

http://postmaster.aol.com/guidelines/standards.html and http://help.yahoo.com/l/us/yahoo/mail/postmaster/basics/postmaster-15.html

Other sites may also be on the lookout for dynamic-looking rDNS host names when deciding whether to accept or reject direct-to-MX email. Have rDNS which looks dynamic? Again, for many sites, that means "Sorry, but we won't be accepting email directly from you, send it via your provider's official SMTP servers…"

Examples of "Dynamic Looking" rDNS:

adsl.nuria.telefonica-data.net
cable.mindspring.com
dhcp.vt.edu
dialup.hawaii.edu
dorm.ncu.edu.tw
dsl.telesp.net.br
dyn.columbia.edu
dynamic.hinet.net
dynamicip.rima-tde.net
user.msu.edu
wireless.indiana.edu

There are efforts underway in the IETF to encourage consistent use of rDNS, and to standardize rDNS naming practices. Two drafts we should be aware of:

- Considerations for the Use of DNS Reverse Mapping http://www.ietf.org/internet-drafts/draft-ietf-dnsop-reverse-mapping-considerations-06.txt (expired)
- Suggested Generic DNS Naming Schemes for Large Networks and Unassigned hosts http://tools.ietf.org/id/draft-msullivan-dnsop-generic-naming-schemes-00.txt (also now expired).

DNS interacts with lots of other things. For example, how do hosts learn which DNS servers they should be using? Users of static IP addresses may be given static DNS server configuration information, but most users who are using dynamic addresses will get their DNS server information from **DHCP** at the same time they receive an IP address to use.

Thus, if we care about the security of DNS, we really want to pay attention to the security of DHCP, too. Why? If we don't pay attention to the security of DHCP, the bad guys and gals can attack the security of our DNS indirectly, by **attacking DHCP**.

The attack would not have to be hard: for example, imagine a rogue DHCP server sitting on the wire and listening for DHCP requests… first server to respond to a DHCPDISCOVER with a DHCPOFFER typically "wins”.

Sample DHCP malware: isc.sans.org/diary.html?storyid=6025

DNS also interacts with NTP (time). Just as DNS and DHCP are tightly coupled, we should also know that DNS can also rely critically on accurate system clocks.

Two examples:

- From the the BIND FAQ (http://www.isc.org/software/bind/faq):

  "**Q:** I'm trying to use TSIG to authenticate dynamic updates or zone transfers. I'm sure I have the keys set up correctly, but the server is rejecting the TSIG. Why?

  "**A:** This may be a clock skew problem. Check that the clocks on the client and server are properly synchronised (e.g., using ntp)."

- If we're trying to identify who was using a dynamic IP address at a given time, it can be critical to have accurate time stamps (including time zone information!)

DNS may control access to resources. Consider, for example, a site-local resource, like a USENET News server, or a site-licensed database. Access to those resources may be controlled by password, or by limiting access to a particular network range, **but** many times access is controlled by limiting access to a particular domain, e.g., "If the connection is coming from an IP address which has the rDNS of *.nato.int, allow access to that resource."

Of course, it is entirely possible that a bad guy or bad gal might create a bogus in-addr for a non-institutional address, thereby pretending to be part of a domain to which they really don't belong; checking to make sure that the forward address and the reverse addresses seen agree helps reduce the magnitude of this issue, but this is still a fundamentally weak approach to the problem of controlling access.

Relying on rDNS means that location can be a replacement for identity (all I need is an open jack somewhere and I'm OK).

DNS may play an infrastructural role. For example, DNS can be used for traffic management and load balancing, perhaps with DNS selectively returning different dotted quads based on a query's geographical or organizational source.

Yes, for most of us this is inconsistent with the goal of having consistent information returned regardless of query source, but highly tailored non-uniform DNS operation is highly valued by some commercial sites which may want to do things like:

- send users to a topologically "close" server farm
- serve a internationalized, language appropriate version of their web site, perhaps in German for users coming from IP's known to be located in Germany, French for users coming from IP's known to be in France, etc.

- display a specially tailored version of their web site for particularly important customers, or a version that has had unacceptable content removed for particular cultural venues.

Round Robin DNS vs. Load Balancers.

Another example of how DNS may be used to manage traffic can be seen in the use of round robin DNS, where multiple IPs are bound to a single fully qualified domain name (FQDN).

When doing round robin DNS, name servers sequentially return each defined dotted quads in turn, providing a sort of crude (and potentially multi-site) alternative to dedicated load balancers such as Ultramonkey (see http://www.ultramonkey.org/).

The down side to doing round robin DNS instead of something more sophisticated? Potentially many things, including:

- caching can screw things up (delay changes in configurations)
- load division is crude at best, and not load aware in any way
- if we "lose" a host in an N-host round robin, every 1-in-N times someone tries to access that site, there will be a failure
- failed hosts do not get automatically removed from the rotation
- debugging round robin DNS issues can be a real pain

DNS can affect network planning. How much load will our DNS servers (and network) see? Choice of DNS TTLs (time to live) may directly impact that…

Speaking of DNS TTLs, if our DNS servers are temporarily down, how long will sites on the network continue to use cached values? (And is this caching good, or does it just help us conceal (rather than fix) substandard DNS infrastructure?)

Still thinking about DNS TTLs, if we experience a disaster and need to move servers, how long will it take for cached values to "cook down" so that new DNS values can be noticed?

What about dynamic addresses? How long should dynamic address leases be? How big should DHCP pools be?

Planning on doing IPv6? How we handle DNS is an integral part of that, whether that's numbering plans, provisioning quad A records, making local DNS servers available via IPv6, etc.

DNS can interact with and impact policy issues in myriad interesting ways.

For example, what does our organisation privacy policy say about DNS server logs? Has our site even thought about why DNS server logs may be sensitive? (Perhaps some member of our community has an embarrassing health condition, and the DNS server logs expose that condition by documenting visits to a site for those suffering from chronic hemorrhoids (or acute leukemia)).

Or what if a key employee is suddenly resolving domain names associated with executive recruiters or online job sites?

Some DNS policy areas:

Who/what organization does DNS for our organisation?

Who can get DNS service from that organization?

Is there a charge for this service?

What's an acceptable DNS name?

What if the FQDN I want is already taken? Can I "bump" them?

Can I get a subdomain?

What determines if I get a static or dynamic address?

Can institutional FQDNs point at non-institutional IPs?

Can non-institutional FQDNs point at institutional IPs?

Does it matter if a domain is a .edu instead of a .com or .org or .net or .us or something else?

And many more areas…

An international policy example: **IDN.**

Since we're westerners and use a Roman alphabet, we probably give scant thought to all the people abroad who may wish they could use accented characters, or Greek letters, or Kanji, or Hangul, or Cyrillic letters as part of domain names…

Surely accommodating the diverse needs of those with non-Roman character sets can only be good, right? Why would that raise policy issues? There are many reasons, including:

- can all name servers technically accommodate non-Roman names?
- what representation should be used for foreign character sets? Choices are potentially legion (and sometimes highly political)
- what about internationalized names which look *almost* the same as already registered names belonging to banks or other phishing targets? (this is often called a homographic attack; see http://www.shmoo.com/idn/homograph.txt for more info).

IDNs have come a long way in the last few years.

Most web browsers now support for IDNs, and 19 internationalized TLDs representing 11 languages have been requested as of April 2010 (see http://icann.org/en/topics/idn/fast-track/ and http://icann.org/en/topics/idn/fast-track/string-evaluation-completion-en.htm).

IDNs are currently available for some existing TLDs (e.g., in dot com one can register punycoded domains: http://xn--hq1bp8p1yi.com/).

Conclusion: DNS is a very important and timely area that "punches through" a lot of background noise.

Important DNS characteristics:

Be available (remember, if the domain name system is unavailable, for most users, the "Internet is down").

Be trustworthy (if the domain name system returns untrustworthy values, we may be sent to a site that will steal confidential data, or to a site that could infect our computer with malware).

Be fast (rendering even a single web page may require tens – or hundreds! -- of domain name system queries; can we imagine waiting even a second for each of those queries to get resolved?)

Be scalable (there are billions of Internet users who rely on DNS, all around the world).

Be flexible (different sites may have different DNS requirements).

Be extensible (there are still many things that DNS will be called upon to do, but we don't know what all those things are yet! We need to have the flexibility to evolve DNS as time goes by).

II. HOW DNS CURRENTLY WORKS

We do need to agree on some terminology and provide a little historical background.

Pretty much everyone (IT specialist) conceptually understands how the Domain Name System (DNS) works, but let just begin with a brief (and very incomplete) functional definition:

"DNS is the network service that translates a fully qualified domain name, such as *www.act.nato.int*, to a numeric IP address, such as *128.223.142.89*. DNS can also potentially do the reverse, translating a numeric IP address to a fully qualified domain name."

Whenever we use the Internet we're using DNS, and **without DNS, using the Internet would become very inconvenient**. Can we imagine having to remember to go to http://66.102.7.147/ instead of http://www.google.com/ for example?

How does the DNS system *currently* work? While the fine points can vary, the basic process is:

1) An application (such as a web browser) requests resolution of a fully qualified domain name, such as www.act.nato.int
2) If the desktop operating systems includes a caching DNS client, the DNS client checks to see if that FQDN recently been resolved and cached (stored locally) -- if yes, it will use that cached value.
3) If not, the desktop DNS client forwards the request for resolution to a recursive DNS server which has been manually pre-configured (or to a recursive DNS server which may have been designated as part of DHCP-based host configuration process)
4) If the recursive DNS server doesn't have a recently cached value for the FQDN, the recursive DNS server will begin to make queries, if necessary beginning with the DNS root zone, until it has resolved a top level domain (e.g., .int), primary domain name (nato.int), and finally a FQDN (such as www.act.nato.int)

What we should understand is that DNS is **inherently distributed service** – every sites doesn't need to store a copy of the the complete Internet-wide mapping of FQDN's to IP addresses.

This differs dramatically from **pre-DNS** days, when mappings of host names to IP addresses happened via **hosts files**, and each server would periodically retrieve updated copies of the hosts file.

Fortunately, because DNS is distributed, it scales very well, far better than replicating host files!

Unfortunately, because DNS is distributed, it is more complex than the conceptually simple (if practically unworkable) hosts file solution, and there can be substantial variation in how, and how well, sites and DNS administrators do DNS-related activities.

DNS efficiencies

Most common DNS queries do not require re-resolving the TLD (.edu, .com, .net, .org, .biz, .info, .ca, .de, .uk, etc.) name servers, or even the name servers for 2nd level domains such as google.com or microsoft.com -- those name servers change rarely if ever, and will typically be statically defined via "glue" records, and cached by the local recursive name server. (Glue records assist with the DNS bootstrapping process, providing a static mapping of name server's FQDNs to its associated dotted quad.)

Cached data which has been seen by a DNS server will be reused until it "cooks down" or expires; cache expiration is controlled by the TTL (time to live) associated with each data element. TTL values are expressed in seconds.

Negative caching (the server may remember that a FQDN **doesn't** exist) may also help reduce query loads.

Notes:

- The DNS entries for domains are contained in **zones**. For example, there would normally be one zone for nato.int and another zone for rto.nato.int
- The **primary** or "master" DNS server for a given domain normally is augmented by a number of **secondary** (or "slave") DNS servers. Secondary servers are deployed to help insure domains remains resolvable even if a primary server becomes unreachable.
- Secondary DNS servers periodically retrieve updated zone data for the zones they secondary from the primary DNS server. Most sites limit who can download a complete copy of their zone file because having a definitive listing of all hosts in a given domain may be useful for cyber reconnaissance and attack purposes.
- It is common for organisations to agree to provide secondary DNS service for each other, e.g., NATO ACT, Norfolk VA does runs a secondary for JFTC Bydgoszcz, Poland.

Some are becoming interested in DNS because of new potential roles, including:

- … as a new way of **identifying** infected systems
- … as a new way of **mitigating** infected systems
- … as a new way of "**monetizing"** typos and other domain name resolution "misses"
- … as something which will **needs to be fixed** after miscreant name servers get taken off the air.

And then there's everyone else, who just wants DNS to keep working…

Let's talk about one of the biggest threats to DNS, spoofed traffic used as a denial of service attack tool.

III. SPOOFED (DNS AND OTHER) TRAFFIC AND DISTRIBUTED DENIAL OF SERVICE ATTACKS

III.1 Distributed Denial of Service (DDoS) Attacks

In a distributed denial of service (DDoS) attack, network traffic from thousands of hacked computer systems -- often systems located all over the Internet -- gets used in a coordinated way to overwhelm a targeted network or computer, thereby preventing the target from doing its normal work.

Unlike that earlier general talk, today we **do** need to talk a little about a specific technical vulnerability. We need some quick background, first.

III.2 TCP and UDP Traffic

There are basically two types of network application traffic: TCP and UDP.

TCP traffic is associated with relatively persistent connections (such as ssh sessions, web traffic, email, etc.), and has a variety of characteristics which are desirable from a network application programmer's point of view, including retransmission of lost packets, congestion control, etc.

UDP traffic, on the other hand, is designed for "send-it-and-forget-it" applications where we don't want to/can't afford to maintain state or we don't want a lot of connection setup overhead.

DNS, NFS, and IP video traffic all normally run as UDP.

III.3 The Spoofability of UDP Connections

Unlike a fully established TCP connection (which only gets established after a bidirectional handshake is negotiated and which is therefore robust to spoofing attempts), * UDP traffic can be created with virtually **any** apparent source address -- including IP addresses which have no relationship to the traffic's actual origin.

Network traffic that's intentionally created with a bogus source address is said to be "spoofed."

If allowed to reach the global Internet, spoofed traffic is generally indistinguishable from legitimate traffic.

Why Would Anyone Bother to Spoof Traffic?

If we don't spend time "thinking like an attacker," we might not immediately "get" why an attacker would be interested in spoofing his attack traffic. The answer is actually quite simple: the attacker wants the systems he's using as part of his attack to stay online and unblocked as long as possible.

Spoofing the source of the attack traffic…

- hinders backtracking/identification/cleanup of the system that's sourcing the traffic;
- makes it harder for the attack victim to filter the attack traffic (the spoofed source addresses may be constantly changed by the attacker, and thus do not provide any sort of stable "filterable characteristic").

"So Why Not Just Block All UDP Traffic?"

Given that UDP can be easily spoofed by the bad guys/bad gals, sometimes we'll hear people naively propose simply blocking all inbound or outbound UDP traffic (or at least heavily rate limiting all UDP traffic).

Unfortunately, because some pretty basic services (like DNS) require support for UDP, blocking (or heavily rate limiting) all inbound or outbound UDP traffic is generally **not** a good idea.

Want or not, we have no choice but to learn to live with UDP traffic.

"Well, Can We Block SOME UDP Traffic?"

For once, the answer is positive: yes, we can block some UDP traffic.

For example, if you're DRESMARA and your Department has been assigned the IP address range 128.223.0.0-128.223.255.255 there's no reason for systems on your network to be sourcing packets that pretend to be from some other IP address range. You'd filter that spoofed traffic before it leaves your organisation.

III.4 Subnet-Level Filtering

While it is great to prevent spoofing at the organisation-wide level, that sort of border router anti-spoofing filter does not prevent a miscreant from forging an IP address taken from one of our subnets for use on another of our subnets.

We KNOW that hosts on each subnet should ONLY be originating packets with IP addresses legitimately assigned to that subnet, so at the uplink from each subnet, drop/block outbound packets that appear to be "from" any other IP address – another very basic sanity check.

III.5 Filtering at Other Levels of Granularity

Although we've talked about filtering at our border and at each subnet uplink, we could also filter all the way upstream at the regional optical network (“RON”) level/the gigapop level, or all the way downstream at the host level.

Obviously, the closer we get to the traffic source, the more effective the anti-spoofing filter will be. That said, catching at least some problematic traffic at the RON/gigapop level is better than nothing if we can't get our downstream customers to do the right thing closer to the traffic source (but the larger our gigapop, the harder it will be to keep accurate track of all the prefixes that may be in use).

III.6 BCP38/RFC2827

Ingress filtering of traffic with spoofed IP addresses is not new and is not my idea – it is Best Current Practice (BCP) 38/RFC2827, written by Ferguson and Senie in May 2000.

Unfortunately, despite being roughly ten years old, **many** sites still do **NOT** do BCP38 filtering -- currently 15-24% Internet wide depending on whether we count netblocks, dotted quads or ASNs (see http://spoofer.csail.mit.edu/summary.php).

"So Why Doesn't Everyone Do BCP38 Filtering?"

"Too hard given the complexity of our network"

Asymmetric costs/benefits: filtering our network protects us (which is nice), but filtering that traffic "costs" us w/o any tangible/economic "benefits." So what are these “horrible” "costs?"

- engineer time to configure and maintain the filters (one time/negligible for most relatively static networks)
- overhead on the routers (but if that overhead is material enough to be a "show stopper," we should upgrade our hardware anyway)
- too busy (or other excuses).

Some may question why others should care what they do with their networks?

However in this case, remember that if we're NOT doing BCP38 filtering, our network may be getting used to generate spoofed attack traffic that's pretending to be "from" someone else's network, and that's the point at which what we do (or don't do) potentially affects a lot of other people including the attack target itself, the entity whose IP addresses are being spoofed, etc.

IV. OPEN RECURSIVE DNS SERVERS AND DNS AMPLIFICATION ATTACKS

Since we just got done covering UDP spoofing, talking a little about open recursive domain name servers and DNS amplification attacks seems like a "nice" segue/practical example of why BCP38 filtering is important, while also pointing out another specific vulnerability we should be addressing.

Again, let's begin with a little more background, however, first.

IV.1 Authoritative and Recursive DNS Servers

There are different types of name servers, with "authoritative" and "recursive" DNS servers being the two most important types:

- Authoritative servers are definitive for particular domains, and should provides information about those domains (and ONLY those domains) to anyone.
- Recursive servers are customer-facing name servers that should answer DNS queries for customers (and ONLY for customers) concerning any domain.

DNS servers that aren't appropriately limited can become abused.

For example, consider a situation where a DNS server is recursive AND is open for use by anyone (a server that's cleverly termed an "open recursive DNS server").

While it might seem sort of "neighbourly" to share our name server with others, in fact it is a really bad idea (the domain name system equivalent of running an open/abusable SMTP relay, in fact).

The problem? Well, there are actually **multiple** problems, but one of the most important ones is associated with spoofed UDP traffic.

IV.2 Example of Spoofed DNS Attack Scenario

- Attacker, who's working from non-BCP38 filtered network. Let's call him/her "A"
- Attack target – let's refer to that entity as "T"
- Open recursive domain name server on large, high bandwidth pipe, denoted below as "NS"

Scene:

- "A" generates spoofed DNS queries with "T"'s address as the "source" address of the queries
- "NS" receives the spoofed queries and dutifully returns the "responses" for those queries to "T"
- "A" repeats as desired, thereby DoS'ing "T" via "NS"

Notes

- From "T"'s point of view, the attack comes from "NS" not from "A"
- DNS queries are small and use UDP, so an attacker can readily generate a "large" query volume
- DNS response traffic is also UDP, which means that it is insensitive to net congestion.
- DNS responses can be **large** relative to size of DNS queries (output/input ratios can run over 8X on most DNS servers, and on servers supporting RFC2671 EDNS0 extensions, observed amplification can exceed 70X).
- "A" can employ **multiple spoofed query sources**, and **use multiple NS's** for more traffic.

This is a well known/documented threat:

"The Continuing Denial of Service Threat Posed by DNS Recursion, "

see http://www.us-cert.gov/reading_room/DNS-recursion121605.pdf

"DNS Amplification Attacks,"

see http://www.isotf.org/news/DNS-Amplification-Attacks.pdf

"DNS Distributed Denial of Service (DDoS) Attacks,"

see http://www.icann.org/committees/security/dns-ddos-advisory-31mar06.pdf

IV.3 Open Domain Name Servers Worldwide

Unfortunately, despite this being a well known problem, at one point it was estimated that 75% of all name servers worldwide run as open recursive name servers (see http://dns.measurement-factory.com/surveys/sum1.html)

How Can We Find Open Recursive DNS Servers At Our Organisation?

Team Cymru can send us notifications about open recursive DNS resolvers on our organisation; to sign up to receive these notifications: http://www.team-cymru.org/Services/Resolvers/

If we want to test things ourself, one tool which we can use to scan our network for open recursive DNS servers is dnsscan, see http://monkey.org/~provos/dnsscan/

NOTE: Please do **NOT** scan for open recursive DNS servers on any network unless you are explicitly authorized by that network's owner/administrator to do so. Unauthorized scans will likely be considered hostile/illegal and may be treated as a computer intrusion and result in legal action against you.

What About Google's Public DNS Servers?

Some people may be aware that Google announced that it would be running publicly available recursive DNS servers that anyone could use by changing their name servers to point to 8.8.8.8 and/or 8.8.4.4 (see http://code.google.com/speed/public-dns/)

Google is explicitly aware of the risks associated with the service that they're offering, and we can read the discussion of how they address/plan to address that issue at http://code.google.com/speed/public-dns/docs/security.html

I should mention that Google is NOT the only site intentionally making recursive name servers available; other examples include:

- http://www.opendns.com/ (free and paid versions are available)
- https://www.dns-oarc.net/oarc/services/odvr (intended for those who want to try using a DNSSEC-enabled name server)

Coming Back to The General Problem of Open Recursive DNS Servers, The Problem Isn't "Just" About DDoS, Either

If we are not yet sufficiently motivated to fix our DDoS-exploitable domain name servers by the discussion I've provided about DNS amplification, let me add a little more thrust to help launch that hog: if we are not controlling access to our domain name servers, we may also be leaving ourself vulnerable to **DNS cache poisoning attacks**, whereby vulnerable caching name servers can be made to return bogus results for a user's name service queries: www.secureworks.com/research/articles/dns-cache-poisoning.

IV.4 Cache Poisoning Attack

In a nutshell, in cache poisoning attacks, the attacker "primes" the caching name server to respond to queries with an IP address of his/her choice, rather than the real/normal IP address for that site. An innocent victim then asks the caching name server for the IP address of a site of interest, such as the IP address of their bank's website. If the domain name of that site happens to be one that the attacker has poisoned, the victim is automatically and transparently misdirected to a website of the attacker's choice, rather than to their bank's real web site, and confidential data can then end up being lost.

Another cache poisoning scenario uses cache poisoning to redirect queries for popular sites (such as google.com or hotmail.com) to a site that contains a virus or other malware. If our caching name server has been poisoned, when we try to visit one of these popular sites, we can unknowingly be redirected to another site that stealthily tries to infect our PC with malware. Blocking open access to our recursive name servers won't completely eliminate the possibility of our servers participating in such attacks, but it will reduce the likelihood of that sort of abuse.

IV.5 Recommendations to Deal With Open Recursive DNS Servers

- Insure that you're running a current version of BIND (or whatever DNS software you use)
- Insure that you've separated your Internet-facing authoritative name server from your customer-facing recursive name server
- Protect your customer-facing recursive name server from access by non-customers
- Consider implementing the additional DNS server hardening measures described in the Team Cymru BIND Template (see http://www.cymru.com/Documents/secure-bind-template.html)

V. MALWARE AND DNS

It's time to start thinking about how malware interacts with DNS, and what will happen when DNS hijacking malware gets disrupted.

V.1 Spam-Related Malware Relies on DNS

Much of the most virulent malware out there has been deployed to facilitate spamming, and spam-related malware is notorious for generating large numbers of DNS queries for MX host information (so the spamware can determine where it should connect to deliver its spam).

Spam related malware may also refer to its upstream command and control hosts using their FQDNs, thereby making it possible for the miscreants to repoint their malware's command and control host from one dotted quad to another should the systems currently "hosting" their C&Cs get filtered or cleaned up.

At the same time that malware critically **relies** on DNS, ironically other malware may **also** be actively working to block or interfere with legitimate DNS uses.

Authors of viruses, trojan horses and other malware may interfere with user DNS for a variety of reasons, including:

- attempting to block access to remediation resources (such as system patches, AV updates, malware cleanup tools)
- attempting to redirect users from legitimate sensitive sites (such as online banks and brokerages) to rogue web sites run by phishers
- attempting to redirect users from legitimate sites to malware-tainted sites where the user can become (further) infected

- attempting to redirect users to pay-per-view or pay-per-click web sites in an effort to garner advertising revenues

V.2 Examples of Malware Interfering with DNS

• **Trojan.Qhosts** (discovered 10/01/2003)

http://www.sarc.com/avcenter/venc/data/trojan.qhosts.html

"Trojan.Qhosts is a Trojan Horse that will modify the TCP/IP settings to point to a different DNS server."

• **MyDoom.B** (published 1/28/2004)

http://www3.ca.com/securityadvisor/virusinfo/virus.aspx?id=38114

"The worm modifies the HOSTS files every time it runs to prevent access to the following sites [list of sites deleted]"

• **JS/QHosts21-A** (11/3/2004) http://www.sophos.com/virusinfo/analyses/jsqhosts21a.html

"JS/QHosts21-A comes as a HTML email that will display the Google website. As it is doing so it will add lines to the Windows Hosts file that will cause requests for the following websites to be redirected: www.unibanco.com.br, www.caixa.com.br, www.bradesco.com.br"

• **Win32.Netmesser.A** (published 2/1/2005):

http://www3.ca.com/securityadvisor/virusinfo/virus.aspx?id=41618

"[the trojan] then enumerates the following registry entry:

HKLM\SYSTEM\CurrentControlSet\Services\Tcpip\Parameters\Adapters

checking for references to dial up adapters. If found, the adapters' DNS servers are changed by altering the value 'NameServer' in the referenced key." […]

"Computer Associates have seen the following DNS server IPs used by these trojans in the wild: 69.50.166.94, 69.50.188.180, 69.31.80.244, 195.225.176.31"

• **Trojan.Flush.A** (discovered 3/4/2005)

http://www.sarc.com/avcenter/venc/data/trojan.flush.a.html

'Attempts to add the following value […]: "NameServer" = "69.50.176.196,195.225.176.37"'

• **DNSChanger.a** (added 10/20/2005) http://vil.mcafeesecurity.com/vil/content/v_136602.htm

"Symptoms: […] Having DNS entries in any of your network adaptors with the values: 85.255.112.132, 85.255.113.13"

• **DNSChanger.c** (added 11/04/2005) http://vil.nai.com/vil/Content/v_136817.htm

"This program modifies registry entries pertaining to DNS servers to point to the following IP address: 193.227.227.218"

• **Trojan.Flush.K (1/18/2007)**

http://www.symantec.com/enterprise/security_response/writeup.jsp?docid=2007-011811-1222-99&tabid=2 states:

'The Trojan then creates the following registry entries: […]

HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Services\Tcpip\Parameters\Interfaces\[RANDOMCLSID]\"DhcpNameServer" = "85.255.115.21,85.255.112.91"

HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Services\Tcpip\Parameters\Interfaces\[RANDOMCLSID]\"NameServer" = "85.255.115.21,85.255.112.91"'

- **DNSChanger.F (3/27/2007)**

http://vil.mcafeesecurity.com/vil/content/v_141841.htm states that "the main objective of this trojan is to change the default DNSentries to its own [preferred] DNS server."

#HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Services\Tcpip\Parameters\NameServer: "85.255.115.46 85.255.112.154" (This is just an example and IP can vary)

#HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Services\Tcpip\Parameters\DhcpNameServer: "85.255.115.46 85.255.112.154" (This is just an example and IP can vary)

And there are many, many more… The bad guys ARE attempting to accomplish their goals via our users' reliance on DNS.

V.3 Architectural Changes Among ISPs

Confronted with malware that's targeting user DNS settings, providers are forced to think about scalable ("network-centric") ways to deal with those threats.

Coming up with a solution requires understanding the mechanics of how DNS is transported across the network.

The Mechanics: 53/UDP and 53/TCP

Most DNS queries are made over port 53/UDP, but some queries may return more data than would fit in a normal single DNS UDP packet (512 bytes). When that limit is exceeded, DNS will normally truncate, and retry the query via 53/TCP.

Occasionally we may run into a site where either 53/**UDP** or 53/**TCP** has been blocked outright for all IP addresses (including real name servers!) at a site. That's a really bad idea.

Blocks on **all** 53/**TCP** traffic sometimes get temporarily imposed because of the misperception that "all" normal DNS (at least all traffic except for zone transfers) happens "only" via UDP; that is an incorrect belief. Real DNS traffic (other than zone transfers) **can, may and will** actually use 53/TCP from time to time.

Blocks on **all** 53/**UDP** may sometimes get installed because of concerns about spoofed traffic, or worries about the non-rate adaptive nature of UDP traffic in general, or simply by mistake.

Because of the high cost of handling user support calls, some ISPs may attempt to avoid user support calls (and associated costs) by actively "managing" user DNS traffic at the network level.

What does "managing" mean?

- **blocking/dropping all** port 53 traffic, **except** to/from the DNS server(s) that the ISP provides for their customers (this will often be implemented via router or firewall filters)
- **redirecting** some or all user DNS traffic that isn't destined for the ISP's customer DNS servers at Layer 4 (e.g., see:

  http://www.cisco.com/en/US/docs/routers/10000/10008/configuration/guides/ancp/isbl4rdt.pdf at PDF pages 12-13)

But, it is not a great idea of **redirecting or rewriting all customer DNS traffic, or limiting users to just their provider's DNS servers** as a "solution." Why?

- doing DNS filtering/redirection breaks Internet transparency in a very fundamental and bad way
- if the provider's designated DNS servers end up having issues, DNS filtering/redirection substantially reduces customer options
- port-based filtering/redirection can be surmounted by technically clued people thru use of non-standard ports for DNS
- port-based filtering/redirection (or even deep packet inspection approaches) can be overcome by VPN-based approaches
- some services (such as commercial DNSBLs) may be limited to just subscribing DNS servers; the DNS server that we may redirect users through may not be allowed to access that data.

It could be better to consider passive DNS monitoring as an alternative way of identifying systems which need attention.

What About Blocking *JUST* Malicious DNS Servers at the Network Level?

Assume we succeed in identifying one or more malicious name servers being used by our users. Most security people would then be inclined to do the "logical" thing and block access to those name servers. Good, right? We're protecting our users by blocking access to just those servers. Well… *yes*, we are, but when we do so, when we block those malicious name servers, ALL name resolution for those infested users (crumby though it may be), will typically suddenly cease. "The Internet is down!"

Suggestion: IF we DO decide to block specific malicious DNS servers, we have to notify our support staff so that they can add DNS checks to their customer troubleshooting processes.

V.4 MS Windows and DNS Cache Pollution

While we're talking about DNS and Windows, some early versions of MS Windows, such as Windows NT and pre-SP1 versions of Windows 2000, are vulnerable to what Microsoft refers to as "cache pollution" (for Microsoft's description of this vulnerability, see: http://support.microsoft.com/kb/316786).

What about Windows 2003? With 2003 we are protected by default, but make sure that Windows Server 2003 admins **do NOT uncheck** the pre-checked "prevent cache pollution" box!

VI. HARDENING DNS

If we are running a DNS server, what steps can we take to help harden or protect it?

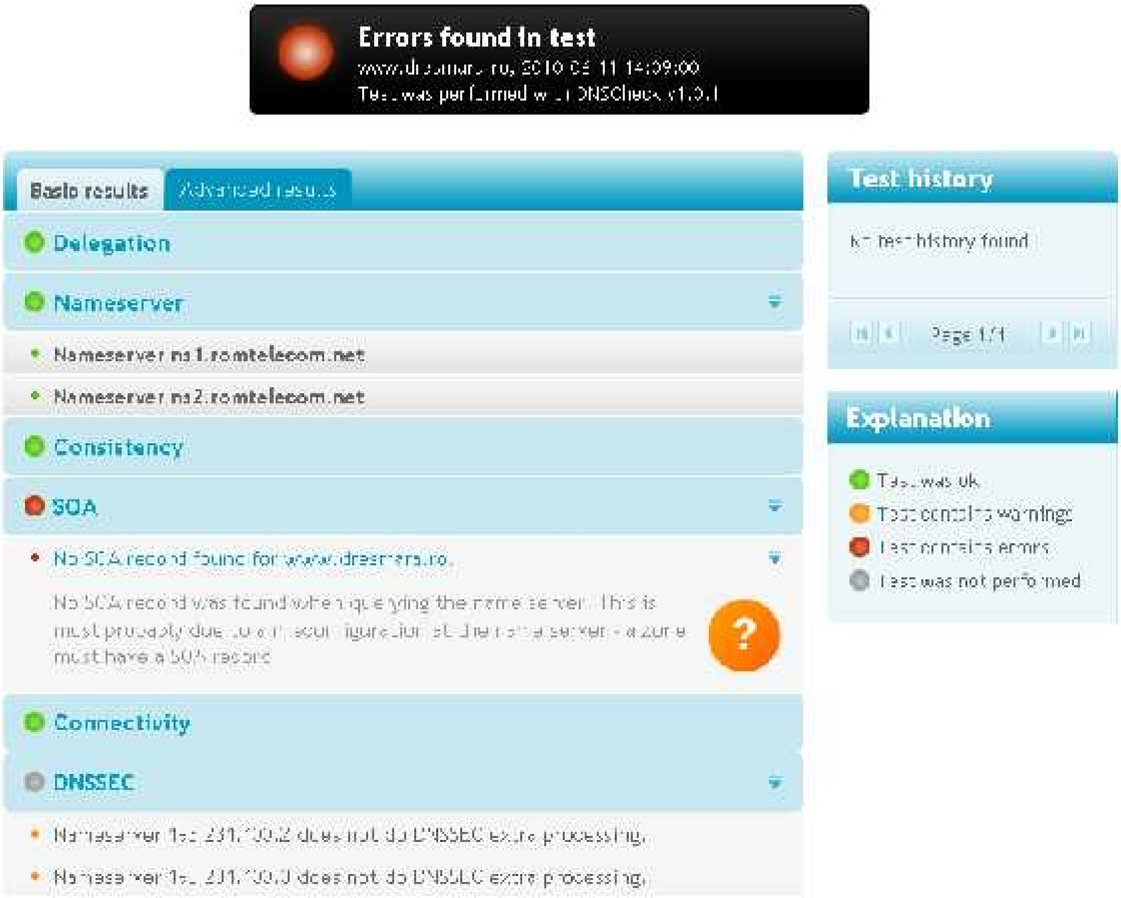
VI.1 Basic DNS Sanity Check

If we do NOTHING else recommended in this paper, I strongly encourage everyone to at least go to http://dnscheck.iis.se/ and conduct a basic test of his organisation's DNS.

That free DNS check will do many basic tests, reporting many DNS-related inconsistencies and DNS-related security issues. The output is easy to understand, and once we know an issue exists, we can then work on getting it fixed.

There are also other online DNS checking tools we can use -- try several and see which works best for us.

Example Output



Errors found in test

www.dreamers.ro, 2010-06-11 14:09:00

Test was performed with DNSCheck v1.0.1

Basic results | Advanced results

● Delegation

● Nameserver

- Nameserver ns1.romtelecom.net
- Nameserver ns2.romtelecom.net

● Consistency

● SOA

- No SOA record found for www.dreamers.ro.

No SOA record was found when querying the name server. This is most probably due to a misconfiguration of the name server - a zone must have a SOA record.

● Connectivity

● DNSSEC

- Nameserver 193.231.100.2 does not do DNSSEC extra processing.
- Nameserver 193.231.100.3 does not do DNSSEC extra processing.

Test history

No test history found

Page 1/1

Explanation

● Test was ok

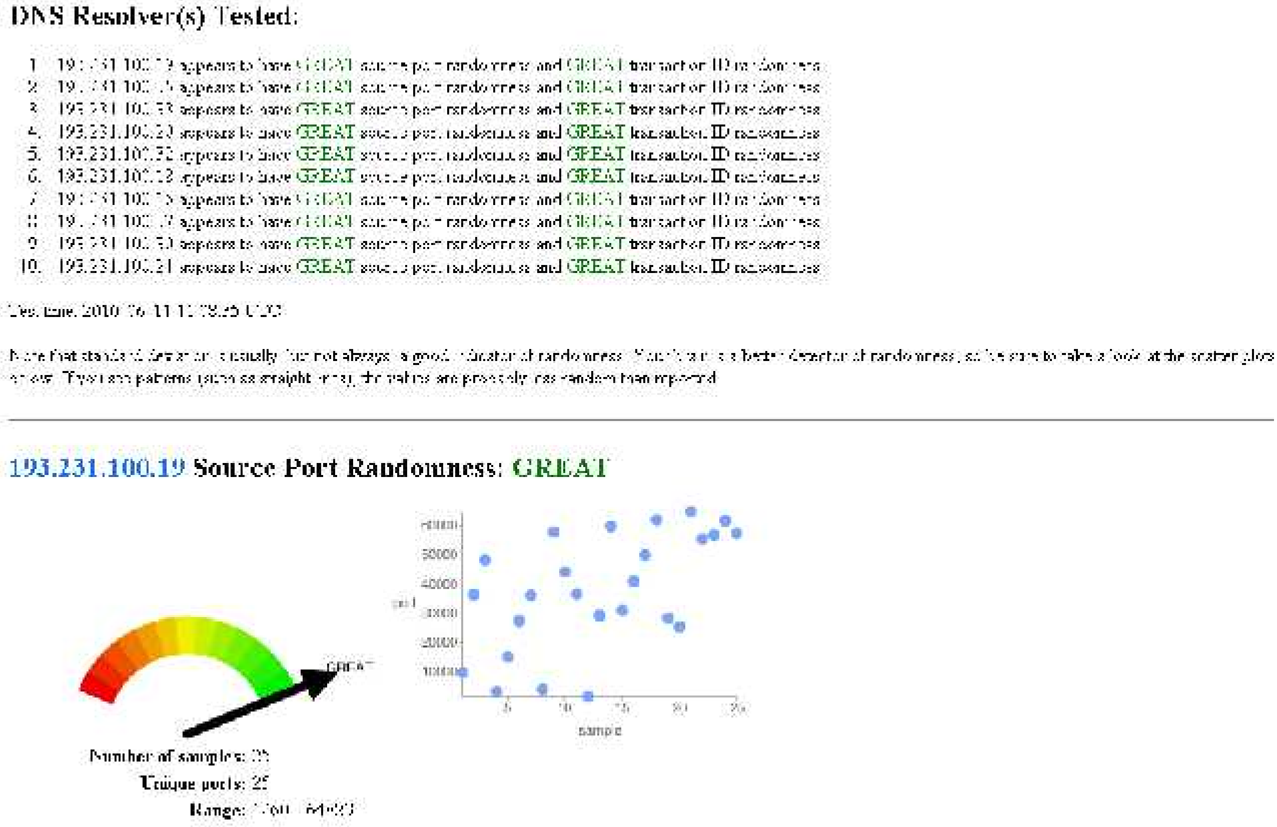
● Test contains warnings

● Test contains errors

● Test was not performed

One Other Test We Should Do

https://www.dns-oarc.net/oarc/services/dnsentropy



DNS Resolver(s) Tested:

1. 193.231.100.19 appears to have GREAT source port randomness and GREAT transaction ID randomness
2. 193.231.100.15 appears to have GREAT source port randomness and GREAT transaction ID randomness
3. 193.231.100.33 appears to have GREAT source port randomness and GREAT transaction ID randomness
4. 193.231.100.20 appears to have GREAT source port randomness and GREAT transaction ID randomness
5. 193.231.100.32 appears to have GREAT source port randomness and GREAT transaction ID randomness
6. 193.231.100.18 appears to have GREAT source port randomness and GREAT transaction ID randomness
7. 193.231.100.[illegible] appears to have GREAT source port randomness and GREAT transaction ID randomness
8. 193.231.100.17 appears to have GREAT source port randomness and GREAT transaction ID randomness
9. 193.231.100.30 appears to have GREAT source port randomness and GREAT transaction ID randomness
10. 193.231.100.21 appears to have GREAT source port randomness and GREAT transaction ID randomness

Test time: 2010-06-11 11:[illegible] UTC

Note that standard deviation is usually (but not always) a good indicator of randomness. Your brain is a better detector of randomness, so be sure to take a look at the scatter plots below. If you see patterns (such as straight lines), the values are probably less random than expected.

193.231.100.19 Source Port Randomness: GREAT

Number of samples: 25

Unique ports: 25

Range: [illegible]

VI.2 DNS Server Software Versions

Unless we have compelling reason to do otherwise, **run the latest version of the DNS server software we're using**.

For BIND users, this means 9.7.0-P1

- If we're on an earlier version, it is highly desirable that we upgrade to the current version
- Updated versions of BIND can be downloaded from http://www.isc.org/downloads

Note: some vendors may not do a great job of keeping their vendor customized versions up to date. If we are using a vendorsupplied version of Bind, we need to carefully weigh the convenience of running an older vendor supported version of BIND against the strong desirability of running the latest version.

It Isn't Just The Name Server Software

If/when we upgrade BIND, we may notice that BIND isn't the **only** thing that may needs upgrading – how about the status of OpenSSL, for example? Problems with stale versions of OpenSSL are so common that BIND explicitly checks OpenSSL as part of the build process! Note that OpenSSL-1.0.0 was released March 29th, 2010, for example. Updated versions of OpenSSL are available from http://www.openssl.org/source/

VI.3 OS Hardening

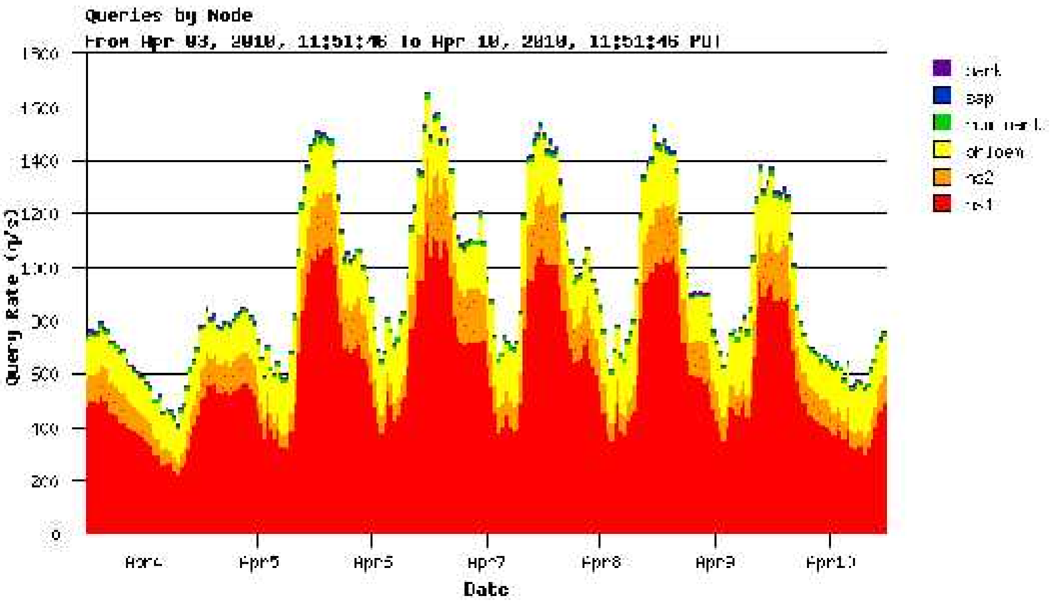
It does little good to run a secure version of the name server software if the operating system that system is running is insecure. Making sure that we're running current versions of OS software and applications are part (but not all) of that picture.

OS hardening is generally beyond the scope of this paper, however a few good starting points include:

- Center for Internet Security "Benchmarks" (checklists), see http://cisecurity.org/en-us/?route=downloads.benchmarks
- See also the National Security Agency's Operating System Guides, http://www.nsa.gov/snac/

VI.4 DNS Monitoring

We should graphically monitor DNS query traffic just as we monitor things like transit bandwidth. A nice tool for this is DNS Stats Collector (DSC), see http://dns.measurement-factory.com/tools/dsc/ (sample below)



Additional DNS Tools

Beyond doing graphical DNS monitoring with DSC, there are additional DNS tools that we may find helpful listed at

- https://www.dns-oarc.net/oarc/tools
- http://dns.measurement-factory.com/tools/
- http://www.dns.net/dnsrd/tools.html

VI.5 Security-As-Availability: Avoid Single Points of Failure

A key step to hardening our DNS service is to look at our architecture with an eye to any single points of failure:

- Do we have multiple physical DNS servers, or just one?
- Assuming we have multiple servers, are they on different subnets?
- Are at least some of our name servers at a different physical location, preferably in a different part of the country?
- If our site uses a border firewall, have we taken steps to make sure all our name servers are not behind a single common firewall?
- Are all of our servers running the same operating system and the same name server software?
- Don't forget our DNS admin, either – do we have at least two people who can handle DNS responsibilities at our site?

VI.6 Network and System Capacity

Because DNS servers may be the target of a denial of service attack, we want to insure that those systems and the connectivity that services them are overprovisioned. While normal traffic loads may require trivial levels of connectivity, if our name server is the target of an attack, we'll find that fast ethernet is better than regular ethernet, and gigabit ethernet is better still. Similarly, a server class system with redundant power supplies an redundant power sources, running as multicore system with plenty of RAM, is also a good idea.

We should run our name servers on dedicated hardware. No other services should be delivered from the name servers – our name servers should be dedicated to just delivering name service!

VI.7 Dynamic DNS (Commercial and RFC2135)

"Dynamic DNS" can refer to two completely different things:

- commercial dynamic DNS service provided by a third party, designed to allow a user to map a vanity domain name or other hostname to a dynamic (rather than static) IP address
- RFC 2135 "Dynamic Updates in the Domain Name System" either as implemented by BIND or Microsoft.

Commercial dynamic DNS service should generally not be needed at most organisations (if someone wants a static IP address, they should generally be able to request and receive one from the organisation); some commercial providers actually forbid use of 3rd party commercial dynamic DNS services.

RFC2135 dynamic updates can cause issues with unnecessary traffic under some circumstances, particularly when they occur in conjunction with NAT'd users, see Section 2.8 of "Observed DNS Resolution Misbehavior" (RFC4697, October 2006).

While it is quite tempting to simply recommend avoiding dynamic DNS updates for philosophical reasons, dynamic updates can have a role in some special circumstances (IPv6, IP mobility, and Active Directory come to mind).

Note that dynamic updates and DNSSEC are also incompatible.

VII. DNSSEC

VII.1 DNSSEC "By the [RFC] Numbers"

DNSSEC is defined by three RFC's:

- RFC4033, "DNS Security Introduction and Requirements,"
- RFC4034, "Resource Records for the DNS Security Extensions,"
- RFC4035, "Protocol Modifications for the DNS Security Extensions"

VII.2 DNSSEC in a Nutshell

DNSSEC uses public key asymmetric cryptography to guarantee that if a DNS resource record (such as an A record, or an MX record, or a PTR record) is received from a DNSSEC-signed zone, and checks out as valid on a local DNSSEC-enabled recursive name server, then we know:

- it came from the authoritative source for that data
- it has not been altered in it’s route
- if the server running the signed zone says that a particular host does not exist, we can believe that assertion.

But what about other things, like insuring that no one's sniffing our DNS traffic, or making sure that DNS service is always available?

DNSSEC Intentionally Focuses on Only One of The Three Traditional Information Security Objectives

While there are three "C-I-A" information security objectives:

- Information **Confidentiality**
- Information **Integrity**, and
- Information **Availability**

DNSSEC is intentionally **NOT** designed to keep DNS data confidential, and it is also intentionally **NOT** designed to improve the availability of DNS data -- it's sole focus is on insuring the **integrity** of DNS data.

And, to the extent that DNSSEC is not an end-to-end protocol, it’s ability to even insure information integrity is imperfect.

VII.3 DNSSEC As A Non-"End-to-End" Protocol

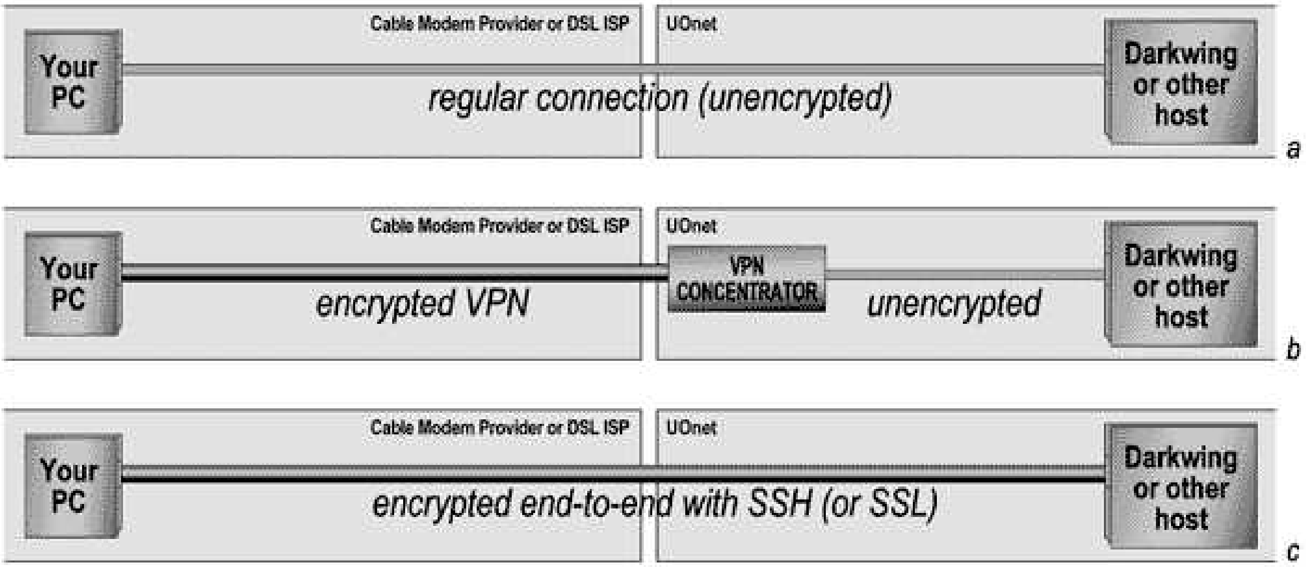
To understand the difference between an end-to-end protocol and one that works only along part of a complete path (e.g., to or from some intermediate point), consider the difference between using SSH and using a typical VPN.

SSH secures traffic all the way from one system (such as a laptop) to the other system connected to (perhaps a server running Linux) – it is "end-to-end."

A VPN, however, may terminate on a hardware firewall or VPN concentrator, and from that point to the traffic's ultimate destination, traffic may travel unsecured. This is NON end-to-end.

DNSSEC is more like the VPN example than the SSH example: **DNSSEC only secures traffic to the local recursive name server**, it typically cannot and will not secure traffic all the way down to the desktop. Thus, a bad guy can still attack DNS traffic that is in flight from the local recursive name server to the endhost.

Non-End-to-End and End-to-End Protocols


Cable Modem Provider or DSL ISP
UOnet
Your PC
regular connection (unencrypted)
Darkwing or other host
a
Cable Modem Provider or DSL ISP
UOnet
Your PC
encrypted VPN
VPN CONCENTRATOR
unencrypted
Darkwing or other host
b
Cable Modem Provider or DSL ISP
UOnet
Your PC
encrypted end-to-end with SSH (or SSL)
Darkwing or other host
c

What About Using TSIG To Secure The Last Hop for DNSSEC?

TSIG **(Transaction SIGnature)** is defined by RFC2845, and was originally created to improve the security of zone transfers, and to provide a secure way by which trusted clients could dynamically update DNS.

For the purpose of providing DNSSEC with last hop integrity, TSIG has a number of potential shortcomings, including:

- it uses a form of symmetric cryptography, so all clients need to be given a copy of a shared secret key
- the only hashing mechanism defined for TSIG in the RFC is HMAC-MD5, which is no longer particularly robust
- clocks need to be roughly in sync (user laptops or desktops often have system clocks which aren't very well synchronized)

The DNSSEC data validation check could be moved from the local recursive DNS server all the way down to the laptop or desktop itself, **IF** the DNS server running on the laptop or desktop knew how to do DNSSEC (but that would probably be painful).

VII.4 Windows DNS Client Support for DNSSEC

Quoting http://technet2.microsoft.com/WindowsServer/en/library/264820c4-55c7-42d6-9747-432af9556acc1033.mspx?mfr=true, Client support for DNSSEC, "The DNS client does not read and store a key for the trusted zone and, consequently, it does not perform any cryptography, authentication, or verification. When a resolver initiates a DNS query and the response contains DNSSEC resource records, programs running on the DNS client will return these records and cache them in the same manner as any other resource records. This is the extent to which Windows XP DNS clients support DNSSEC. When the DNS client receives the SIG RR relating to the RRset, it will not perform an additional query to obtain the associated KEY record or any other DNSSEC records."

Speaking of Client Layer Stuff, What Would a DNSSEC User See If a DNS Resource Record Failed DNSSEC Validation?

Answer: nothing. Users would see nothing that would indicate a DNSSEC validation failure had occurred. Such a failure is normally "silent" and indistinguishable (to the user) from many other types of DNS failures. It is strange to know that DNSSEC validation failures are opaque to users. Instinctively, we know that DNSSEC validation might fail due to:

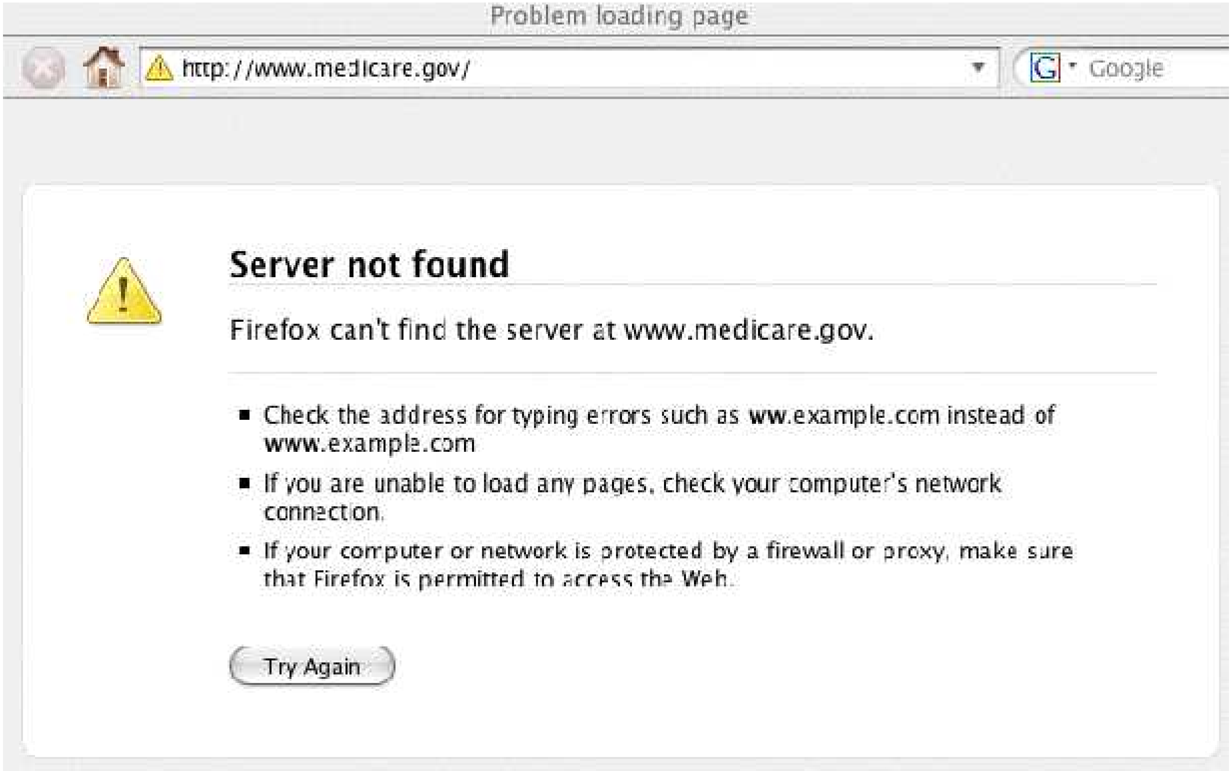
- operational error: it would be good to make sure that's noticed and corrected, and users could act as "canaries in the coal mine"
- an active attack; it would be REALLY good to know that's happening!
- something completely unrelated to DNSSEC might be busted

Silent failure modes that confound several possible issues is just a bad idea.

What Would a DNSSEC User See If a DNS Admin Screws Up Signing DNSSEC Signing Their Zone?

The zone wouldn’t resolve. Thus web pages under that zone would be inaccessible to user doing DNSSEC, although users who AREN’T doing DNSSEC would still be just fine.

Example: try to access www.medicare.gov on 4/10/2010 :



Server not found

Firefox can't find the server at www.medicare.gov.

- Check the address for typing errors such as ww.example.com instead of www.example.com
- If you are unable to load any pages, check your computer's network connection.
- If your computer or network is protected by a firewall or proxy, make sure that Firefox is permitted to access the Web.

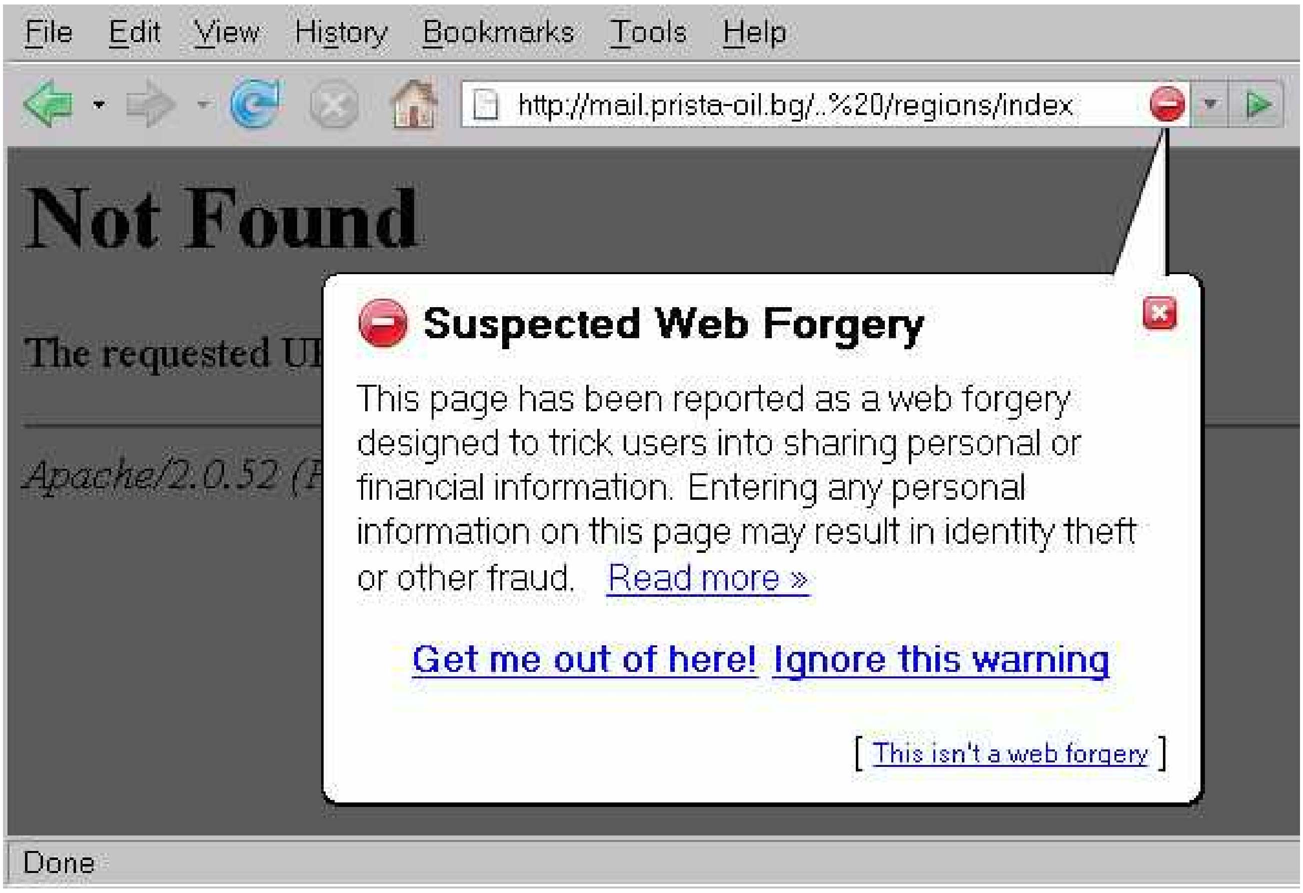
VII.5 DNSSEC and Application Layer Visibility

DNSSEC **needs** application layer visibility for all the times when it works, kin to the little padlock icon for SSL encrypted secure web sessions (or certificate failure notices for when things are self signed, expired, or otherwise not trustworthy).

If DNSSEC similarly "just works" (except for when it silently breaks attempts to do bad things, or someone screws up and it breaks attempts to do legitimate things), will people even know they're using DNSSEC?

Contrast invisible DNSSEC protection with the anti-phishing protection that Firefox delivers, something that's FAR more "in our face" and visible…

What A Firefox User Sees When Attempting to Visit A Phishing Site



Not Found

The requested UI

Apache/2.0.52 (F

Suspected Web Forgery

This page has been reported as a web forgery designed to trick users into sharing personal or financial information. Entering any personal information on this page may result in identity theft or other fraud.

VII.6 Another Issue: The DNSSEC Trust Model

Trust models focus on the question of, "Why should I believe you're really you?" "Why should I accept 'your' credentials as being authentic?" This is a pivotal question in cryptography.

Some crypto protocols, such as GPG/PGP, are decentralized, and employ a "web-of-trust" trust model where we trust their public key because it has been signed by other keys which we recognize/trust.

Other crypto protocols, such as PKI, are more centralized or "top down." In the PKI model, we trust a particular PKI certificate because it has been signed by a trusted certificate authority ("CA").

DNSSEC was originally intended to use a centralized top-down trust model, with a signed root. The trusted signed root would then sign immediately subordinate TLDs (top level domains); those TLDs would sign second level domains immediately below them, etc.

One slight problem: the root still hasn't been fully signed.

What About The TLDs? Are The TLDs At Signed and Supporting DNSSEC?

Signed TLD domains include .arpa (the in-addrs), .bg (Bulgaria), .br (Brasil), .cz (Czech Republic), .gov, .li (Liechtenstein), .na (Namibia), .nu (Niue), .org, .pr (Puerto Rico), .se (Sweden), .th (Thailand), .tm (Turkmenistan), .uk (the United Kingdom) and .us

There are also trust anchors for a number of IDN'd TLDs.

Most other TLDs (including .edu, .com, .net, .info, .mil, .biz, .int, .ca, .cn, .de, .fr, .jp, etc.) are still NOT signed at this time.

This does not prevent domains under those TLDs from doing DNSSEC, but when a domain under one of those TLDs does do DNSSEC, they exist as an "island of trust."

VII.7 Islands Of Trust

Remember, DNSSEC was designed to work using a **centralized, top-down trust model**. If the root isn't signed, or the TLD above them isn't signed, all the stuff below that point must establish **alternative trust anchors**. In some cases (such as .se), the trust anchor may be the TLD, but in other cases, the trust anchor may be 2nd-level domain (such as nanog.org).

If there is **no central trust anchor**, unless we can come up with an alternative way of establishing a chain of trust, **we must obtain trustworthy keys for each of those individual islands of trust**. (Key management is the 2nd thing, after trust models, to always scrutinize when considering about a crypto effort!)

If each site that wants to do DNSSEC has to do a "scavenger hunt" for each island of trust's DNSSEC keys, that's **rather inconvenient** particularly if (1) trust islands periodically **rekey**, (2) there are **thousands** of domains, and (3) given that if a site **fails** to keep each trust island's keys current, then that zone will "do a medicare.gov"

VII.8 DLV (Domain Lookaside Validation)

To avoid these problems, ISC has proposed DLV (Domain Lookaside Validation) as a temporary/transitional model.

In the DLV model, even if the root or a TLD isn't ready to support DNSSEC and sign its zone, perhaps a trusted third party can collect, authenticate and deliver the required keys. Someone attempting to do DNSSEC then has only to configure the DLV server or servers as an anchor of trust, thereafter automatically trusting domains that are anchored/validated via the DLV.

DLV is described at http://www.isc.org/solutions/dlv

DLV is supported in current versions of BIND.

DLV is the most popular approach to dealing with the problem of maintaining trust anchors until the root and TLDs are signed.

VII.9 The Zone Enumeration Issue And NSEC3

As originally fielded, DNSSEC made it possible to exhaustively enumerate, or "walk," a zone, discovering all known hosts.

Zone enumeration give miscreants a real "boost up" when it comes to reconnoitering a domain, and this was a real problem for some TLDs in countries with strong privacy protections.

NSEC3 as defined by RFC5155, addresses the zone enumeration issue through use of salted hashes, which handles both the zone enumeration concern as well as the problem that "the cost to cryptographically secure delegations to unsigned zones is high for large delegation-centric zones and zones where insecure delegations will be updated rapidly."

For our purposes, it is sufficient to know that NSEC3 effectively eliminates the zone enumeration problem.

Are Name Servers (the Software Programs) DNSSEC-Ready?

Another potential stumbling block might be the name server software. If the name server software we use doesn't support DNSSEC, our ability to do DNSSEC will obviously be limited.

First, what name server products do people run?

BIND Dominates The DNS Server Market

http://dns.measurement-factory.com/surveys/200910.html

Using dataset II, authoritative 2nd level com/net/org servers:

Recent **BIND 9**	173,590	69.23%
Other versions of **BIND**	11,583	4.62%
		73.85% total

Using dataset I, nameservers found on random IPv4 addresses:

Recent **BIND 9**	235,358	31.44%
Other versions of **BIND**	16,828	2.26%
		33.70% total

Current Versions of BIND Support DNSSEC

The good news for people interested in deploying DNSSEC is that the current version of BIND supports DNSSEC, and BIND has the lion's share of the current DNS server market, as shown by the table.

VII.10 Microsoft's DNS Servers and DNSSEC

Quoting technet2.microsoft.com/WindowsServer/en/library/264820c4-55c7-42d6-9747-432af9556acc1033.mspx?mfr=true (updated January 31st, 2005): "Windows Server 2003 DNS provides basic support of the DNS Security Extensions (DNSSEC) protocol as defined in RFC 2535." *[however, note that RFC2535 dated March 1999, was made obsolete by RFC4033, RFC4034, and RFC4035 ca. March 2005]* "**The current feature support allows DNS servers to perform as secondary DNS servers for existing DNSSEC-compliant, secure zones.** DNS supports the storing and loading of the DNSSECspecific resource records (RRs). **Currently, a DNS server is not capable of signing zones and resource records (creating cryptographic digital signatures) or validating the SIG RRs.** The DNSSEC resource records are KEY, SIG, and NXT." [the March 2005 RFC's deprecated those earlier DNSSEC record types]

DNSSEC and Windows Server 2008 R2

The situation is less dire for Windows Server 2008 R2. Windows Server 2008 R2 now provides at least basic DNSSEC support, and an 87 page guide to Windows DNSSEC deployment guide released in October 2009 is now available from Microsoft (see http://tinyurl.com/windows-and-dnssec).

The current Microsoft DNSSEC implementation has not exactly won universal acclimation, unfortunately. See, for example, “DNSSEC: Will Microsoft Have Enough Time?”, Jan 29, 2010, www.circleid.com/posts/dnssec_will_microsoft_have_enough_time/

See also “NIST SP 800-81r1 Checklist Items and Microsoft Windows Server 2008 R2,” http://www.dnsops.gov/vendors/MS-Win2008R2-SP800-81r1-Checklist.pdf (note, for example, that NSEC3 support is still lacking)

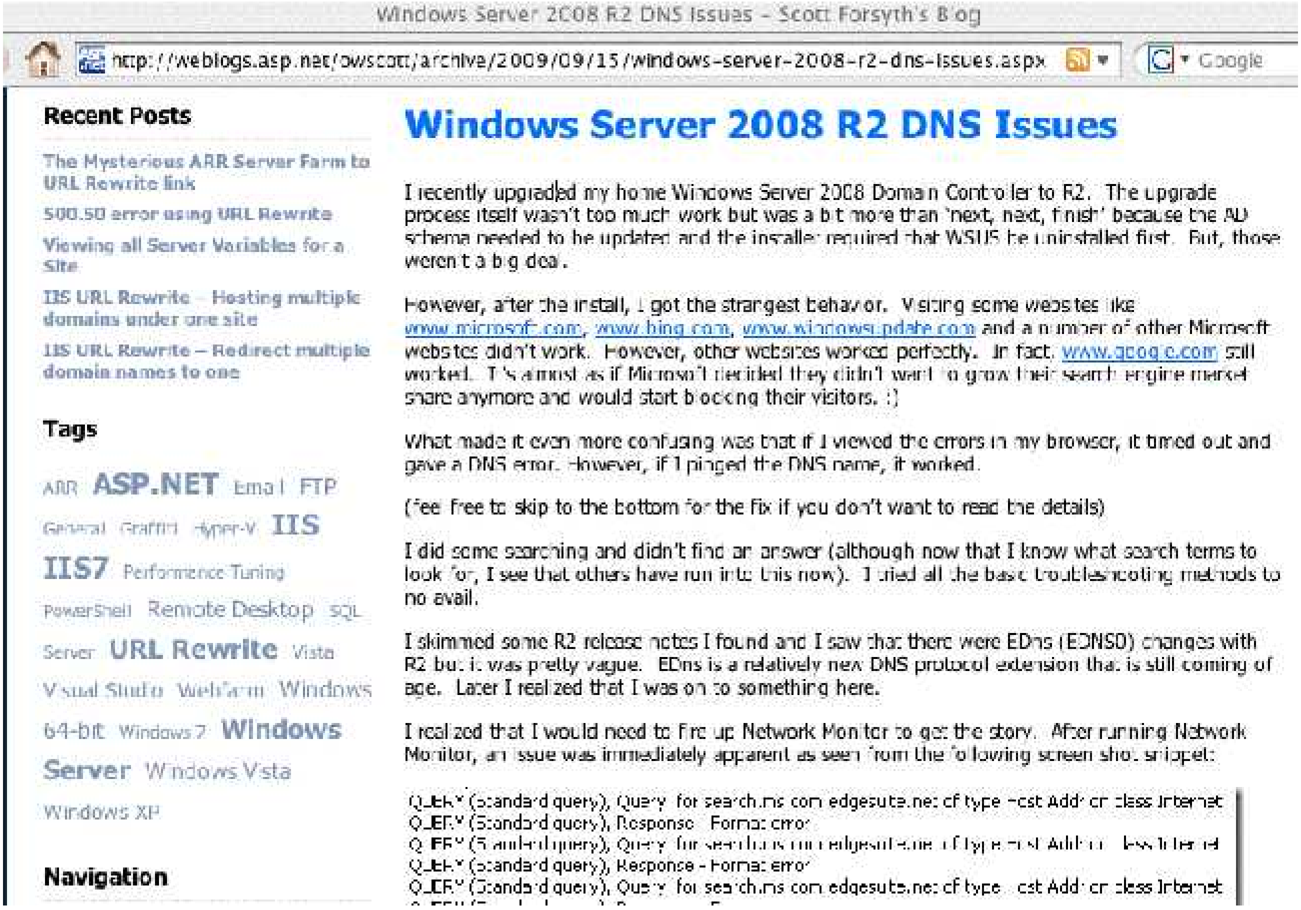
VII.11 EDNS0

We should know that name servers doing DNSSEC requires a feature known as EDNS0, as defined in RFC2671, "Extension Mechanisms for DNS (EDNS0)," August 1999.

Normally, DNS UDP responses are limited to just 512 bytes, a size that's too small for the much larger DNSSEC records. To better handle delivery of DNSSEC records, EDNS0 allows clients and servers to negotiate the maximum size datagram which they can handle, with the expectation that at least some hosts might negotiate datagram sizes as high as 4KB. Name servers doing DNSSEC **must** also do EDNS0.

Why's that a problem? Well… some firewalls may be configured to block UDP DNS traffic > 512 bytes. If we've got a firewall in front of our DNS server, we have to test to see if we're broken: https://www.dns-oarc.net/oarc/services/replysizetest

EDNS0 In Some MS Windows Environments



Windows Server 2008 R2 DNS Issues - Scott Forsyth's Blog

http://weblogs.asp.net/owscott/archive/2009/09/15/windows-server-2008-r2-dns-issues.aspx

Recent Posts

The Mysterious ARR Server Farm to URL Rewrite link

500.50 error using URL Rewrite

Viewing all Server Variables for a Site

IIS URL Rewrite – Hosting multiple domains under one site

IIS URL Rewrite – Redirect multiple domain names to one

Tags

ARR ASP.NET Email FTP General Graffiti Hyper-V IIS IIS7 Performance Tuning PowerShell Remote Desktop SQL Server URL Rewrite Vista Visual Studio Web Farm Windows 64-bit Windows 7 Windows Server Windows Vista Windows XP

Navigation

Windows Server 2008 R2 DNS Issues

I recently upgraded my home Windows Server 2008 Domain Controller to R2. The upgrade process itself wasn't too much work but was a bit more than 'next, next, finish' because the AD schema needed to be updated and the installer required that WSUS be uninstalled first. But, those weren't a big deal.

However, after the install, I got the strangest behavior. Visiting some websites like www.microsoft.com, www.bing.com, www.windowsupdate.com and a number of other Microsoft websites didn't work. However, other websites worked perfectly. In fact, www.google.com still worked. It's almost as if Microsoft decided they didn't want to grow their search engine market share anymore and would start blocking their visitors. :)

What made it even more confusing was that if I viewed the errors in my browser, it timed out and gave a DNS error. However, if I pinged the DNS name, it worked.

(feel free to skip to the bottom for the fix if you don't want to read the details)

I did some searching and didn't find an answer (although now that I know what search terms to look for, I see that others have run into this now). I tried all the basic troubleshooting methods to no avail.

I skimmed some R2 release notes I found and I saw that there were EDns (EDNS0) changes with R2 but it was pretty vague. EDns is a relatively new DNS protocol extension that is still coming of age. Later I realized that I was on to something here.

I realized that I would need to fire up Network Monitor to get the story. After running Network Monitor, an issue was immediately apparent as seen from the following screen shot snippet:

```
QUERY (standard query), Query for search.ms com edgesuite.net of type Host Addr on class Internet
QUERY (Standard query), Response - Format error
QUERY (Standard query), Query for search.ms com edgesuite.net of type Host Addr on class Internet
QUERY (standard query), Response - Format error
QUERY (Standard query), Query for search.ms com edgesuite.net of type Host Addr on class Internet
```

CONCLUSIONS

Why Aren't People Currently Using DNSSEC?

Do people simply not know DNSSEC exists? Well at least that's no longer an excuse for the IT specialists.

Are people willing to try DNSSEC, but simply don't know the "recipe" to get going? If so, there are (not only) three resources:

- Olaf Kolkman/NLNet Lab's "DNSSEC HOWTO, a tutorial in disguise," see http://www.nlnetlabs.nl/dnssec_howto/

- Geoff Huston's three part exploration of DNSSEC: http://www.potaroo.net/ispcol/2006-08/dnssec.html, http://www.potaroo.net/ispcol/2006-09/dnssec2.html, http://www.potaroo.net/ispcol/2006-10/dnssec3.html and
- The RIPE NCC's DNSSEC Training Course: http://www.ripe.net/training/dnssec/material/dnssec.pdf

Are people waiting for the root zone (or major TLDs) to be signed? Or are people waiting for more of their peers to take the plunge and report back, first?

Or Are There More Fundamental Problems?

Are people just really busy, with slow uptake just the normal resistance to yet one more thing – *ANYTHING* MORE! – to handle without substantial additional resources?

Does DNSSEC solve what's perceived by the community to be a **"non-existent" or "unimportant" problem?**

Are there **critical administrative tools** missing? (if that's the issue, then see http://www.dnssec-tools.org/ and http://www.ripe.net/disi/dnssec_maint_tool/)

Are people waiting to see what the other people do DNSSEC?

Something to Note: DNSSEC Adoption Doesn't Need to Be Symmetric

When deploying DNSSEC (just as when deploying SPF or DK/DKIM for email), adoption doesn't need to be symmetric:

- we can sign our own zones with DNSSEC on our authoritative name servers, yet **not** check DNSSEC on our recursive customer-facing name servers, or
- we can check DNSSEC on our recursive customer-facing facing name servers, yet **not** publish DNSSEC records for our own domains on our authoritative name servers

Most sites will eventually want to "take the whole plunge" (or skip the technology entirely), but sometimes different people have decision making authority for different parts of the organization, and we should recognize that asymmetric adoption is a possibility.

REFERENCES

http://www.uoregon.edu/~joe/secprof10-dns/

http://www3.ietf.org/proceedings/06nov/slides/plenaryt-2.pdf

https://www.dns-oarc.net/

http://www.enum.org

http://www.ops.ietf.org/lists/namedroppers/namedroppers.2001/msg00247.html

http://www.isc.org/software/bind/faq

http://www.dnssec-deployment.org/

http://spoofer.csail.mit.edu/summary.php http://www.dnssec-tools.org/

http://www.ripe.net/disi/dnssec_maint_tool/

http://code.google.com/speed/public-dns/

http://www.cisco.com/en/US/docs/routers/10000/10008/configuration/guides/ancp/isbl4dt.pdf

http://support.microsoft.com/kb/316786

http://dnscheck.iis.se/

https://www.dns-oarc.net/oarc/services/dnsentropy

http://dns.measurement-factory.com/tools/dsc/

http://www.cymru.com/Documents/

http://technet2.microsoft.com/WindowsServer/en/library/

http://www.isc.org/solutions/dlv

http://dns.measurement-factory.com/surveys/

https://www.dns-oarc.net/oarc/services/replysizetest

NIST SP 800-81

THE SELF-HACK AUDIT

LT Eduard GHIU

1. Introduction

Hacker n. **1.** A person who enjoys exploring the details of programmable systems and how to stretch their capabilities, as opposed to most users, who prefer to learn only the minimum necessary. **2.** A malicious meddler who tries to discover sensitive information by poking around. Hence 'password hacker', 'network hacker', or 'cracker'.

One of the greatest fears of the system administrator is the thought of their network being compromised. There are many threats, constantly bombarding the defenses of computer networks. If an intruder has physical access to a machine, they will be able to remove or damage parts of the system. If a hacker already has a low-privilege user account on the system, and the latest security patches have not been applied, there is a good chance he will be able to use known exploits in order to gain additional privileges.

Finally, remote intrusion involves a hacker who attempts to penetrate a system across a network. He starts with no privileges and must gain entry by bypassing the network's defenses. In order to combat these threats, one must put oneself in the mind of the attacker, and assess their own vulnerabilities from that point of view. System administrators must not only be aware of the potential vulnerabilities inherent in their operating system and applications software, but they must know how to protect the network from these dangers, and they must be able to assess their defenses before a successful attack is carried out.

In today's electronic environment, the threat of being hacked is no longer an unlikely incident, occurring in a few unfortunate organizations. New reports of hacker incidents and compromised systems appear almost daily. As organizations continue to link their internal networks to the Internet, system managers and administrators are becoming increasingly aware of the need to secure their systems. Implementing basic password controls is no longer adequate to guard against unauthorized access to data. Organizations are now looking for more up-to-date techniques to assess and secure their systems. The most popular and practical technique emerging is the **Self-Hack Audit (SHA)**. The **SHA** is an approach that uses hacker methods to identify and eliminate security weaknesses in a network before they are discovered by a hacker.

This document provides a methodology for the **SHA** and presents a number of popular hacker techniques that have allowed hackers to penetrate various systems in the past. Each description is followed by a number of suggested system administration steps or precautions that should be followed to help prevent such attacks. Although some of the issues discussed are specific to UNIX systems, the concepts can be applied to all systems in general.

2. Objectives of the Self-Hack Audit

The basic objective of the **SHA** is to identify all potential control weaknesses that may allow unauthorized persons to gain access to the system. The network administrator must be familiar with and use all known hacker techniques for overcoming system security. Depending on the nature of the audit, the objective may be either to extend a user's current levels of access (which may be no access) or to destroy (i.e., sabotage) the system.

2.1. Overview of the Methodology

To perform a useful **SHA**, the different types of hackers must be identified and understood. The stereotype of a hacker as a brilliant computer science graduate sitting in a laboratory in a remote part of the world is a dangerous misconception. Although such hackers exist, the majority of security breaches are performed by staff members of the breached organization. Hackers can be categorized into four types:

- Persons within an organization who are authorized to access the system. An example may be a legitimate staff member in the Accounting department who has access to Accounts Payable application menu functions.
- Persons within an organization who are not authorized to access the system. These individuals may include personnel such as the cleaning staff.
- Persons outside an organization who are authorized to access the system. An example may be a remote system support person from the organization's software vendor.
- Persons outside an organization who are not authorized to access the system. An example is an Internet user in an overseas country who has no connection with the organization.

The objective of the **SHA** is to use any conceivable method to compromise system security. Each of the four hacker types must be considered to assess fully all potential security exposures.

3. Inside / Outside Attacks

It seems that the main focus of computer security is protection from the malicious outside hacker, but, in reality, your network is more likely to be compromised by people inside your organization. Even the best security policy is no match for an unhappy employee. According to the Computer Security Institute/FBI and Ernst & Young, nearly 50% of all network attacks come from the inside. In a NetVersant survey, 82% reported spotty or no compliance with their company’s security policies. 85% say a properly implemented firewall would still be vulnerable to a disgruntled employee. And 75% say the firewall is at risk because of employee incompetence. It is far beyond the scope of the system administrator’s responsibilities to keep employees happy, but there are several things one can do to prevent these inside hackers from causing network problems. You can start with the physical location of your servers. They should be located in a locked, air-conditioned room. Access to this room should be limited to those who actually need physical access to the servers. The main console may also be locked away in this room. The rules that you setup for your users can limit such things as which files they can access, what they can do with those files, the time window in which they can log onto the system, and which workstations they can use to logon. A system administrator must have auditing turned on, and must review these logs on a regular schedule that has been spelled out in the security policy.

Log filters can be put in place to weed-out unusual or suspect traffic, and can alert an administrator via email, pager, etc. Employees must be taught to lock their workstations before walking away from a terminal session. A partial fix to this problem is to set password protected screen savers to be invoked at a maximum of 15 minutes of idle time. This will still leave a window of opportunity for the would-be hacker, but it is better than the alternative. Essentially the system administrator must take on the role of security guard, periodically “rattling doorknobs”, and watching the logs to see who has been where and what they were doing.

4. Popular Hacker Techniques

The following sections describe the techniques most commonly used by hackers to gain access to various corporate systems. Each section discusses a hacker technique and proposes basic controls that can be implemented to help mitigate these risks. The network administrator should attempt each of these techniques and should tailor the procedures to suit the organization’s specific environment.

4.1. Protecting the Operating System

For all hosts, whether they are servers or clients, the first step toward security is hardening the operating system. The numerous holes and security exploits on Windows and UNIX systems are widely publicized on hacker websites and on the sites of those trying to improve network security. One of the most common types of these exploits is the buffer overflow. In a buffer overflow attack, the hacker would typically send more logon characters than the operating system is prepared to handle. These extra characters may be treated like executable code. When run, this code allows the hacker to gain access to the network. This is just one of the many methods used by attackers. By limiting the number of open ports and services on your system, it follows that you would be limiting the number of available exploits. So, one of the first things to do after your installation is to turn off or disable every network service that is not essential to your operation and disable any unnecessary open ports that are running. **Nmap** is a free utility, downloadable from the internet that you can use to scan your entire network for open ports. Keep in mind that Nmap may also be used by attackers to scan your network. You should also keep patches updated on your system. It is recommended once a month to download patches and keep abreast of the latest security news and alerts.

Know the enemy. One of the most common hacker profiles is a 13 year old “script kiddie”. The script kiddie hunts for vulnerabilities with no specific target in mind. They simply download a malicious script from the Internet and fire it off, in hopes of finding a victim. By keeping your patches up to date and knowing the inherent problems with your operating system and your application software, problems caused by script kiddies can usually be avoided. To test your network against script kiddies, after obtaining written permission to do so, download a recent script from any number of hacker web sites, and run it against your system.

Note: This is a very dangerous thing to do. Results are unpredictable, and your system could be damaged. Always have a good backup, and know how to use it.

Once your operating system has been hardened, patches have been installed, and application software has been installed and updated accordingly, you should make a backup of your entire system and lock it away in a safe place. If something catastrophic should happen to your network, you will always be able to restore to your original baseline. Backup and Restore procedures must define when a backup must be performed, who will perform it, and what method will be used. The disaster recovery plan should define who is responsible for the restore, and the step-by-step procedure for completing a thorough recovery. The disaster recovery plan should be tested periodically to ensure success.

4.2. Accessing the Log-In Prompt

One method of gaining illegal access to a computer system is through the log-in prompt. This situation may occur when the hacker is physically within the facility or is attempting to access the system through a dial-in connection.

An important step in securing corporate information systems is to ensure that physical access to computer resources is adequately restricted. Any internal or external person who gains physical access to a terminal is given the opportunity to attempt to sign on at the log-in prompt.

To reduce the potential for unauthorized system access by way of a terminal within the organization’s facility, the network administrator should ensure that:

- Terminals are located in physically secure environments.
- Appropriate access control devices are installed on all doors and windows that may be used to access areas where computer hardware is located.
- Personal computers that are connected to networks are password-protected if they are located in unrestricted areas. A hacker trying to access the system would be required to guess a legitimate password before gaining access through the log-in prompt.
- Users do not write their passwords on or near their work areas.

4.3. Obtaining Passwords

Once the hacker has gained access to an organization's log-in prompt, he or she can attempt to sign on to the system. This procedure requires a valid user ID and password combination.

4.3.1. Choosing good passwords

A good password is one that is kept secret and is difficult for someone to guess or crack through traditional brute force methods. So the first rule is "Keep it a secret"! Do not disclose your password to anyone. Many people routinely write down their passwords on a sticky note and put it on their monitor or under their keyboard. This makes it extremely easy for any passerby to obtain access and use their account. If a person must write down a password, it should be kept in a secure, locked place. This should be spelled out in your security policy and can be audited by randomly making a quick sweep of a person's office or cubicle, looking for written passwords.

The second rule is to choose a good password. Passwords should not be comprised of a person's name, their pets name, their birthday, social security number, or anything else that could be easily discovered and used to gain access. There are many parameters that can be set to ensure good passwords. Free applications, such as **npasswd** or **yppapasswd**, or commercial applications such as **PowerPassword** by **Symark**, (http://www.symark.com/) can do automatic password checking every time a password is changed. A good password should have at least 6 characters. More characters increase the difficulty of cracking or guessing the password. A good password will also contain a combination of upper and lower case characters, numbers, and special characters like *, &, $, #, etc.

4.3.2. Brute Force Attacks

Brute force attacks involve manual or automated attempts to guess valid passwords. A simple password guessing program can be written in approximately 60 lines of C code or 40 lines of PERL. Many password guessing programs are available on the Internet. A dictionary attack simply uses all of the words in a dictionary to guess a password, including hybrid attacks that check for backwards spelling and using numbers to replace certain letters, such as using 1's for l's in the word "hello" (i.e. "he11o"). The brute force attack is a guaranteed method of cracking a password, but it may take more time than is practical for the hacker. The brute force attack tries every character in every position of a password, until it gets the correct combination. The longer and more complicated passwords may take years for a single pc to crack.

Most hackers have a "password hit list," which is a collection of default passwords automatically assigned to various system accounts whenever they are installed. For example, the default password for the guest account in most UNIX systems is "guest."

To audit the passwords that your users have chosen, you can use cracking tools like **L0phtCrack** or **Crack**. These tools attempt to guess passwords using dictionary attacks and brute force attacks. Before running any of these tools on your network, get written permission from anyone in management that may be concerned with this, and give your users at least a weeks notice before your test.

To protect the network from unauthorized access, the network administrator should ensure that:

- All user accounts are password protected.
- Password values are appropriately selected to avoid guessing.
- Default passwords are changed once the system is installed.
- Failed log-in attempts are logged and followed up appropriately.
- User accounts are locked out after a predefined number of sign-on failures.
- Users are forced to select passwords that are difficult to guess.
- Users are forced to change their passwords periodically throughout the year.
- Unused user accounts are disabled.
- Users are educated and reminded regularly about the importance of proper password management and selection.

4.3.3. Password Cracking

Most UNIX sites store encrypted passwords together with corresponding user accounts in a file called /etc/passwd. Should a hacker gain access to this file, he or she can simply run a password cracking program such as **Crack**. **Crack** works by encrypting a standard dictionary with the same encryption algorithm used by UNIX systems (called **crypt**). It then compares each encrypted dictionary word against the entries in the password file until it finds a match. **Crack** is freely available via an anonymous FTP from **ftp.cert.org at /pub/tools/crack**.

To combat the hacker's use of password-cracking software, the network administrator should ensure that:

- Encrypted passwords are stored in a shadow password file and that the file is adequately protected.
- All "weak" passwords are identified by running **Crack** against the password file.
- Software such as **Npasswd** or **Passwd+** is used to force users to select passwords that are difficult to guess.
- Users do not write their passwords on or near their work environments.
- Only the minimum number of users have access to the command line to minimize the risk of copying the /etc/passwd file.

4.3.4. Keystroke Logging

It takes less than 30 seconds to type in a short script to capture sign-on sessions. A hacker can use a diskette to install a keystroke-logging program onto a workstation. Once this Trojan horse is installed, it works in the background and captures every sign-on session, based on trigger key words. The hacker can read the captured keystrokes from a remote location and gain access to the system. This technique is very simple and almost always goes unnoticed.

To prevent a hacker's access to the system by way of a keystroke-logging program, the network administrator should ensure that:

- Privileged accounts (e.g., root) require one-time passwords.
- The host file system and individual users' workstations are periodically scanned for Trojan horses that could include keystroke-logging programs.
- Adequate physical access restrictions to computer hardware are in place to prevent persons from loading Trojan horses.

4.3.5. Packet Sniffing

The Internet offers a wide range of network monitoring tools, including network analyzers and "packet sniffers." These tools work by capturing packets of data as they are transmitted along a communications segment. Once a hacker gains physical access to a PC connected to a LAN and loads this software, he or she is able to monitor data as it is transferred between locations. Alternatively, the hacker can attach a laptop to a network port in the office and capture data packets.

Remembering that network traffic often is not encrypted, there is a high chance that the hacker will capture valid user account and password combinations, especially between the hour of 8:00 a.m. and 9:00 a.m. **Tcpdump** is a tool for UNIX systems used to monitor network traffic and is freely available via an anonymous FTP from **ftp.ee.lbl.gov**.

To reduce the possibility of account and password leaks through packet sniffers, the network administrator should ensure that:

- Communications lines are segmented as much as practical.
- Sign-on sessions and other sensitive data are transmitted in an encrypted format by using software such as **Kerberos**.
- Privileged accounts (e.g., root) sign on using one-time passwords.
- Physical access to communications lines and computer hardware is restricted.

4.3.6. Perimeter Security

The surest way to keep hackers from stealing passwords and wreaking havoc on your network is to build a wall that won't allow them to get close enough to touch your system. Typically the first line of defense against outside penetration is the firewall. The firewall acts as a single point of access, where all traffic coming into a network can be audited, authorized and authenticated. Based on the rules used to configure it, a firewall can block any suspicious activity. Common types of firewalls are:

- Routers, which simply look at a packet and decide whether or not its destination is inside the network;
- Packet filters, which examine the source and destination of an IP packet, as well as the source and destination TCP/UDP ports, and accept or reject the packet based on user-defined rules;
- Stateful packet systems, which are similar to packet filters, but they actually examine the contents of a packet to determine if it should accept or reject it, based on the rules defined by the user;
- Application proxies, which force all network traffic to be examined before they decide which data is to be passed on, and which to drop.

When creating the rule sets for firewalls, one must consider the effect certain limitations will have on the performance of the network. The more stringent the rules are, the more processing power is required. There are several methods used to evaluate the effectiveness of your firewall. Auditing tools can log all traffic that hits the firewall, and these logs may be sorted by many variables, such as: what traffic was allowed to pass, what traffic was rejected, where the traffic originated, and what is its destination. The audit logs can provide important forensic evidence. After an attack, all of the traffic related to the attack can be analyzed to give the administrator a better understanding of how the attack was carried out, and how to protect the network against future attacks. Intrusion detection is a type of network management tool that gathers information from various places in a network to identify possible security breaches. These breaches include intrusions (attacks from outside the organization) and misuse (attacks from within an organization).

Intrusion detection functions include:

- Monitoring and analyzing both user and system activities
- Analyzing system configurations and vulnerabilities
- Assessing system and file integrity
- Ability to recognize patterns typical of attacks
- Tracking user policy violations10

One of the most popular Network Intrusion Detection Systems is **Snort**. **Snort** is an open source application that is capable of performing real-time traffic analysis and packet logging on IP networks.

4.3.7. Social Engineering

Social Engineering is hacker-speak for tricking a person into revealing their password. This is the most common and easiest way for an attacker to gain access to a network. There are several ways to do this. A hacker may send a bogus email message claiming to be the system administrator needing your password for some administrative task. Another method is "shoulder surfing", simply looking over someone's shoulder as they type in their password. Another problem is sharing accounts. This should never happen. If a person needs to be on the system, they should have their own account. It only takes one of these mistakes to bypass all of the password policies, firewalls, and whatever else you have implemented for your perimeter security. To fight social engineering, you must educate your users and frequently test them to ensure awareness. Send your own emails and make your own phone calls asking for passwords. If someone makes a mistake, let him or her know how serious it is, and, chances are, it will not happen again.

Hackers often select a user account that has not been used for a period of time (typically about two weeks) and ensure that it belongs to a user whom the administrator is not likely to recognize by voice. Hackers typically target accounts that belong to interstate users or users in another building. Once they have chosen a target, they assume a user's identity and call the administrator or the help desk, explaining that they have forgotten their passwords. In most cases, the administrator or help desk will reset passwords for the hackers over the telephone.

In an effort to keep the network safe from this type of infiltration, the network administrator should ensure that:

- All staff are regularly reminded and educated about the importance of data security and about proper password management.

- The organization has documented and controlled procedures for resetting passwords over the telephone.
- Staff do not fall prey to social engineering attacks. Staff members must be aware of the possibility that a hacker may misrepresent himself or herself as a member of the information systems department and ask for a password.

4.4. General Access Methods

4.4.1. Internet Protocol Address Spoofing

In a typical network, a host allows other "trusted" hosts to communicate with it without requiring authentication (i.e., without requiring a user account and password combination). Hosts are identified as trusted by configuring files such as the **.rhost** and **/etc/hosts.equiv** files. Any host other than those defined as trusted must provide authentication before being allowed to establish communication links.

Internet protocol (IP) spoofing involves an untrusted host connecting to the network and pretending to be a trusted host. This access is achieved by the hacker changing his IP number to that of a trusted host. In other words, the intruding host fools the host on the local network into not challenging it for authentication.

To avoid this type of security violation, the network administrator should ensure that:

- Firewalls and routers are appropriately configured so that they reject IP spoofing attacks.
- Only appropriate hosts are defined as trusted within **/etc/hosts.equiv**, and file permissions over this file are adequate.
- Only appropriate hosts are defined within users' **/.rhost** files. If practical, these files should be removed.

4.4.2. Unattended Terminals

It is quite common to find user terminals left signed on and unattended for extended periods of time, such as during lunch time. Assuming that the hacker can gain physical access to users' work areas (or assuming that the hacker is an insider), this situation is a perfect opportunity for a hacker to compromise the system's security. A hacker may use an unattended terminal to process unauthorized transactions, insert a Trojan horse, download a destructive virus, modify the user's .**rhost** file, or change the user's password so that the hacker can sign on later.

The network administrator can minimize the threat from access through unattended terminals by ensuring that:

- User sessions are automatically timed out after a predefined period of inactivity, or password-protected screen savers are invoked.

- Users are regularly educated and reminded about the importance of signing off their sessions whenever they expect to leave their work areas unattended.
- Adequate controls are in place to prevent unauthorized persons from gaining physical access to users' work areas.

4.4.3. Writeable Set User ID Files

UNIX allows executable files to be granted root privileges by making file permissions set user ID (SUID) root. Hackers often search through the file system to identify all SUID files and to determine whether they are writeable. Should they be writeable, the hacker can insert a simple line of code within the SUID program so that the next time it is executed it will write to the **/etc/passwd** file and this will enable the hacker to gain root privileges. The following UNIX command will search for SUID root files throughout the entire file system:

find /-user root -perm -4000 -print.

The network administrator can reduce the possibility of illegal access through SUID files by ensuring that:

- Only a minimum number of programs are assigned SUID file permissions.
- Programs that are SUID are not writeable by users other than root.
- Executables defined within the system **cron** tables (especially the root **cron** table) are not writeable by users other than root because they are effectively SUID root.

4.4.4. Computer Emergency Response Team Advisories

The Computer Emergency Response Team (**CERT**) issues advisories whenever a new security exposure has been identified. These exposures often allow unauthorized users to gain root access to systems. Hackers always keep abreast of the latest CERT advisories to identify newly found bugs in system software. CERT can be accessed via an anonymous FTP at **info.cert.org.**

The network administrator should ensure that:

- All CERT advisories have been reviewed and acted on in a controlled and timely manner.
- Checksums are used to ensure the integrity of CERT patches before they are implemented.

4.4.5. Hacker Bulletin Boards

The Internet has a large number of hacker bulletin boards and forums that act as an invaluable source of system security information. The most popular hacker bulletin board is the **"2600" discussion group**. Hackers from around the world exchange security information relating to various systems and often publish security-sensitive information relating to specific organizations or hacker techniques relating to specific programs.

The network administrator should ensure that the organization's data security officer regularly reviews hacker bulletin boards to identify new techniques and information that may be relevant to the organization's system environment.

4.4.6. Internet Software

The Internet offers a large number of useful tools, such as **SATAN**, **COPS**, and **ISS**, which can assist data security officers and administrators in securing computer resources. These tools scan corporate systems to identify security exposures. However, these tools are also available to hackers and can assist them in penetrating systems.

To identify and resolve potential security problems, the network administrator should ensure that:

- The latest version of each security program is obtained and run in a regular manner. Each identified exposure should be promptly resolved.
- The system is subject to regular security audits by both the data security officer and independent external consultants.

5. Summary

No network is 100% secure. A system administrator must determine how secure the network must be, and implement a security policy that conforms to that philosophy. One thing that will play a major role in the security of a network is the monetary cost versus the cost of a security breach. Is the threat from a malicious source dangerous enough to justify the cost of protecting against an intrusion? One must continuously assess the network's vulnerabilities and keep abreast of the latest threats. I believe employee education plays a major role in network security. If your employees don't know what is expected of them, and they are unaware of the consequences of their actions, what motivation could they have to do their parts in keeping the network secure?

Anti-virus software must be kept up-to-date, **operating systems and applications must be patched**, **passwords must be checked**, **penetration scans must be performed**, and **backups must be done**. The list goes on and on. When all of this has been done, and no holes have been found, you, as a system administrator, can finally relax… then start from the top and do it all over again.

Hacker activity is a real and ongoing threat that will continue to increase as businesses connect their internal corporate networks to the Internet. This chapter has described the most common hacker techniques that have allowed unauthorized persons to gain access to computer resources. The Self-Hack Audit is becoming an increasingly critical technique for identifying security weaknesses that, if not detected and resolved in a timely manner, could allow hackers to penetrate the corporate system. System administrators and data security officers should keep abreast of the latest hacker techniques by regularly reading all CERT publications and hacker bulletin boards.

References

1. http://www.tuxedo.org/~esr/jargon/html/entry/hacker.html
2. http://www.ccert.edu.cn/education/cissp/hism/119-123.html
3. http://www.sans.org/reading_room/whitepapers/basics/basic-self-assessment-hack_467
3. http://www.robertgraham.com/pubs/network-intrusion-detection.html
4. http://networks.depaul.edu/security/winnt.htm
5. http://www.tuxedo.org/~esr/jargon/html/entry/script-kiddies.html
6. http://networks.depaul.edu/security/passwords.htm
7. http://www.seas.rochester.edu:8080/CNG/docs/Security/node9.html
8. http://www.snort.org/about.html

ENCRYPTION ALGORITHMS OVERVIEW

CPT Tiberiu MOLDOVAN

INTRODUCTION

Encryption is the process of transforming information (referred to as plaintext) using a mathematical algorithm (called cipher) to make it unreadable to anyone except those possessing special knowledge, usually referred to as a key. The result of the process is encrypted information (in cryptography, referred to as ciphertext). In many contexts, the word encryption also implicitly refers to the reverse process, decryption (e.g. “software for encryption” can typically also perform decryption), to make the encrypted information readable again (i.e. to make it unencrypted).

In short we could define:

- Cryptology - The art and science of making and breaking “secret codes”
- Cryptography - making “secret codes”
- Cryptanalysis - breaking “secret codes”
- Symmetric key cryptosystem uses the same key to encrypt as to decrypt
- Asymmetric key cryptosystem uses a public key to encrypt and a private key to decrypt

The encryption “key” is the password used to encrypt or decrypt (lock or unlock) the secure data. Key size (alternatively key length) is the size of the digits used to create an encrypted text; it is therefore also a measure of the number of possible keys which can be used in a cipher, and the number of keys which must be tested to 'break' the encryption if no faster means is available.

In an ideal encryption system, the key length is therefore a measure of how secure the data is, and the effort and time needed to decrypt it by force. The length of a key is therefore critical in determining the susceptibility of a cipher to exhaustive search attacks.

Because modern cryptography uses binary keys, the length is usually specified in bits. The time and effort needed to break a cipher of a given key size varies according to the cipher; therefore a 128 bit key size in one system may be deemed equivalent in security to a 1024 bit key size in another. For example, on average, a typical home computer can crack a 40-bit key in a little less than two weeks. However, a 128-bit key could potentially take years, depending on the cipher. More complex keys take longer for the computer to process during encryption and decryption, effectively limiting practical key length.

HISTOY OF CRIPTOGRAPHY

The history of cryptography can be broadly divided into three phases:

(1) From ancient civilizations to the nineteenth century and the first part of the twentieth century, with relatively simple algorithms that where designed and implemented by hand.

(2) Extensive use of encrypting electro-mechanical machines, around the period of the Second World War.

(3) Ever more pervasive use of computers, about in the last fifty years, supported by solid mathematical basis.

Encryption has been used to protect communications since ancient times, but only organizations and individuals with extraordinary need for confidentiality had bothered to exert the effort required to implement it. As an example in 1586 Mary Stuart, Queen of Scotland, was sentenced to death for having conspired against her cousin Elizabeth, Queen of England. That was possible because Sir Francis Walsingham, Secretary of State, proved that Mary had taken part in the conspiracy by deciphering her communications with Sir Babington. The secret messages were hidden inside beer barrels and were written making use of several symbols that substituted letters, words or phrases, and also some more symbols with no real meaning, just to confuse other people.

Al-Kindi was a pioneer in cryptanalysis and cryptology. He gave the first known recorded explanation of cryptanalysis in approx. 850 AD. Encryption, and successful attacks on it, played a vital role in World War II. Many of the encryption techniques developed then were closely-guarded secrets (Kahn). In the mid-1970s, with the introduction of the U.S. Data Encryption Standard and public key cryptography, strong encryption emerged from the preserve of secretive government agencies into the public domain. Cryptography has since become a renowned and respected scientific field.

TYPES OF ALGORITHMS

The essential concept underlying all automated and computer security application is cryptography; the two ways of going about this process are conventional (or symmetric) encryption and public key (or asymmetric) encryption.

Different encryption algorithms use proprietary methods of generating these keys and are therefore useful for different applications.

Encryption schemes are based on **block** or **stream** ciphers.

The type and length of the keys utilized depend upon the encryption algorithm and the amount of security needed.

Block cipher: these algorithms work on chunks of specific sized data along with a key resulting in blocks of cipher text.

Stream cipher: a symmetric encryption algorithm that processes the data a bit or a byte at a time with a key resulting in a randomized ciphertext or plaintext. Some commonly used stream cipher algorithms are RC4 and W7.

Here are some nitty gritty details about some of these encryption algorithms. Strong encryption is often discerned by the key length used by the algorithm.

RSA

In 1977, shortly after the idea of a public key system was proposed, three mathematicians, Ron Rivest, Adi Shamir and Len Adleman gave a concrete example of how such a method could be implemented. To honor them, the method was referred to as the RSA Scheme. The system uses a private and a public key. To start two large prime numbers are selected and then multiplied together; **n=p*q**. If we let **f(n) = (p-1) (q-1)**, and **e>1** such that **GCD(e, f(n))=1**. Here **e** will have a fairly large probability of being co-prime to **f(n)**, if **n** is large enough and **e** will be part of the encryption key. If we solve the Linear Diophantine equation; **ed congruent 1 (mod f(n))**, for **d**. The pair of integers **(e, n)** is the public key and **(d, n)** form the private key. Encryption of **M** can be accomplished by the following expression; **Me = qn + C** where **0<= C < n**. Decryption would be the inverse of the encryption and could be expressed as; **Cd congruent R (mod n)** where **0<= R < n**. RSA is the most popular method for public key encryption and digital signatures today.

DES/3DES

The Data Encryption Standard (DES) was developed and endorsed by the U.S. government in 1977 as an official standard and forms the basis not only for the Automatic Teller Machines (ATM) PIN authentication but a variant is also utilized in UNIX password encryption. DES is a block cipher with 64-bit block size that uses 56-bit keys. Due to recent advances in computer technology, some experts no longer consider DES secure against all attacks; since then Triple-DES (3DES) has emerged as a stronger method. Using standard DES encryption, Triple-DES encrypts data three times and uses a different key for at least one of the three passes giving it a cumulative key size of 112-168 bits.

BLOWFISH

The Blowfish is a fast encryption algorithm designed by Bruce Schneier. Bruce Schneier is well known as the president of Counterpane Systems, a security consulting firm, and the author of Applied Cryptography: Protocols, Algorithms, and Source Code.

The Blowfish encryption algorithm was specially designed to encrypt data on 32-bit microprocessors. Blowfish is significantly faster than DES and GOST when implemented on 32-bit microprocessors, such as the Pentium or Power PC.

The original Blowfish paper was presented at the First Fast Software Encryption workshop in Cambridge, UK (proceedings published by Springer-Verlag, Lecture Notes in Computer Science #809, 1994) and in the April 1994 issue of Dr. Dobbs Journal. In addition, "Blowfish--One Year Later" appeared in the September 1995 issue of Dr. Dobb's Journal.

Additional information about the Blowfish algorithm is also available on World-Wide-Web at: http://www.counterpane.com/blowfish.html

Blowfish is a symmetric block cipher just like DES or IDEA. It takes a variable-length key, from 32 to 448 bits, making it ideal for both domestic and exportable use. Bruce Schneier designed Blowfish in 1993 as a fast, free alternative to the then existing encryption algorithms. Since then Blowfish has been analyzed considerably, and is gaining acceptance as a strong encryption algorithm.

IDEA

International Data Encryption Algorithm (IDEA) is an algorithm that was developed by Dr. X. Lai and Prof. J. Massey in Switzerland in the early 1990s to replace the DES standard. It uses the same key for encryption and decryption, like DES operating on 8 bytes at a time. Unlike DES though it uses a 128 bit key. This key length makes it impossible to break by simply trying every key, and no other means of attack is known. It is a fast algorithm, and has also been implemented in hardware chipsets, making it even faster.

SEAL

Rogaway and Coppersmith designed the Software-optimized Encryption Algorithm (SEAL) in 1993. It is a Stream-Cipher, i.e., data to be encrypted is continuously encrypted. Stream Ciphers are much faster than block ciphers (Blowfish, IDEA, DES) but have a longer initialization phase during which a large set of tables is done using the Secure Hash Algorithm. SEAL uses a 160 bit key for encryption and is considered very safe.

RC4

RC4 is a cipher invented by Ron Rivest, co-inventor of the RSA Scheme. It is used in a number of commercial systems like Lotus Notes and Netscape. It is a cipher with a key size of up to 2048 bits (256 bytes), which on the brief examination given it over the past year or so seems to be a relatively fast and strong cypher. It creates a stream of random bytes and 'XORing' those bytes with the text. It is useful in situations in which a new key can be chosen for each message.

AES (Rijndael)

The algorithm was invented by Joan Daemen and Vincent Rijmen. The National Institute of Standards and Technology (http://www.nist.gov) has recently selected the algorithm as an Advanced Encryption Standard (AES).

The cipher has a variable block length and key length. Authors of the algorithm currently specify how to use keys with a length of 128, 192, or 256 bits to encrypt blocks with a length of 128 bits. To get more information on the algorithm, visit the Rijndael Home Page: http://www.esat.kuleuven.ac.be/~rijmen/rijndael/.

CAST

CAST-128 (described in RFC-2144 document http://www.faqs.org/rfcs/rfc2144.html) is a popular 64-bit block cipher allowing key sizes up to 128 bits. The name CAST stands for Carlisle Adams and Stafford Tavares, the inventors of CAST.

GOST 28147-89

The Government Standard of the USSR 28147-89, Cryptographic protection for Data Protection Systems, appears to have played the role in the former Soviet Union (not only in Russia) similar to that played by the US Data Encryption Standard (FIPS 46). When issued, GOST bore the minimal classification 'For Official Use,' but is now said to be widely available in software both in the former Soviet Union and elsewhere. The introduction to GOST 28147-89 contains an intriguing remark that the cryptographic transformation algorithm "does not put any limitations on the secrecy level of the protected information."

The GOST 28147-89 standard includes output feedback and cipher feedback modes of operation, both limited to 64-bit blocks, and a mode for producing message authentication codes. Additional information about the GOST 28147-89 algorithm is also available at the Jetico Web site: http://www.jetico.com/gost.htm

RC-6

RC6 block cipher was designed by Ron Rivest in collaboration with Matt Robshaw, Ray Sidney, and Yiqun Lisa Yin from RSA Laboratories. RSA's RC6 encryption algorithm was selected among the other finalists to become the new federal Advanced Encryption Standard (AES). Visit RSA Laboratories site (http://www.rsasecurity.com/rsalabs/node.asp?id=2512) to get more information on the algorithm.

Serpent

Serpent is a block cipher developed by Ross Anderson, Eli Biham and Lars Knudsen. Serpent can work with different combinations of key lengths. Serpent was also selected among other five finalists to become the new federal Advanced Encryption Standard (AES).
Additional information about the Serpent algorithm is also available on World-Wide-Web from: http://www.cl.cam.ac.uk/~rja14/serpent.html

Twofish

The Twofish encryption algorithm was designed by Bruce Schneier, John Kelsey, Chris Hall, Niels Ferguson, David Wagner and Doug Whiting.
Twofish is a symmetric block cipher; a single key is used for encryption and decryption. Twofish has a block size of 128 bits and accepts keys of any length up to 256 bits.
The National Institute of Standards and Technology (NIST) investigated Twofish as one of the candidates for the replacement of the DES encryption algorithm. As the authors of the algorithm state, "we have spent over one thousand hours cryptanalyzing Twofish, and have found no attacks that go anywhere near breaking the full 16-round version of the cipher".
Additional information about the Twofish algorithm is available also on the World-Wide-Web from: http://www.counterpane.com/twofish.html

TYPES OF ATTACKS

This section briefly overviews common terminology related to types of attacks.
These terms are used throughout the document.

- **Brute-force Attack**: Brute-force is the ultimate attack on a cipher, by which all possible keys are successively tested until the correct one is encountered. A brute-force attack cannot be avoided but it can be made infeasible.
- **Codebook Attacks**: Codebook attacks are attacks that take advantage of the property by which a given block of plaintext is always encrypted to the same block of ciphertext as

long as the same key is used. There are several types of codebook attacks. The most typical ones are using character occurrence probabilities in plaintext.

- **Differential Cryptanalysis**: Differential cryptanalysis is the attempt to find similarities between various cipher-texts that are derived from similar (but not identical) plaintexts. This similarity may assist in recovering the key.
- **Known Plaintext Attacks**: These are attacks in which the attacker knows the plaintext as well as the ciphertext of an encryption operation and attempts to recover the key.
- **Linear Cryptanalysis**: Linear cryptanalysis is the attempt to find linear dependency of high probability between the plaintext, the ciphertext and the key, by which the key may be retrieved.
- **Man-In-The-Middle Attack** (MIM, or MITM): A "man-in-the-middle" attack is an attack that is placed by an active attacker who can listen to the communication between two entities and can also change the contents of this communication. While performing this attack, the attacker pretends to be one of the parties in front of the other party.
- **Oracle Attack**: An Oracle attack is in attack during which the attacker can be assisted by a machine or user who will perform encryption or decryption for him at will. The attacker can use multiple encryptions and decryptions of data of his choice to recover the key.
- **Related-Key Cryptanalysis**: Related-key cryptanalysis refers to attacks based on encrypting plaintexts with various similar (but not identical) keys and analyzing the differences in output.

ALGORITHMS VULNERABILITIES

The following chapter presents the identified weaknesses for each of the evaluated algorithms.

SYMMETRIC BLOCK CIPHERS

The following section presents the weaknesses that were found in algorithms for *symmetric encryption*. A symmetric encryption algorithm is an algorithm by which decryption and encryption are performed using the same key. Such algorithms are often called "*Private key algorithms*". These algorithms, being noticeably faster, are used for bulk encryption.

DES (Data Encryption Standard)

The DES algorithm, other than its short key (which can be brute-forced quite easily), is known to be secure. Triple-DES was developed to overcome this limitation.

Chaum and Evertse demonstrated an attack, on six rounds of DES, in a time of 2^{54}. This attack cannot be applied on more than eight rounds (let alone sixteen) so is not alarming in practice.

Davies known plain-text attack, using S-box pairs, succeeded in cracking eight rounds of DES using 2^{40} known plaintexts and with 2^{40} operations work. This is not too alarming since it has almost no effect on the strength of the full sixteenround DES. Extension of this particular attack to sixteen rounds will take more than the entire codebook of plaintexts and is therefore impractical. Yet, an improvement to this attack could crack full-DES (sixteen rounds) with a time of 2^{50} for data collection plus a time of 2^{50} for the attack itself. There is a tradeoff between success rate, time and amount of required plaintexts. The abovementioned figure presents the best "deal" in these terms. Alternatively, the attack can be performed to the extent it reveals 24 bits of the 56-bit key using 2^{52} known plaintexts, with very little work (complexity). These results may be alarming for systems in which bulk data is encrypted with unchanged keys.

The DES algorithm suffers from *Simple Relations* in its keys. In DES, the simple relationship is of a complementary nature. This means that the complementary relationship between keys results in a complementary relationship between the resulting ciphertexts. This vulnerability reduces the algorithm strength by one bit. Other relationships are existent for some other specific keys as well.

With regards to weak keys, DES has at least four of them. When encrypting using one of these weak keys, all sixteen rounds will be using the same sub-keys, making the algorithm as strong as a single round. Therefore, use of these keys must be avoided. In addition to these four keys, there are twelve more weak keys by which two rounds are running using the same sub-keys. In addition to these weak keys, DES also has keys that are defined as weak[1] and keys that are defined as semi-weak[2]. All these keys should be avoided so as not to harm the strength of the implementation when using the algorithm.

The key schedule that DES uses is not one-way. This results in the attacker being able to recover most of the master-key by compromising the sub-keys of few rounds. This vulnerability is hardly practical since the round keys are not easily available. Yet, this feature does assist in optimizing differential attacks.

The DES algorithm is vulnerable to linear cryptanalysis attacks. By such an attack, the algorithm in its sixteen rounds can be broken using 2^{43} known plaintexts. This vulnerability raises a notable risk when encrypting bulk data that may be predictable with keys that are constant.

Eli Biham and Adi Shamir presented a differential attack, by which a key can be recovered in 2^{37} time using 2^{37} ciphertexts taken from a pool after encrypting 2^{47} chosen plaintexts, even if these

[1] A key is called "weak" if when using it the encryption function is similar to the decryption function

[2] A pair of keys is called "semi-weak" if for one key the encryption function acts as the decryption function of the other

ciphertexts were encrypted with various keys. This attack, although very interesting academically, is hard to mount in most circumstances.

DES has several modes of operation, most commonly used one being CBC. Yet, if such modes (other than ECB) are used, it must be verified that the IV (Initialization Vector) is either fixed or not transferred in the clear. Otherwise, the implementation is highly vulnerable in the existence of an active attacker[3].

The 3DES algorithm, at least theoretically, is vulnerable to linear and differential attacks.

Triple-DES (3DES), other than being slow, is vulnerable to a variant of a meet-inthe- middle attack together with differential related-key attack. We did not find accurate figures for the cost of such attacks.

RC2

RC2 is an algorithm for which little cryptanalysis is available. However, it is known to have two weaknesses.

First, RC2 is vulnerable to differential attacks. An implementation with r mixing rounds (including the accompanying mashing rounds) will require at most 2^{4r} chosen plaintexts for a differential cryptanalysis attack. Commonly, the RC2 runs with 16 mixing rounds, making this attack less feasible than it may seem.

Second, the algorithm is vulnerable to a differential related-key attack requiring only 2^{34} chosen plaintexts and one related-key query.

RC4

The RC4 algorithm was not reviewed publicly to the extent of the others. The main weakness in this algorithm is that due to a weak key-mixing phase, 1/256 of the keys belong to a class of weak keys. These keys are detectable. After detection of a key belonging to this class, it is fairly easy to reveal 16 bits of the key with a 13.8% probability. In any implementation of this algorithm, a test to assure these keys are not used must be performed.

RC5

RC5 was not extensively reviewed either. This algorithm is, however, known to suffer from several weak keys. For RC5 running with r rounds, each key has a probability of 2^{-10r} of being a weak key. This weakness is not highly risky in practice although the weak keys should be avoided in the implementation.

[3] An attacker who can manipulate the data.

One attack, demonstrated by the designers of RC6, can break both RC5 and RC6 when running with up to 15 rounds, requiring less time than needed to perform an exhaustive key search.

RC6

RC6 is considered to be a strong algorithm with a fast and easy hardware implementation. It was submitted as an AES candidate and reached the second AES evaluation round. The RC6 has the following (mostly impractical) documented vulnerabilities.

For RC6 with 15 rounds or less, running on input blocks of 128 bits, it has been shown that the resulting ciphertext could be distinguished from a random series of bits. One of the conditions for an encryption algorithm to be secure is that its output resembles a completely random series of bits. Several applications check for randomness of bit streams to indicate strong encryption. Moreover, the writers of the algorithm have shown an attack against RC6 running with up to 15 rounds that is faster than an exhaustive key search. For one class of weak keys, it was shown that full randomness is not accomplished for up to 17 rounds of the algorithm.

For RC6 with 16 rounds, a linear cryptanalysis attack is possible, but requires 2^{119} known plaintexts, which makes this attack quite infeasible.

The RC6 algorithm is robust against differential cryptanalysis, provided that it applies more than 12 rounds.

CMEA (Cellular Message Encryption Algorithm)

The CMEA algorithm has been used for encryption of control messages (and any other messages) in cellular phones. This algorithm is highly criticized for its lack of strength. This algorithm is by far the weakest algorithm that was examined and most of the criticism to which it is subject is justifiable. The following are brief descriptions of the most alarming vulnerabilities that were encountered.

With regards to weak keys in their original definition, every single key of CMEA is a weak key. In other words it may be said that the CMEA function is its own inverse.

Since the CMEA algorithm does not support a CBC[4] mode or anything similar, nor does it support the use of IVs (Initialization Vectors), codebook attacks are feasible and easy to mount. Codebook attacks are one of the most primitive attacks and require hardly any facilities or computational power. Since the algorithm is also used to encrypt dialed digits that can be easily revealed (directly or by using side-information), codebook attacks are made easy. The encryption algorithm does not protect the entire plaintext that is provided as input, but rather, it protects

[4] The CBC (Cipher Block Chaining) mode is a mode of operation that causes dependency between every encrypted block and a predecessor block, making codebook attacks ineffective.

everything except the last bit. The last bit of the plaintext is always changed to its complement, regardless of the key that is used.

An important element of the CMEA algorithm is the T-Box, which does most of the transformation. The T-Box generates many equivalent outputs for various keys. The number of equivalents is very large, making four bits of the key simply meaningless. Therefore, the effective key length of the algorithm is only 60 bits.

The T-Box makes use of a fixed table of substitution, called the “Cave Table”. This table provides substitution that is not uniform but is skewed. Due to this property, the output of the T-Box is skewed as well, making some values appear more often and some values never appear.

Recovery of the T-box structure for a given key is enough to break the algorithm and the key itself does not need to be revealed at all. Recovery of the complete T-Box (for any block size) using a chosen-plaintext attack requires only 338 chosen plaintexts and very little computation time. Alternatively, if the block size is three bytes, as in most implementations, complete T-box recovery can be done using known-plaintext attacks requiring as little as 40-80 known plaintexts and 2^{32} operations (that can be done in parallel). If the block size is two bytes (as sometimes used), a complete recovery requires only four known plaintexts and 2^{32} operations. Alternately, the same recovery can be done using only two known plaintexts and some unknown ciphertext blocks.

There is absolutely no doubt that the CMEA algorithm is not suitable for any application that requires even the minimal level of security. This algorithm should be implemented only to satisfy specific needs such as interoperability with existing systems, compatibility with existing infrastructures, etc.

Blowfish

Blowfish was written by Bruce Schneier, a well-known cryptographer. The algorithm is designed to resist all known attacks on block ciphers. Blowfish is known for its high security and is available in several common encryption products.

Blowfish has some classes of weak keys. For these weak keys, separate rounds end up using the same round-keys. Keys belonging to these classes can be detected only in reduced-rounds versions of the algorithm and not on the full blowfish. Blowfish is known to successfully make (almost) every bit of the key affect most of the round-keys.

Blowfish is immune against differential related-key attacks because of the fact that every bit of the master key affects many round keys. The round-keys are highly independent, making related-key attacks very difficult or infeasible. Such independence is highly desirable.

Twofish

Twofish was also invented by Bruce Schneier, and was submitted as an AES candidate. The algorithm is considered to be secure after having been reviewed by several respectable cryptographers.

Mirza and Murphy published an attack by which guessing the 64 bits of the key that form the round keys (for 128-bit key encryption) can reveal information about the S-box, which is key-dependent, due to non-uniform distribution of round keys. According to the creators of the algorithm, the loss of entropy due to this attack is only 0.8 bit (for the specific round-key). Moreover, this attack cannot be applied, as it is, to keys of more than 128-bits. The same technique is claimed to apply to DES and to Triple-DES as well, although further experiments were not published to demonstrate this.

Further, Mirza and Murphy claimed that the number of sub-key pairs possible from a 128-bit key is not 2^{128}, as one would expect (so as not to lose entropy) but only 0.632 of that. The designers of the algorithm claim that the number of unique sub-key pairs is 2^{117} but that this quality does not affect the security of the algorithm.

The Twofish algorithm seems to be quite robust against known types of cryptanalysis according to its authors' claims, which have been proven to some limited extent and have not yet proven to be false.

According to the authors, the algorithm is resistant to related key attacks such as the slide-attack[5] and the related key differential attack. Also, according to its authors, the algorithm does not have any weak keys in the sense that using these specific keys will result in predictable sub-keys or in predictable round outputs. Related-key characteristics are not existent either. The authors, however, state some specific issues that have not as yet been covered by their analysis, mainly the resistance of the algorithm to chosen-key attacks. Possible vulnerability to these attacks may harm the algorithms security when used in some specific implementations, such as a hash function.

CAST

CAST is often perceived as one of the strongest algorithms available, and is deployed in several common encryption applications. CAST is relatively secure and is built to resist most of the known types of attacks. Following are its known weaknesses.

Whereas CAST is known to be quite resistant against linear cryptanalysis, it's key can be recovered by linear cryptanalysis using a known-plaintext attack. Such an attack on CAST-256

[5] The Sliding-Attack is an attack by which the opponent shifts rounds forward by key manipulation.

(with 256-bit key) requires 2^{122} known plaintexts (for 48 rounds). This attack is definitely not feasible even when considering only the amount of time it is likely to take. On 8-round CAST, linear cryptanalysis will require only 2^{34} known plaintexts. On 12-round CAST, such an attack will require 2^{50} known plaintexts, which is infeasible due to space constraints (requiring more than 9,000 terabytes of storage space). Similar infeasibility applies for 16-round CAST, which requires 2^{66} known plaintexts.

The 64-bit key version of CAST is somewhat vulnerable to differential related-key cryptanalysis. It can be broken by 2^{17} chosen plaintexts along with one related key query in offline work of 2^{48}. This result may be alarming if the implementation of the algorithm gives the attacker the chance to feed in ample amounts of chosen-plaintexts to be encrypted with keys that differ by values that are known to the attacker. However, it is not likely that CAST with 64-bit keys will ever be implemented.

Regarding differential-cryptanalysis, CAST-256 is considered secure. This is, of course, only true when the correct (specified) number of rounds is used. Differential cryptanalysis attack on CAST-256 when employing 48 rounds requires 2^{140} chosen plaintexts, which is clearly infeasible.

Rijndael

Rijndael was the NIST finalist in the AES competition and was declared the new standard symmetric cipher that is expected to replace DES. Briefly, it may be said that it was possible to break only up to eight rounds of the cipher with less work than of an exhaustive key-search (both for 192 and 256-bit keys). Nine rounds can be attacked using related-key attacks, but this is still impractical.

Four attacks were discovered to work against reduced-rounds versions of Rijndael: Square Attack, Improved Square Attack, Impossible Differential Attack and Reversed Key Schedule Attack.

A Square Attack breaks four rounds of Rijndael in 2^9 time, requiring 2^9 chosen plaintexts. An Improved Square Attack does that the same in 2^8 time. The same Square Attack when running on 5-rounds requires 2^{40} time, with 2^{11} chosen plaintexts. The Improved Square Attack does the same in 2^{39} time. The Sqaure Attack on six rounds requires 2^{72} time using 2^{32} chosen plaintexts while the Improved Square Attack will take 2^{71} time. The Impossible Differential Attack handles five rounds in 2^{31} time using $2^{29.5}$ chosen plaintexts, whereas a Reversed Key Schedule attack requires only 2^{11} chosen plaintexts for the same job. An attack on six rounds can be done using the Reversed Key Schedule Attack in 2^{63} time using 2^{32} known plaintexts. Attacks on six rounds were also shown using $6*2^{32}$ known plaintexts in 2^{44} operations.

For seven rounds of Rijndael, attacks on a 192-bit key were shown using 19*2^{32} known plaintexts in 2^{155} time. For a 256-bit key this attack would require 21*2^{32} known plaintexts and 2^{172} time. These attacks are, of course, infeasible due to length of time required and due to the fact that Rijndael employs more than seven rounds when in practical use.

Attacks on eight rounds will take the entire codebook of known plaintexts and 2^{188} time for a 192-bit key (2^{204} time for a 256-bit key) and are therefore totally inapplicable. A related-key attack on nine rounds of Rijndael with 256-bit key will require 2^{77} plaintexts to be encrypted using 256 related keys and 2^{224} time. This attack is clearly infeasible.

IDEA (IPES)

IDEA is an algorithm that uses a technique of multiple-group operations to gain its strength and to defeat most common attacks. Some vulnerabilities were found for reduced-rounds versions of the algorithm, and several classes of weak keys were also detected.

A successful differential attack was presented for 2.5 rounds[6] of IDEA, requiring 2^{10} chosen plaintexts and 2^{32} time (one day on a standard PC). This attack is believed to work for 3.5 rounds of IDEA in a time that is shorter than the time required for an exhaustive search. It is quite definite, though, that this attack cannot apply to eight rounds of IDEA. The attack is not based in any way on IDEA's key schedule, so it will not be made infeasible even if the key schedule mechanism is improved. Yet, it is not considered as a real threat since it works only on versions that are significantly reduced.

Another attack produced by Borst, Knudsen and Rijmen on reduced IDEA (3 rounds) required 2^{29} chosen plaintext pairs and 2^{44} encryptions. A third attack recovered the key of 3.5 rounds with a probability of 86%, using 2^{56} chosen plaintexts and 2^{67} encryptions. A truncated differential attack on 3.5 rounds of IDEA was found to recover the key using 2^{32} words of memory, 2^{56} chosen plaintexts and 2^{67} operations with a probability of greater than 83%. However, all these attacks, as the first one, do not apply for the full 8.5 rounds algorithm. According to Borst, Knudsen and Rijmen, the attacks can be extended to more than 3 rounds, though it is not likely that 8.5 rounds (full) IDEA is at any risk.

Full IDEA has 8 rounds, but the first 3 rounds seem to be highly vulnerable to related-key attacks such as key-schedule attacks and related-key differential timing attacks (by comparing the time it takes to perform decryption using multiple related keys), in addition to the attacks on reduced rounds that were presented above.

The algorithm also has a few classes of weak keys. Detection of a key belonging to these classes requires only two chosen plaintexts encryptions. Additionally, IDEA is claimed to have many

[6] One half round in IDEA represents the exit permutation.

other weak key classes, given its present keyschedule mechanism. Such weak keys should be avoided in any implementation.

Furthermore, with regards to weak keys, 2^{23} keys exhibit a linear factor. A linear factor is a linear equation between the plaintext, the key and the ciphertext, that applies for all input. Membership in this group can be detected by observing some plaintext-ciphertext pairs. In addition to this class, 2^{35} keys have global characteristics[7]. Additionally, in another group of 2^{51} keys, these are easily recoverable if detected as belonging to this group. Such membership detection takes two encryptions and the solution of 16 linear equations. All these weak keys must be detected and avoided by the implementation.

There also exists a related-key ciphertext-only attack on IDEA that requires $5*2^{17}$ related-keys encryptions, each on 2^{20} random (unknown) plaintexts.

ASYMMETRIC CIPHERS

The following sub-chapter presents the weaknesses that were found in algorithms for asymmetric encryption. An asymmetric encryption algorithm is an algorithm by which decryption and encryption are performed using two separate (but mathematically related) keys. Such algorithms are often called “Public key cryptosystems”.

RSA

RSA (named after Rivest, Shamir and Adelman) is the most commonly used asymmetric encryption algorithm, being adopted by many products since the 1970’s. Its strength relies on the mathematical complexity of prime-factoring, which is still high enough to offer robustness when large primes are used. No severe inherent vulnerabilities have been found in RSA and all vulnerabilities that have ever been presented refer to a specific implementation of the algorithm. Although the algorithm is safe by its academic nature, it is probably the most susceptible to weaknesses due to weak implementation. These restrictions on implementations of the algorithm are hereby presented, under the assumption that the reader is familiar with the RSA basics and notations.

RSA private keys (hereafter called “Private Exponents”) are likely to be weak if their value is less than N0.292. This figure is occasionally adjusted and it is believed that practical secure implementations require the private exponent to be larger than N0.5. This latter figure has not been proven.

[7] For such keys, a known relationship between inputs is mapped to a known relationship between the outputs

The system (N,d,e) is likely to be insecure if (p-1), for the p that is one of the factors of N, is a product of small primes.

When implementing an RSA system with several key-pairs, the implementer often chooses to use the same N for all key-pairs, thus saving computation time. However, since the private and public exponents together always assist in factoring N, every single member of the system will be able to factor N with his key-pair and use the result to invert any public exponent to the corresponding private exponent. Therefore, it is necessary to generate a new N value for each key-pair.

When using RSA for signature, the blinding feature[8] of the algorithm may be used by an attacker to cause a person to sign a message that he/she would not willfully sign. Applying a hash function on the message prior to signing can effectively solve this problem. This will allow the blinding feature but will make the blinding-based attack infeasible.

Whereas it is obvious that the system is weak when the private exponent is small, it has also been proven that the system is weak if the public exponent is too small. This claim is backed up by the proven Coppersmith Theorem and the Hastad Broadcast attack. Low public exponents are also risky when related messages are encrypted with the same public key. Such related messages are messages for which there is a polynomial that converts one message to the other. Furthermore, the related-messages attack was expanded so that it applies to two messages with different random padding (even when the padding values are unknown), as presented by Coppersmith in the Short-pad attack.

It is important to keep the entire private exponent (private key) secure as a whole. There are attacks that enable an adversary to recover the remaining bits of a private key using known bits. If the public exponent is smaller than N0.5, then the private key can be recovered from a fraction of its bits.

As with many other algorithms, RSA is vulnerable to timing attacks that are easily implemented, as well as to power consumption attacks. The “Repeated Squaring Algorithm” can be used to effectively mount a timing attack. The solutions against timing attacks are either to add artificial delays or to use blinding.

When RSA is implemented using the Chinese Remainder Theorem (CRT) to save computation time, a breach can be formed if the signature algorithm fails for some reason. In this case, the attacker may be able to factor N using the mistaken signature that is generated. This problem can be solved either by adding random padding to the message, so the attacker never knows the exact

[8] A description of the blinding feature is beyond the scope of this document. It will only be noted that this feature enables an entity to sign a message without knowing its content.

message that was signed, or by verifying the signature prior to presenting it, in order to ensure that it was calculated properly.

When RSA is implemented according to PKCS#1, a standard block represents an RSA encrypted message. This block starts with a fixed constant ("02" in this case) that symbolizes an encrypted block. When a server receives a message, it decrypts it and issues an error message if these first two bytes don't match this fixed value. According to Bleichenbacher's oracle attack, this "hint" in the form of an error message is enough for an attacker to mount some sort of a brute-force attack. It is therefore important to implement a system that does not generate public error messages when a fixed block is not encountered.

If p and q that are used to generate N are too close to each other, then Fermat's factoring is possible, making the system highly insecure. Thus, the difference between the two primes should be at least $N^{0.25}$.

Diffie-Hellman (DH)

Diffie-Hellman is the one of the most common asymmetric algorithms. It is mainly used for two anonymous parties, who do not have a secure channel between them, to efficiently and securely exchange symmetric keys that will be used for the symmetric encryption of session data. Diffie-Hellman is considered to be secure. However, its nature of being an algorithm which retains anonymity implies that it is highly vulnerable to man-inthe- middle attacks. Diffie-Hellman, when not combined with adequate authentication methods, may therefore be highly risky due to these kinds of attacks.

HASH FUNCTIONS

Hash functions are functions that receive an arbitrary-length input and produce a fixed-length output. These functions are usually used for authentication and for digital signatures. Cryptographic hash functions are required to be collision free[9] and non-invertible[10].

MD5 (Message Digest 5)

MD5 is one of the most commonly used message digest (hash) functions. The MD5 algorithm is also known to be secure. No one has yet presented an efficient method for inverting the function. The function is also known to be collision-free.

[9] A hash function is regarded as "collision-free" if one cannot find two messages for which the function generates the same digest.

[10] Inverting a hash function refers to the action of recovering parts of the original message (input) from the digest (output). Complete inversion is, of course, impossible by the definition of a hash function as a function from a group of size aleph-zero to a group of a finite size.

The closest attack found was the ability to generate a pseudo-collision[11] in 2^{16} operations. However, this does not seem to cause any notable risk.

CONCLUSIONS

Whereas most of the algorithms do not suffer from real vulnerabilities that make their implementation useless, some suffer from some kind of vulnerability that may put the user at risk in some circumstances. Still, it must be remembered that cryptographic forums are mostly academic, so they often publicize weaknesses that have purely academic value. Most of the weaknesses that were detected in algorithms that are known to be secure, have a high educational and professional value, but usually cannot be exploited in practice. Yet, such academic evaluation is perhaps the most reliable source for information about an algorithm’s strength. Therefore, algorithms for which no weaknesses were published are not necessarily secure algorithms but most likely are algorithms that were never examined by professional entities. In our view, the algorithms can be divided into three groups with respect to their strength, in ascending order as follows.

1. Algorithms that were examined by the academia (or by respectable cryptographers) and in which serious exploitable flaws were found (such as the CMEA). Also included in this group are algorithms that were not examined by the academia for reason of being too weak to start with. Algorithms belonging to this group are the Cellular Message Encryption Algorithm, Single DES, RC2 and RC4. Implementation of algorithms of this group is highly discouraged.
2. Algorithms that were not examined by the academia (or by any other group of respectable cryptographers) either for their lacking interest to the public or for their not being open-sourced. Algorithms belonging to this group are (by the definition of this group) not present in this document. Ciphers belonging to this group should be implemented only if there is a need for these specific ciphers, or if there is a basis for the assumption that the lack of publications about the strength of the algorithm is temporary.
3. Algorithms that were examined by the academia (or by any other group of respectable cryptographers) and for which no weaknesses were found (an uncommon situation) or for which the only weaknesses that were detected are not exploitable. Weaknesses are usually not exploitable if they require infeasible work time, amount of required known plaintext which exceeds the number of possible plaintext blocks, enormous amounts of related-key searches or chosen plaintexts, etc. Another category of non-exploitable weaknesses is weaknesses in

[11] A pseudo-collision is a case in which an attacker can predict two keys that will generate the same hash value for a single given message, when using keyed hash.

reduced-round variants, in cases where it is clear that the weakness cannot be extended to the full-version of the cipher. Clearly, algorithms of this group are recommended for use in practice.

REFRENCES

- Dan Boneh, *Twenty Years of Attack on RSA*
- Dan Boneh and Glenn Durfee, *Cryptanalysis of Low-Exponent RSA*
- Fauzan Mirza, *Linear and S-Box Pair Cryptanalysis on DES*
- http://www.computerworksnorthwest.com
- www.cdc.informatik.tu-darmstadt.de
- Niels Ferguson, John Kelsey, Stefan Lucks, Bruce Schneier, Mike Stay, David Wagner, and Doug Whiting, *Improved Cryptanalysis of Rijndael*
- Bruce Schneier, John Kelsey, Doug Whiting, David Wagner, Chris Hall, and Niels Ferguson, *On the Twofish Key Schedule*
- Lars R. Knudsen, Vincent Rijmen, Ronald L. Rivest, and M.J.B. Robshaw, *Cryptanalysis of RC2*
- www.cc.gatech.edu
- www.neiu.edu/~ncaftori/school
- Lars R. Knudsen and Willi Meier, *Correlations in RC6*
- Sean Murphy and Matt Robshaw, *New Observations on Rijndael*
- John Kelsey, Bruce Schneier, and David Wagner, *Key Schedule Cryptanalysis*
- http://computer.howstuffworks.com
- Henry Gilbert and Marine Minier, *A Collision Attack on Seven Rounds of Rijndael*
- John Kelsey, Bruce Schneier, and David Wagner, *Related-Key Cryptanalysis*
- http://www.discretix.com
- Eli Biham and Adi Shamir, *Differential Fault Analysis on Secret-Key Cryptosystems*
- Eli Biham and Nathan Keller, *Cryptanalysis on Reduced Rijndael*
- http://citeseerx.ist.psu.edu
- Benne De Weger, *Cryptanalysis of RSA with Small Prime Difference*
- David Wagner, Bruce Schneier, and John Kelsey, *Cryptanalysis of the Cellular Message Encryption Algorithm*
- Doug Whiting, John Kelsey, Bruce Schneier, David Wagner, Niels Ferguson, and Chris Hall Further, *Observations on the Key-Schedule of Twofish*
- http://www.logicalsecurity.com
- Eli Biham and Adi Shamir, *Differential Cryptanalysis of Full DES*

- Eli Biham, Alex Biryukov, *An Improvement of Davies Attack on DES*
- H.M. Heys and S.E. Tavares, *On the Security of the CAST Encryption Algorithm*
- http://www.networksorcery.com
- C. Adams, H.M. Heys, S.E. Tavares, M. Wiener, *An Analysis of the CAST- 256 Cipher*
- Peter C. Weiner, *Using Content-Addressable Search Engines to Break DES*
- Joan Daemen, Rene Govaerts, and Joos Vandewalle, *Cryptanalysis of 2.5 Rounds of IDEA*
- Bruce Schneier, *MD5 Cryptanalysis*

RFID –SECURITY AND PRIVACY ISSUES

LT. Octavian PALEACU

INTRODUCTION

This paper it meant to provide some thoughts on security and privacy issues concerning RFID systems and to highlight some of the areas that have to be considered regarding this topic.

To deal with security and RFID means to deal not only with security aspects of RFID systems but also with security aspects of anything or anyone affected by RFID systems. The widespread dissemination of identification technology and storage devices certainly has side effects and can lead to new threats in other areas and applications. Therefore the use of RFID challenges existing security systems, which have to be reviewed[12].

As with any other security measures, RFID security has to be a process rather than a singular event. This process should start at the technological basis, providing security mechanisms for applications built on this basis.

RFID security is not limited to technology but also has to deal with the question how secure it is to rely on information provided by RFID.

UNDERSTANDING RFID

PRINCIPLES

Although the foundation of the Radio Frequency Identification (RFID) technology was laid by past generations, only recent advances opened an expanding application range to its practical implementation.

RFID is only one of numerous technologies grouped under the term Automatic Identification (Auto ID), such as bar code, magnetic inks, optical character recognition, voice recognition, touch memory, smart cards, biometrics etc. Auto ID technologies are a new way of controlling information and material flow, especially suitable for large production networks.

The RFID technology is a means of gathering data about a certain item without the need of touching or seeing the data carrier, through the use of inductive coupling or electromagnetic waves. The data carrier is a microchip attached to an antenna (**together called transponder or tag**), the latter enabling the chip to transmit information to a reader (or transceiver) within a given range, which can forward the information to a host computer. The middleware (software

[12] RFID Security Issues by Andreas Krisch Version: 1.0

for reading and writing tags) and the tag can be enhanced by data encryption for security-critical application at an extra cost, and anti-collision algorithms may be implemented for the tags if several of them are to be read simultaneously.

RFID COMPONENTS (TAGS & READERS)

RFID tags and readers can be grouped under a number of categories.

Classification of RFID tags

I. Passive:

- also called ‘pure passive’, ‘reflective’ or ‘beam powered’;
- obtains operating power from the reader;
- the reader sends electromagnetic waves that induce current in the tag’s antenna, the tag reflects the RF signal transmitted and adds information by modulating the reflected signal.

II. Semi-passive:

- uses a battery to maintain memory in the tag or power the electronics that enable the tag to modulate the reflected signal;
- communicates in the same method, as the other passive tags.

III. Active:

- powered by an internal battery, used to run the microchip’s circuitry and to broadcast a signal to the reader- generally ensures a longer read range than passive tags;
- more expensive than passive tags (especial because usually are read/write);
- the batteries must be replaced periodically.

By the tag’s memory type

I. Read-only:

- the memory is factory programmed, can not be modified after its manufacture;
- its data is static;
- a very limited quantity of data can be stored, usually 96 bits of information;
- can be easily integrated with data collection systems;
- typically are cheaper than read-write tags.

II. Read-write:

- can be as well read as written into;
- its data can be dynamically altered;
- can store a larger amount of data, typically ranging from 32 Kbytes to 128 Kbytes;

- being more expensive than read-only chips, is impractical for tracking inexpensive items.

By the method of wireless signal used for communication between the tag and reader

I. Induction:

- close proximity electromagnetic or inductive coupling—near field;
- generally use, LF and HF frequency bands.

II. Propagation:

- propagating electromagnetic waves—far field;
- operate in the UHF and microwaves frequency bands.

Classification of readers

By design and technology used

I. Read:

- only reads data from the tag;
- usually a micro-controller-based unit with a wound output coil, peak detector hardware, comparators, and firmware designed to transmit energy to a tag and read information back from it by detecting the backscatter modulation;
- different types for different protocols, frequencies and standards exist.

II. **Read/write:**

- reads and writes data from/on the tag.

By fixation of the device

I. Stationary:

-the device is attached in a fixed way, for example at the entrance gate, respectively at the exit gate of products.

II. Mobile:

- in this case the reader is a handy, movable device.

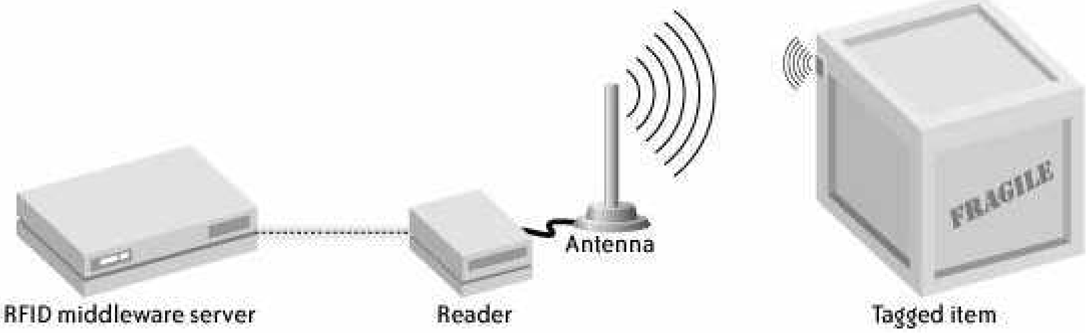


Fig.1. Working of RFID

ADVANTAGES OF RFID

ADVANTAGES & BENEFITS

Though RFID is not likely to entirely replace commonly used barcodes in the near future, the following advantages suggest to additionally apply RFID for added value of identification[13]:

- Tag detection not requiring human intervention reduces employment costs and eliminates human errors from data collection;
- As no line-of-sight is required, tag placement is less constrained;
- RFID tags have a longer read range than, e. g., barcodes;
- Tags can have read/write memory capability, while barcodes do not;
- An RFID tag can store large amounts of data additionally to a unique identifier;
- Unique item identification is easier to implement with RFID than with barcodes;
- Tags are less sensitive to adverse conditions (dust, chemicals, physical damage etc.);
- Many tags can be read simultaneously;
- RFID tags can be combined with sensors;
- Automatic reading at several places reduces time lags and inaccuracies in an inventory;
- Tags can locally store additional information; such distributed data storage may increase fault tolerance of the entire system;
- Reduces inventory control and provisioning costs;
- Reduces warranty claim processing costs.

For utility locating and management purposes, RFID offers several significant features and advantages to facility owners:

Able to Operate in Challenging Environments: RFID tags for utility marking can be easily designed to withstand a variety of challenging environmental operating conditions. Factors such as temperature variations, moisture and dirt have little or no impact on their ability to function and, assuming that passive RFID tags are used, the tags do not even require an on-board battery. They are instead powered by the reader - a very important factor for long-term underground installation.

- **Enables Flexible Strategies for Locating Assets**: Depending on the type of tags and readers used and future advances in RFID technology, field-locating utilities using RFID has the potential to be useful on multiple levels. The most common scenario may be to use handheld readers to locate utilities horizontally and mark them in the field. However, in theory, readers may also be installed on excavating and construction equipment, in

[13] "In proceedings of The Modern Information Technology in the Innovation Processes of the Industrial Enterprises" – MITIP 2006

order to proactively warn the operator, in real-time, that utilities are in close proximity to the excavation area.

- **Implements Database Back-End Storage**: As mentioned earlier, RFID tags typically are able to store only 2 kilobytes of data on-board - just enough to store a unique identification code and not much else. Some extended-capability RFID tags are capable of storing up to 64 kilobytes of data; however, it is much more common to implement a back-end database for additional storage. This makes RFID a powerful tool indeed, as the database can be designed to be as simple or as extensive as the user desires. The database records are associated with the individual tags using the unique identification number and the database can be used to store all of the information that is pertinent to the tagged asset. The database can store information about the contractor, installation conditions, installation date, installation depth, dimensions, specifications, CAD drawings, shop drawings, installation photographs, and maintenance records. Maintenance records are discussed in more depth below.
- **Facilitate Maintenance Tasks**: A technology such as RFID would not only facilitate location of utilities, it would also ensure that regular maintenance and replacement tasks are more accurately associated with specific assets, over their entire lifetime. Because each RFID tag has a unique identification number associated with it, these tags can be used to mark specific pipe sections, splices, valves, or other appurtenances within the utility network. These different parts may have varying expected life spans, and will likely be serviced individually, over time. With RFID, facility owners would be able to more easily maintain part-specific maintenance records than existing practices permit.
- **Serve as a Basis for Future Decision Support Tools**: In addition to the uses mentioned above, RFID has the potential to be leveraged as a basis for a variety of innovative decision support tools. The data can easily be integrated into GIS systems, served over the Internet, or manipulated through handheld devices.

IMPLEMENTATION

The RFID industry is segmented into application categories such as:

a) Manufacturing and supplies chain management:

- Inventory management
- Tracking management
- Quality control management
- Resource management
- Distribution management

- Material processing management
- Safety management
- Picking management
- Receiving management
- Shipping management
- Inventory management
- Shelf-stock management
- Checkout management

b) Traffic transportation:

- Public transport ticket
- Toll collection
- Smart card key
- Automatic vehicle location

c) Monitoring and tracking:

- Parcels, mall bags
- Luggage
- Digital signature
- Library inventory

d) Agriculture:

- Animal tracking
- Animal diagnostic
- Crop identification

e) Environment:

- Waste haulage
- Recycling

f) Sports and games:

- Sports event timing
- Tracking golf balls
- Gaming chips

g) Government and military:

- Military logistics

h) Healthcare:

- Pharmaceutics
- Hospital equipment and personnel

- Patients medical history
- Implants and prostheses
- Elderly care

j) Human identification:

- Digital ID
- Electronic passport
- Facility access
- Punishment system

FEARS SURROUNDINGS RFID

1. RFID, Big Brother?!

The intrusion of the Radio Frequency Identification, RFID, in our lives is increasing day by day. Just like any other new technology, it is entering our everyday life at a lightning fast speed whether we like it or not.

All though this small chip may look quite harmless, it is not so. The day is not far when everything from your cell phone to your shoes to your car to even your clothes is rigged by an RFID tag. Big Brother would be watching over you always[14].

1.1. Direct monitoring (by vendors)

One fear is that someone in the manufacturing or sales chain will use information gleaned from RFID systems to learn information about or track a consumer contrary to his or her interests and desires. While linking the serial number on an RFID tag back to the purchaser can have many substantial benefits, misuse of that same linkage may constitute a privacy invasion.

For example, RFID could be used to note a customer's purchases and then learn when the customer returns to the store—or at least when the associated RFID tag (perhaps in clothing) has done so. Conceivably, information like this could be used to develop a dossier about a consumer and his or her activities. The mere collection of too-detailed information may offend consumers' sense of privacy. Use of this information for marketing purposes may offend others, and other broader monitoring may compound the offense.

Questions about direct monitoring parallel longstanding debates about what retailers and marketers may do with consumer information they gather through transactions. This is not a new issue, but an extension of an old one[15].

[14] http://www.articlesbase.com/rticles/rfid-and-the-big-brother-effect-1584446.html

[15] "RFID Tags and Privacy: How Bar-Codes-On-Steroids Are Really a 98-Lb. Weakling" by Jim Harper -2004

1.2. Indirect monitoring (by third parties)

The second way RFID systems may be used to compromise privacy is when an outsider to an RFID network uses the existence of RFID tags to read and collect personally identifiable information contrary to the interests of those monitored. Someone may scan an RFID tag and use further reading of the tag elsewhere as a proxy for the presence of the same individual in the second location. Collecting that information, or subsequently using it in various ways, may compromise privacy and threaten other interests.

For example, union operatives could surreptitiously scan for RFID tags on clothing, ID cards, and so on at the entrance to a right-to-work rally. When an RFID tag's serial number that was scanned at the rally arrives with a person at a union hall, he or she could face retaliation from the union.

Of course, for this method to successfully compromise privacy, it is necessary at some point to identify the person associated with the tag. This type of monitoring would be prone to significant error and it entails many challenges. But it is at least a conceivable way, the technology's opponents argue, that RFID could be used to invade privacy.

Concerns with indirect monitoring using RFID are similar to concerns over monitoring using surveillance cameras. Sometimes it is appropriate; other times it is not. Almost always, it is ineffective at deriving much in the way of useable personal information. Photographic surveillance seems much more powerful than RFID-based surveillance because it captures true personal information—an individual's appearance—on each "scan." RFID-based surveillance will capture the presence of a tag, which may or may not correlate to any individual[16].

2. Possible virus attack

RFID technology has few well-known threats, like sniffing, tracking, spoofing, replay attacks & DoS. Though the absence of "in the wild" RFID exploit makes people think that the power constraints of RFID tags make them invulnerable for such attacks. But what about the vulnerability of RFID middleware system which comprises of the combination of RFID reader interfaces, application servers, & back-end databases.

Example of virus attack :

In order to determine if such technology is prone to the same security threats suffered by computer systems, Dr. Gasson's chip was infected with a virus. The researcher found that the corruptive code was then passed onto the main system used to communicate with the chip. Had other devices been connected to the system, the infection might have spread. In summary, Dr.

[16] "RFID Tags and Privacy: How Bar-Codes-On-Steroids Are Really a 98-Lb. Weakling" by Jim Harper -2004

Gasson said the research showed "that implantable technology has developed to the point where implants are capable of communicating, storing and manipulating data. They are essentially mini computers. This means that, like mainstream computers, they can be infected by viruses and the technology will need to keep pace with this so that implants, including medical devices, can be safely used in the future."

PROTECTING COSTUMER PRIVACY

In the near term, RFID tags are unlikely to reach the hands of consumers on a regular basis—but they will eventually. As a result, consumer privacy is becoming a major issue around RFID in the media, and a critical concern for those that use it.

Policy guidelines are in and of themselves insufficient to guarantee consumer privacy. After all, RFID-tag reading is not a visible process. Consumers can have no easy way of knowing when RFID policies are adhered to or breached. In fact, RFID tags can be so small and easily embedded in products, that consumers may not even know when they are carrying them!

1. Countermeasures and Self-Help

Industry approaches to consumer privacy vary. Some enterprises are proposing policy guidelines for use of RFID information, like to enforce clear labeling of RFID-tagged products, among other measures.

So what technologies can help protect consumer privacy? Here are a few approaches proposed by scientists:

- **The “Kill Tag” approach**

The most straightforward approach for the protection of consumer privacy is to “kill” RFID tags before they are placed in the hands of consumers. A killed tag is truly dead, and can never be re-activated.

The standard mode of operation is indeed for tags to be killed upon purchase of the tagged product. With their proposed tag design, a tag can be killed by sending it a special “kill” command (including a short 8-bit “password”).

For example, a supermarket might use RFID tags to facilitate inventory management and monitoring of shelf stocks. To protect consumer privacy, checkout clerks would “kill” the tags of purchased goods; no purchased goods would contain active RFID tags.

Can be the “Kill” approach inadequate? There are many environments, however, in which simple measures like “kill” commands are unworkable or undesirable for privacy enforcement. For example, a Prada store in New York City tracks the RFID tags of items held by customers in order to display related accessories on nearby screens.

Other examples of RFID-tag applications for ordinary consumers include a physical access control1, theft-protection of belongings, and wireless cash cards[17].

- **The Faraday Cage approach**

An RFID tag may be shielded from scrutiny using what is known as a Faraday Cage—a container made of metal mesh or foil that is impenetrable by radio signals (of certain frequencies). Indeed, petty thieves are already known to use foil-lined bags in retail shops to circumvent shoplifting-detection mechanisms.

If high-value currency notes do indeed come supplied with active RFID tags, then it is likely that foil-lined wallets will become big sellers! At least one company already have a Faraday-cage-based product for privacy purposes[18].

RFID tags will inevitably see use, however, in a vast range of objects that cannot be placed conveniently in containers, such as clothing, wrist-watches, and even human beings

- **The Active Jamming Approach**

Active jamming of RF signals is another, related physical means of shielding tags from view. The consumer could carry a device that actively broadcasts radio signals so as to block and/or disrupt the operation of any nearby RFID readers.

This approach may be illegal – at least if the broadcast power is too high – and is a crude, sledgehammer approach. It could cause severe disruption of all nearby RFID systems, even those in legitimate applications where privacy is not a concern.

The approach we propose in this paper is akin to "jamming," but is much more subtle in its operation, interacting cleverly with the RFID "singulation" protocol to disrupt only certain operations[19].

- **The "Smart" RFID Tag Approach**

Another approach is to make the RFID tags "smarter," so they interact in a way that protects privacy better and provide the desired active functionality. This of course will involve the use of cryptographic methods.

These approaches are exceptionally challenging to design with the severe cost constraints on the basic RFID tag. (With a budget of five cents, there is very little to spend on additional logic gates!)

Three instances of the "smart RFID-tag" approach that have been proposed are the hash-lock method, the re-encryption method (in several forms), and silent tree-walking.

[17] http://www.rsa.com/rsalabs/

[18] "The Blocker Tag: Selective Blocking of RFID Tags for Consumer Privacy" – Ari Juels and Ronald L., RSA Laboratories

[19] Idem 7

- **The “Hash-Lock” Approach**

A tag may be “locked” so that it refuses to reveal its ID until it is “unlocked.”[20]

In the simplest scenario, when the tag is locked it is given a value (or meta-ID) y, and it is only unlocked by presentation of a key or PIN value x such that y = h(x) for a standard one-way hash function h.

In the supermarket example, tags may be locked at check-out time. A consumer could provide a meta-ID y for the tags (perhaps on a loyalty card), and then transmit the unlocking PIN x via some special device (perhaps requiring physical contact) to unlock tags on returning home.

- **Blocker Tag**

To ensure its attractiveness as a widespread tool for protection of consumer privacy, the blocker tag must create little or no disruption of normal RFID-based commercial processes. A universal blocker tag would be counterproductive: it would provide privacy protection, but at the cost of indiscriminately disrupting all RFID-tag reading in its vicinity.

For the purpose of practical privacy enhancement, we must instead require the use of selective blocker tags. This involves the special designation of one or more zones for privacy protection. Thus a “privacy zone” consists of a restricted range of tag serial numbers targeted for protection (i.e., simulation) by a selective blocker tag

Supermarkets make use of blocker tags whose privacy zone consists of all serial numbers with a leading ‘1’ bit. Packages in Supermarkets each bear an RFID tag with a unique serial number used for purposes of inventory control. As initially programmed, and while an item is inside the supermarket or its warehouses, the serial number in its RFID tag carries a leading ‘0’ bit. At this point, blocker tags don’t disrupt the reading of tags.

When the RFID-tag reader at a cash register scans an item for purchase by a customer, it also transmits a tag-specific key to the RFID tag on the item.2 This causes the leading bit in the serial number of the tag to flip to a ‘1’. Supermarkets provide its customers with free blocker tags. These are available embedded in shopping bags at registers or as stickers to be placed on items.

2. Existing legal protections against RFID privacy invasion.

Existing law, such as property rights and the common law privacy torts, already substantially delimit the use of RFID and its potential for abuse. They head off many
RFID privacy issues in at least two ways.

[20] “Security and privacy aspects of low-cost radio frequency identification systems.” - S. A. Weis, S. Sarma, R. Rivest, and D. Engels , 2003

First, existing law gives consumers substantial autonomy and control over what goes into their homes, what rides in their cars, and what goes on or in their bodies. Many concerns expressed about RFID omit the almost total power consumers have.

Because of the existence of property rights, people aren't oblige to allow RFID readers into their homes, though they may certainly want them to simplify grocery shopping and cooking—or to locate RFID tags. Many concerns about RFID presume that RFID readers will somehow be able to inventory the contents of homes. Read ranges will simply not be long enough, RFID readers will not be allowed in homes without the permission of homeowners, and correlations between goods and tags will not be publicly available.

Stories of the potential for human implantation of RFID tags have led to a charged atmosphere of concern. But this implantation of an RFID tag into an individual against his or her will would be a tort and probably a crime if done by a private actor, and a violation of constitutional rights if done by a public official.

It takes a lot of imagination, and a lack of legal comprehension, to buy into many of the concerns being aired about RFID. The web of laws protecting human autonomy and property rights sharply limit the chance that RFID will be used in ways consumers does not want.

A second way that law circumscribes RFID is by outlawing harmful uses of it. The genuine harms that potentially could be done to consumers via RFID are illegal already.

A body of state law, the privacy torts, bars various invasions of privacy and gives a cause of action to victims no matter what technology was used to collect the information used in an invasion.7 Various statutes prohibit all variety of harms that may be done with information, whether derived via RFID systems or not. It is illegal (if it is possible) to use RFID in the course of identity fraud, theft, burglary, stalking, murder, or conspiracy to commit any of those crimes. Someone who places an RFID reader surreptitiously on another's property or in another's home must commit trespass, burglary, or both to do it[21].

The mischief that might be made possible by RFID is already against the law. Ignorance of the law allows many to believe that RFID has outsized power to affect consumers' privacy.

Many concerns about RFID also arise from ignorance about the economic constraints in which the RFID user community will operate. As noted above, vendors will be driven to use cheap, dumb tags useful for tracking inanimate objects in controlled environments, but not good at all for tracking humans in our social environments. While self-help is a worthy, perhaps superior, failsafe should economic constraints on RFID deployments fail; existing law represents the final

[21] http://www.privacilla.org/releases/Torts_Report.html

bulwark against abuse of RFID systems. It punishes wrongdoing and empowers consumer to reject uses of RFID that they do not want.

In any event, RFID should not be assumed to have capabilities beyond what the laws of physics and economics will allow.

CONCLUSIONS

True consumer's interests are broader. Along with privacy, consumers want a complex and constantly shifting mix of low prices, convenience, customization, quality, customer service, and other characteristics in their goods and services. Radio frequency identification technology will bring help to producers, marketers, and retailers better understand and serve the mix of interests consumers have.

The components that go into RFID readers and tags are simple radio communications, but their smaller size and broad deployment enhance the power of the technology and raise concerns about the privacy effects of RFID deployment. These concerns are often premised on unlikely assumptions about where the technology will go and how it will be used.

Any inclination to abuse RFID technology will be hemmed in by a variety of social forces, economic forces being one of the most significant: The typical RFID tag in the consumer goods environment will be cheap, dumb, and not good for much more than tracking inventory.

As economic actors, we have substantial power to dictate in the give and take of the market how RFID will be used. We will likely demand tags linking to our identities in certain applications—such as consumer electronics—but may object to the presence of RFID tags in other situations.

MANAGING AN HOST-BASED INTRUSION DETECTION AND PREVENTION SYSTEM

1st LT Bogdan RUSU

Introduction

Network security has been an issue since computers have been connected together. The evolution of the Internet has increased the need for security systems. Important security products that emerged are Intrusion Detection System (IDS)/ Intrusion Prevention System (IPS). These products were developed to detect abnormal behavior of information systems and networks, indicating a breach of the security policy.

With industry's widespread adoption and integration of intrusion detection/ prevention, it has become clear that intrusion detection and prevention systems (IDPSs) are an integral part of an organization's infrastructure. Many organizations and companies have deployed, or are in the process of deploying, enterprise-wide IDPS solutions.

With the development of TCP/IP, security problems have become more frequent and taken very different forms, and have lead to the development of new security techniques. Very early in the development of the Internet, vulnerabilities affecting operating systems have allowed attackers to move from system to system. Detecting attackers has been a necessity for military environments and not only. This paper will explore the challenges of managing an host- based IDS/ IPS in our days.

1 – Intrusion Detection and Prevention Principles

1.1 – Understanding intrusion detection and intrusion prevention

Intrusions are attempts to compromise the confidentiality, integrity, availability, or to bypass the security mechanisms of a computer system or network (illegal access).

Intrusions have many causes, such as malware (worms, spyware, etc.), attackers gaining unauthorized access to systems, and authorized users of systems who misuse their privileges or attempt to gain additional privileges for which they are not authorized.

Although many intrusions are malicious in nature, many others are not; for example: a person might mistype the address of a computer and accidentally attempt to connect to a different system without authorization.

As defined by Rebecca Bace and Peter Mell, "Intrusion detection is the process of monitoring the events occurring in a computer system or network and analyzing them for signs of *intrusions,* defined as attempts to compromise the confidentiality, integrity, availability, or to bypass the security mechanisms of a computer or network."[22]

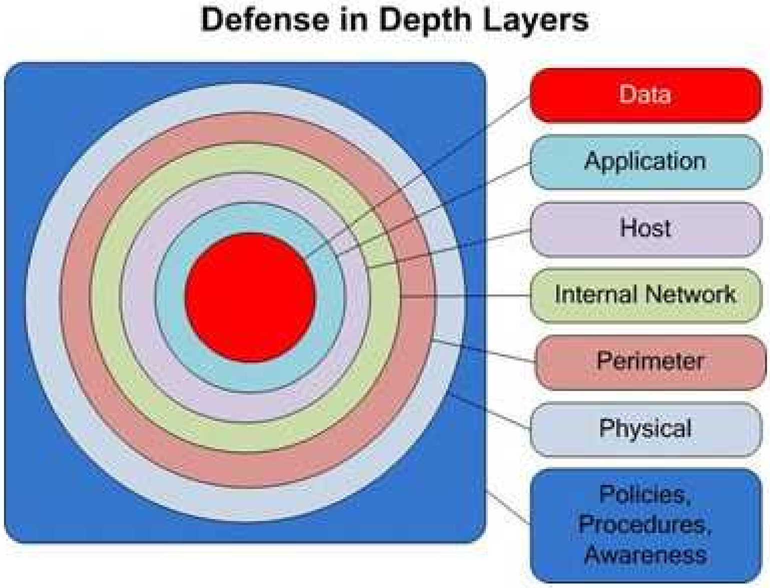
Intrusion prevention is the process of performing intrusion detection and attempting to stop detected possible incidents.

Because IDS and IPS technologies offer many of the same capabilities, and administrators can usually disable prevention features in IPS products, causing them to function as IDSs, intrusion detection and prevention systems (IDPS) concept used throughout the rest of this paper refers to both IDS and IPS technologies.[23]

1.2 – What is Defense in Depth

Defense in depth is an information assurance (IA) strategy in which multiple layers of defense are placed throughout an information technology (IT) system.[24]

Defense in Depth is a commonly used approach in IT security to address data confidentiality, integrity and availability. Defense in Depth utilizes multiple "layers" to protect an organization's critical assets, as shown in figure below.



In other words, an attack that is not stopped by an outer layer will be stopped by an inner layer. Any exploit will have to break through multiple defense layers for it to be successful. Having a Defense in Depth architecture does not assure an organization that it won't be attacked. It does make it as difficult as possible for the attacks that inevitably will occur to succeed. As attacks get more numerous and more complex, organizations need to develop more complex defense strategies.

[22] http://csrc.ncsl.nist.gov/publications/nistpubs/800 800-31/sp800 sp800-31.pdf
[23] http://csrc.nist.gov/publications/nistpubs/800-94/SP800-94.pdf
[24] http://en.wikipedia.org/wiki/Defense_in_depth_(computing)

2 – IDPS Technologies

2.1 – Intrusion Detection System

IDS are the second layer of defense. It detects the presence of attacks within traffic that flows in through the holes punched into the firewall, but doesn't block these patterns. It only reports that they took place.

Types of IDS

There are two main categories of IDS based on the IDS alarm triggering mechanism - action that causes the IDS to generate an alarm. They are: **anomaly detection based IDS** and **misuse detection based IDS**.

With *anomaly detection*, you need to create a profile for each user group on your system. These profiles can be built automatically or created manually. How the profiles are created is not important as long as the profiles accurately define the characteristics for each user group or user on your network. These profiles are then used as a baseline to define normal user activity. If any network activity deviates too far from this baseline, then the activity generates an alarm. Because this type of IDS is designed around user profiles, it is also sometimes known as *profile-based detection*.

Anomaly detection systems offer several benefits. They can detect insider attacks or account theft very easily. If a real user or someone using a stolen account starts performing actions that are outside the normal user-profile, it generates an alarm. Because the system is based on customized profiles, it is very difficult for an attacker to know for sure what activity he can do without setting off an alarm. The intrusive activity generates an alarm because it deviates from normal activity, not because someone configured the system to look for a specific traffic.

Like every IDS, anomaly detection systems also suffer from several drawbacks. The first obvious drawback is that the system must be trained to create the appropriate user profiles. During the training period to define what normal traffic looks like on your network, the network is not protected from attack. Just defining "normal" is a challenge in itself. Maintenance of the profiles can also become time-consuming. However, the biggest drawback to anomaly detection is probably the complexity of the system and the difficulty of associating an alarm with the specific event that triggered the alarm. Also, you have no guarantee that a specific attack will even generate an alarm. If the intrusive activity is too close to normal user activity, then the attack will go unnoticed. It is also difficult for you to know which attacks will set off alarms unless you actually test the attacks against your network using various user-profiles.

The second major category of IDS triggering is known as misuse detection. Misuse detection is also sometimes referred to as signature-based detection because alarms are generated based on specific attack signatures. These attack signatures include specific traffic or activity that is based on known intrusive activity.

Misuse detection provides various benefits. One of the first benefits is that the signature definitions are modeled on known intrusive activity. In addition, the user can examine the signature database, and quickly determine which intrusive activity the misuse detection system is programmed to alert on. Another benefit is that the misuse detection system begins protecting your network immediately upon installation. One final benefit is that the system is easy to understand. When an alarm fires, the user can relate this directly to a specific type of activity occurring on the network.

Along with the numerous benefits, misuse detection systems also have disadvantages. One of the biggest problems is maintaining state information for signatures in which the intrusive activity covers multiple discrete events (that is, the complete attack signature occurs in multiple packets on the network). Another drawback is that your misuse detection system must have a signature defined for all of the possible attacks that an attacker may launch against your network. This leads to the necessity for frequent signature updates to keep the signature database of your misuse detection system up-to-date. One final problem with misuse detection systems is that someone may set up the misuse detection system in their lab and intentionally try to find ways to launch attacks that bypass detection by the misuse detection system.

The other way to classify IDS is by monitoring location:

- Network based IDS (NIDS) sit behind the firewall, on the demilitarized zone (DMZ) or the private network and sniff packets in *promiscuous* mode invisible to the attacker. It monitors and analyzes packets and can use anomaly or misuse detection techniques. While the firewall screens out unwanted traffic, the NIDS will alert to what is "leaking" through the firewall. NIDS need to keep up with the high volume of traffic or else it could miss attacks. High speed is also essential for low latency. Thus, it's usually available as dedicated hardware appliances.
- Host based IDS (HIDS) software is run on each host. The software monitors and detects user and operating system activity and logs. Attacks on a given host are detected using misuse detection. HIDS have a closer and deeper look at the activity of attack tools on the host and should be employed on Web, DNS servers and target hosts.

Although there are various categories of IDS, IDS evasion techniques have also become sophisticated.

Evasion techniques are modifications made to attacks in order to prevent detection by an Intrusion Detection System (IDS)[25]. IDS systems work fine for signature-based attacks but the new breed of stealthy attacks go unnoticed.

2.2 – Intrusion Prevention System

The Intrusion Prevention System (IPS) is designed and developed for more active protection to improve upon the IDS and other traditional security solutions. When discussing Intrusion Prevention Systems, there is a formula that is commonly used:

IPS = IDS + Firewall[26]

While this formula provides a useful means of conceptualizing the basic make-up of an Intrusion Prevention System, it is also a simplistic model, concentrating upon form over substance, and more explanation is needed.

A firewall is a system which applies an access control policy. It checks data traffic passing through, and blocks data packets which do not match its security policies. An Intrusion Detection System (IDS) monitors network or system performance, looks for behavior contrary to its security policies and recognizable attack signatures, and it triggers alarms accordingly. So, a firewall rejects obvious attacks, while suspicious traffic will pass through. In turn, the IDS monitors all the data within the network, notifying the network administrator of attacks at a point where the attack is actually live and inside the network. In other words, neither the IDS nor the firewall is capable of blocking attacks themselves at the point at which an intrusion is identified.

The IPS then, is something more than an IDS plus a firewall. The IPS is designed as an embedded system which creates plenty of filters to prevent different kinds of attacks, such as those from hackers, worms, viruses, DoS and other malicious traffic, in advance so that enterprise networks do not suffer any loss even if the latest security patch has not yet been applied.

2.3 – IDPS implementation methods

The IDPS technologies are divided into four groups based on the type of events that they monitor and the ways in which they are deployed:

- **Network-Based**, which monitors network traffic for particular network segments or devices and analyzes the network and application protocol activity to identify suspicious activity. It can identify many different types of events of interest. It is most commonly deployed at a boundary between networks, such as in proximity to border firewalls or

[25] http://en.wikipedia.org/wiki/Intrusion_detection_system_evasion_techniques

[26] http://www.foursquareinnovations.co.uk/software_development_and_ebusiness_articles/intrusion_prevention_systems_3.html

routers, virtual private network (VPN) servers, remote access servers, and wireless networks.

- **Wireless**, which monitors wireless network traffic and analyzes its wireless networking protocols to identify suspicious activity involving the protocols themselves. It cannot identify suspicious activity in the application or higher-layer network protocols (e.g., TCP, UDP) that the wireless network traffic is transferring. It is most commonly deployed within range of an organization's wireless network to monitor it, but can also be deployed to locations where unauthorized wireless networking could be occurring.
- **Network Behavior Analysis (NBA)**, which examines network traffic to identify threats that generate abnormal traffic flows, such as distributed denial of service (DDoS) attacks, certain forms of malware (e.g., worms, backdoors), and policy violations (e.g., a client system providing network services to other systems). NBA systems are most often deployed to monitor flows on an organization's internal networks, and are also sometimes deployed where they can monitor flows between an organization's networks and external networks (e.g., the Internet, business partners' networks).
- **Host-Based,** which monitors the characteristics of a single host and the events occurring within that host for suspicious activity. Examples of the types of characteristics a host-based IDPS might monitor are network traffic (only for that host), system logs, running processes, application activity, file access and modification, and system and application configuration changes. Host-based IDPSs are most commonly deployed on critical hosts such as publicly accessible servers and servers containing sensitive information.

Some forms of IDPS are more mature than others because they have been in use much longer. Network-based IDPS and some forms of host-based IDPS have been commercially available for over ten years. Network behavior analysis software is a somewhat newer form of IDPS that evolved in part from products created primarily to detect DDoS attacks, and in part from products developed to monitor traffic flows on internal networks. Wireless technologies are a relatively new type of IDPS, developed in response to the popularity of wireless local area networks (WLAN) and the growing threats against WLANs and WLAN clients.

2.4 – IDPS Components

The typical components in an IDPS solution are as follows:

Sensor or Agent. Sensors and agents monitor and analyze activity. The term *sensor* is typically used for IDPSs that monitor networks, including network-based, wireless, and network behavior analysis technologies. The term *agent* is typically used for host-based IDPS technologies.

Management Server. A *management server* is a centralized device that receives information from the sensors or agents and manages them. Some management servers perform analysis on the event information that the sensors or agents provide and can identify events that the individual sensors or agents cannot. Matching event information from multiple sensors or agents, such as finding events triggered by the same IP address, is known as *correlation*. Management servers are available as both appliance and software-only products. Some small IDPS deployments do not use any management servers, but most IDPS deployments do. In larger IDPS deployments, there are often multiple management servers, and in some cases there are two tiers of management servers.

Database Server. A *database server* is a repository for event information recorded by sensors, agents, and/or management servers. Many IDPSs provide support for database servers.

Console. A *console* is a program that provides an interface for the IDPS's users and administrators. Console software is typically installed onto standard desktop or laptop computers. Some consoles are used for IDPS administration only, such as configuring sensors or agents and applying software updates, while other consoles are used strictly for monitoring and analysis. Some IDPS consoles provide both administration and monitoring capabilities.

3 – Host-based Intrusion Detection and Prevention Systems (HIDPS)

3.1 – What is HIPS and how it works?

"A HIPS is like an airport security checkpoint. A variety of technologies look for multiple types of threats, including checking bags and people for weapons and chemical residues, and utilizing facial recognition software to identify wanted individuals. Still to prevent attacks you need some idea of what to look for"[27]

A host-based IDPS monitors the characteristics of a single host and the events occurring within that host for suspicious activity. Examples of the types of characteristics a host-based IDPS might monitor are:

- wired and wireless network traffic (only for that host);
- system logs, running processes, file access and modification;
- system and application configuration changes.

3.2 – Host-based IDPS deployment

Most host-based IDPSs have detection software known as *agents* installed on the hosts of interest. Each agent monitors activity on a single host and if IDPS capabilities are enabled, also

[27] http://www.iss.net/documents/whitepapers/ISS_Preemptive_Host_Protection_Whitepaper.pdf

performs prevention actions. The agents transmit data to management servers, which may optionally use database servers for storage. Consoles are used for management and monitoring.

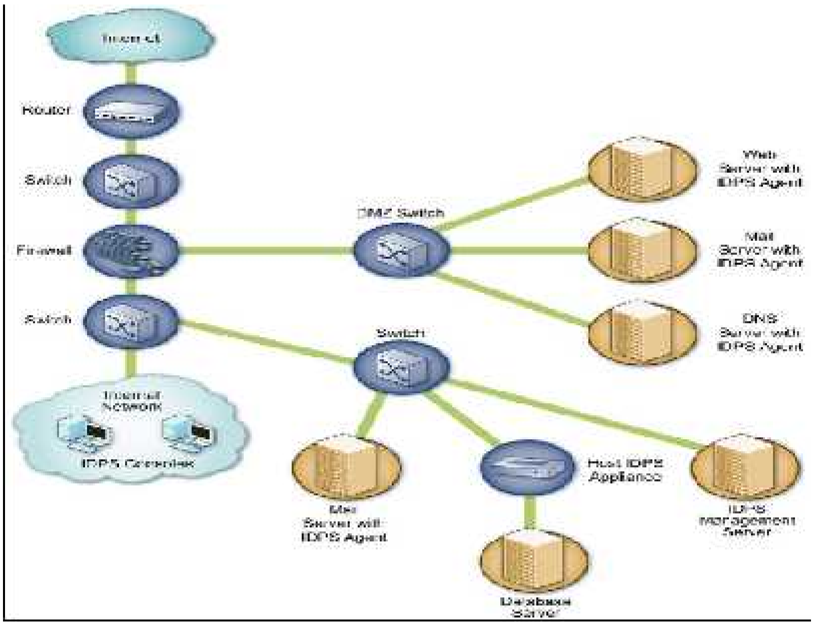
Some host-based IDPS products use dedicated appliances running agent software instead of installing agent software on individual hosts. Each appliance is positioned to monitor the network traffic going to and from a particular host. Technically, these appliances could be considered network-based IDPSs, because they are deployed inline to monitor network traffic. However, they usually monitor activity for only one specific type of application, such as a Web server or database server, so they are more specialized than a standard network-based IDPS. Also, the software running on the appliance often has the same or similar functionality as the host-based agents.

Each agent is typically designed to protect one of the following:

- **A server.** Besides monitoring the server's operating system (OS), the agent may also monitor some common applications;
- **A client host (desktop or laptop).** Agents designed to monitor users' hosts usually monitor the OS and common client applications such as e-mail clients and Web browsers;
- **An application service.** Some agents perform monitoring for a specific application service only, such as a Web server program or a database server program. This type of agent is also known as an *application-based IDPS*.

Most products do not have agents for other types of hosts, such as network devices (e.g., firewalls, routers, switches).

The network architecture for host-based IDPS deployments is typically very simple. Because the agents are deployed to existing hosts on the organization's networks, the components usually communicate over those networks instead of using a separate management network. Most products encrypt their communications, preventing eavesdroppers from accessing sensitive information. Appliance-based agents are typically deployed inline immediately in front of the hosts that they are protecting. Next figure shows an example of a host-based IDPS deployment architecture.



Host-Based IDPS Agent Deployment Architecture Example

Host-based IDPS agents are most commonly deployed to critical hosts such as publicly accessible servers and servers containing sensitive information. However, because agents are available for various server and desktop/ laptop operating systems, as well as specific server applications, organizations could potentially deploy agents to most of their servers and desktops/ laptops. Some organizations use host-based IDPS agents primarily to analyze activity that cannot be monitored by other security controls. For example, network-based IDPS sensors cannot analyze the activity within encrypted network communications, but host-based IDPS agents installed on endpoints can see the unencrypted activity.

Organizations should consider the following additional criteria when selecting agent locations:

- the cost to deploy, maintain, and monitor the agents
- OSs and applications supported by the agents
- the importance of the host's data or services
- The ability of the infrastructure to support the agents (e.g. sufficient network bandwidth to transfer alert data from the agents to centralized servers and to transfer software and policy updates from the centralized servers to the agents).

3.3 – What is a shim?

To provide intrusion prevention capabilities, most IDPS agents change the internal architecture of the hosts on which they are installed. This is typically done through a *shim*, which is a layer of code placed between existing layers of code. A shim intercepts data at a point where it would normally be passed from one piece of code to another. The shim can then analyze the data and determine whether or not it should be allowed or denied. Host-based IDPS agents may use shims for several types of resources, including network traffic, filesystem activity, system calls, Windows registry activity, and common applications (e.g., e-mail, Web).

Some host-based IDPS agents do not modify the host architecture. Instead, they monitor activity without shims, or they analyze the artifacts of activity, such as log entries and file modifications. Although less intrusive to the host, reducing the possibility of the IDPS interfering with the host's normal operations, these methods are also generally less effective at detecting threats and often cannot perform any prevention actions.

One of the important decisions in selecting a host-based IDPS solution is whether to install agents on hosts or use agent-based appliances. From a detection and prevention point of view, installing agents on hosts is generally preferable because the agents have direct access to the hosts' characteristics, often allowing them to perform more comprehensive and accurate detection and prevention. However, agents often support only a few common OSs; if a host does not use a supported OS, an appliance can be deployed instead. Another reason to use an appliance instead of installing an agent on a host is performance; if an agent would negatively impact the performance of the monitored host too much, it might be necessary to offload the agent's functions to an appliance.

3.4 – Benefits of implementing a Host Intrusion Prevention System

First and foremost, enterprise and home users now have increased protection from unknown zero-day attacks. Because HIPSs use anomaly detection, there is a better chance that it will stop an attack trying to exploit an unknown vulnerability as opposed to traditional protective measures.

A second benefit of using HIPS is that the need to be running and managing multiple security applications such as anti-virus, anti-spyware, and software firewalls to protect your PC may be combined into one. Depending on the environment, you may only need to implement HIPS on the workstation, like Proventia Desktop. Users now have a firewall, anti-virus, anti-spyware protection, and application control in one application.

The best part is not having to worry about making sure that multiple security applications work together correctly. Another benefit is Total Cost of Ownership (TCO). In implementing a HIPS only one security application may need to be purchased instead of three *(again depending on your environment).* Therefore, instead of paying three license and support maintenance costs every year there is only one that will need to be paid.

3.5 – Implementing a HIPS

Implementing an Host-based IPS takes a lot of time and preparation. Whoever will be implementing and configuring the HIPS should have a thorough understanding of how the network is designed, know what applications are being used and how they function. Some applications may need to write to the root of the primary drive, others may need to communicate over specific ports. Whatever the case may be, a thorough understanding of the network is needed or serious problems could happen while implementing the HIPS.

Most HIPS systems are managed by a centralized management console. Security Administrator control what the agents will deny and permit on each workstation. Some specific things that are essential to know before configuring the agent's rules and policies are:

- What ports do the applications communicate over?
- Who initiates the communication between the clients and servers? Only the servers, only the clients, or both?
- What protocols do the applications use - UDP, TCP, ICMP, etc?
- Are there branches or remote sites that need to communicate with workstations at the main branch? If so, what IP addresses will need to be permitted?

It is also a good idea to check if the HIPS that will be used comes with its own anti-virus. If it does, determine if the HIPS is able to run concurrently with the anti-virus/ antispyware already being used. Most HIPS systems integrate their own anti-virus/anti-spyware and most likely will not be able to run concurrently with another vendor's anti-virus/anti-spyware software.

Depending on which HIPS is chosen, make sure that it is flexible enough to have the ability to filter by different rules, and control the agents.

3.6 – Configuring a HIPS

Once there is a complete understanding of the applications and their communications the groups can be created. If everybody will have the same rules and policies then only one group is needed. For companies that have mobile or remote workers, HIPS systems are very useful. HIPSs are able to provide relatively the same level of protection as internal workstations.

One precaution to make sure that the agent service never stops is to set the agent protection to prevent unauthorized shutdown of the agent services. This ensures that only administrators can shut down the agent service with a password, if the password option is set.

If the password option is not set, anybody with administrative privileges will be able to shut down the service. Additionally, there are options to encrypt the configuration files and prevent unauthorized changes to the agent files. If the password option is set, the agent password must be entered every time a change needs to be made. These options need to be set for each group and each group’s password can be different. This works out well for those that would like to delegate control. If only one person will be managing the agents, it is probably not a good idea to set different passwords for each group.

An important thing to remember when configuring the HIPS agents is to implement the most restrictive policy allowed and then permit only what is needed. Never permit everything and then scale back. There is always a chance of missing something leaving the workstations vulnerable to attack.

The next step is to create and assign each group’s rules and policies based on the applications that they use. If all the workstations use the same applications or are somewhat similar with just a few additional applications on some of them, copy the policies and modify them so that it will fit specific groups. After the groups have been created with their respective rules and policies start setting up the test environment.

Having multiple workstations will be very useful. During testing, be sure to test all communication that occurs on the workstation within the test environment to avoid any interruptions.

The workstations being used for testing should replicate production workstations. Large enterprises most likely will not be able to replicate all the different workstations, but still need to be sure not to miss any applications being used. If something is missed the HIPS will stop any application not permitted and may cause some disruptions or down time.

Once all applications have been tested thoroughly and made sure that everything will run correctly, start to deploy the agents to your production workstations. While deploying the agents do not deploy them to all the workstations at once. Deploy the agents to workstations that will be least affected if something goes wrong. Also, only deploy the agents to one or two workstations in each department at each site (if there are remote sites). Let them run for a few days or however long is needed to verify that all communication and applications have been permitted. After everything is running smoothly, the agents should be deployed to the rest of the workstations.

3.7 – Tuning HIPS Alerts

Tuning HIPS alerts will take some time as false-positives have always been a problem with IDS and IPS sensors. When tuning any IPS or IDS alerts, the first thing to do is baseline the alerts. Investigate the alerts and find out which alerts are relevant and those that are not.

Begin with the high severity alerts first and then work your way down to the medium and low severity alerts. There may be some alerts that will be triggered and will not apply to the environment which can be excluded quickly.

For example, if there are a lot of Windows attacks being triggered and there are only Linux or Unix operating systems being used, obviously these are false-positives that do not need to be investigated and the signatures could be turned off.

So if an application continuously sets off an alert, but you know it is a false positive, use the option to turn that particular signature off or change the severity on the signature. The general goal of tuning alerts is to not waste time investigating false positives that are triggered repeatedly.

Another feature to help with tuning alerts is viewing events by groups, if groups have been created. Being able to view alerts and events by groups gives an idea of what really may be happening with a certain agent or group.

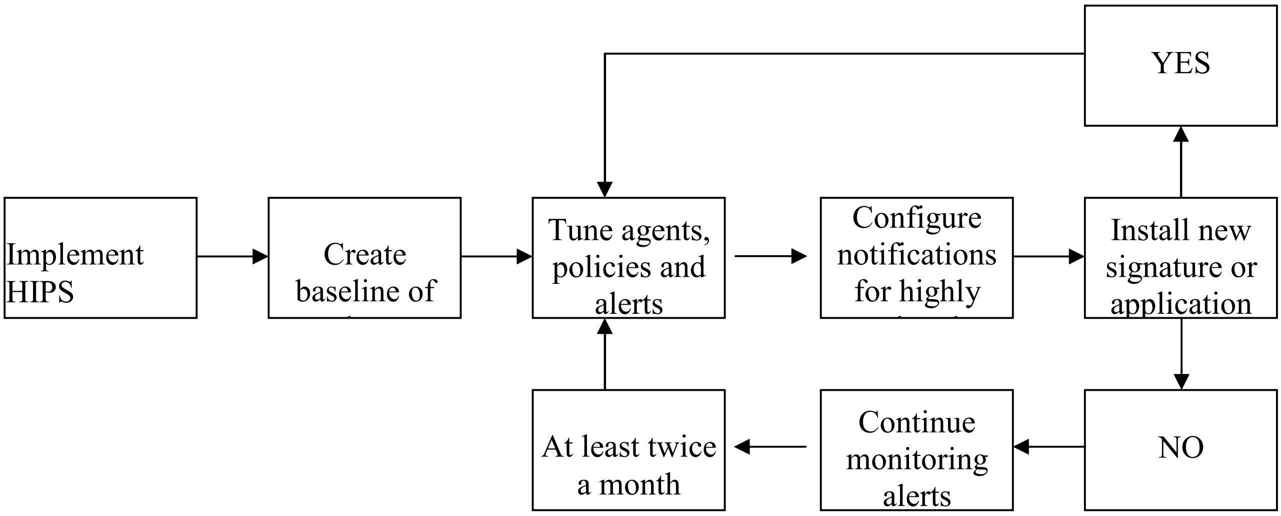
The next step after configuring the alerts is to configure the notifications. It is critical to know when high severity alerts are being triggered. If somebody is attempting a DoS attack on a workstation and high severity alerts are being triggered, it is imperative to know immediately in order to determine the situation and respond appropriately. Therefore, the agent needs to be configured to email a notification when high severity events are triggered.

On the other hand, notifications can be very annoying if the alerts are configured incorrectly. Therefore, administrators need to be careful when configuring the alerts.

Once all of the agents have been deployed and the alerts are tuned, there is the never ending task of monitoring all the alerts, agents, and event logs. Alerts will also have to be continuously tuned as well as the agents to work with new vulnerability patches and new programs. Each time a new program is added or a new vulnerability patch is released, test it in a non-production environment and determine what needs to be permitted, just as with any other program or patch being added.

Understanding the applications and knowing what is traversing the network is an essential part in tuning alerts. The more that is known about the network, the faster the alerts can be tuned. Also, depending on how many and what kind of applications are running will determine how long it will take to tune. The more applications being run, the more false-positives may be triggered. As new vulnerabilities arise, new signatures will follow with the probability that those new signatures will trigger false-positives from your applications. Be sure to check that the new signatures do not overwrite the settings on any current signatures.

To solve the false-positive problem, continue doing what is already being done. Test any new applications or patches in a test environment and see what alerts they trigger. Then make the necessary changes to reduce the false-positives. To help understand the overall process of implementing, configuring, and tuning a HIPS, next figure shows the steps from beginning to end.


YES
Implement HIPS
Create baseline of
Tune agents, policies and alerts
Configure notifications for highly
Install new signature or application
At least twice a month
Continue monitoring alerts
NO

CONCLUSION

With a lot of vulnerabilities out there, security administrators need to constantly mitigate the risks associated with the constantly changing environments and applications being introduced. Host Intrusion Prevention Systems are an invaluable tool, but we need to remember that it is not the "silver bullet" for workstation security. “They can be a great addition to a solid, layered defense including firewalls, NIPSs, IDSs, and anti-virus applications among other things, but should not replace them.”[28]

As each host protection technology possesses strengths and weaknesses, selecting just one technology for comprehensive host protection results in too much risk to the host environment. Any single technology represents a singular point of failure. Employing the different technologies in concert brings risk exposure to threats down to acceptable levels. In addition, combining multiple host protection technologies into a single host protection solution significantly reduces management costs.

Whatever the choice, whether it is host-based, network-based, or a hybrid of the two, it is clear that using intrusion detection and prevention systems is an important and necessary tool in the security manager's arsenal.

[28] http://netsecurity.about.com/cs/firewallbooks/a/aa050804.htm

REFERENCES

1. Intrusion Detection Systems, viewed on 10.06.2010, http://www.ciscopress.com/articles/article.asp?p=25334
2. Dinesh Sequeira, - Intrusion Prevention Systems – Security's silver bullet?, viewed on 12.06.2010 http://www.sans.org/reading_room/whitepapers/detection/intrusion-prevention-systems-securitys-silver-bullet_366
3. INTRUSION PREVENTION SYSTEMS: AN INTRODUCTION..., viewed on 10.06.2010,http://www.foursquareinnovations.co.uk/software_development_and_ebusiness_articles/intrusion_prevention_systems_2.html
4. Defining the Rules for Preemptive Host Protection, viewed on 12.06.2010, http://www.iss.net/documents/whitepapers/ISS_Preemptive_Host_Protection_Whitepaper.pdf
5. Host Intrusion Prevention Systems and Beyond, viewed on 12.06.2010, http://www.sans.org/reading_room/whitepapers/intrusion/host-intrusion-prevention-systems_32824
6. Things to look for in this last line of defense. Host-Based Intrusion Prevention, viewed on 12.06.2010, http://netsecurity.about.com/cs/firewallbooks/a/aa050804.htm
7. NIST Special Publication SP800-31, viewed on 10.06.2010, http://csrc.ncsl.nist.gov/publications/nistpubs/800 800-31/sp800 sp800-31.pdf
8. NIST Special Publication SP800-94, viewed on 10.06.2010, http://csrc.nist.gov/publications/nistpubs/800-94/SP800-94.pdf
9. Wikipedia, http://en.wikipedia.org

INTERNET PROTOCOL SECURITY (IPsec)

Maj. Laurenţiu SPĂTĂROAIA

INTRODUCTION

IPsec is a framework of open standards for ensuring private communications over public networks. It has become the most common network layer security control, typically used to create a Virtual Private Network (VPN)[29]. A VPN is a virtual network built on top of existing physical networks that can provide a secure communications mechanism for data and control information transmitted between networks. VPNs are used most often to protect communications carried over public networks such as the Internet.

A VPN can provide several types of data protection, including confidentiality, integrity, data origin authentication, replay protection and access control. Although VPNs can reduce the risks of networking, they cannot totally eliminate them. For example, a VPN implementation may have flaws in algorithms or software, or a VPN may be set up with insecure configuration settings and values. Both of these flaws can be exploited by attackers.

I. NETWORK LAYER SECURITY

TCP/IP is widely used throughout the world to provide network communications. TCP/IP communications are composed of four layers that work together. When a user wants to transfer data across networks, the data is passed from the highest layer through intermediate layers to the lowest layer, with each layer adding additional information. The lowest layer sends the accumulated data through the physical network; the data is then passed up through the layers to its destination. Essentially, the data produced by a layer is encapsulated in a larger container by the layer below it.

The four TCP/IP layers, from highest to lowest, are shown in Table 1:

Application Layer. This layer sends and receives data for particular applications, such as *Domain Name System* (DNS), *HyperText Transfer Protocol* (HTTP), and *Simple Mail Transfer Protocol* (SMTP).
Transport Layer. This layer provides connection-oriented or connectionless services for transporting application layer services between networks. The transport layer can

[29] National Institute of Standards and Technology – "*Special Publication 800-77, Guide to IPsec VPNs – Recommendations of National Institute of Standards and Technology*", 2005;

optionally assure the reliability of communications. *Transmission Control Protocol* (TCP) and *User Datagram Protoco*l (UDP) are commonly used transport layer protocols.
Network Layer. This layer routes packets across networks. Internet Protocol (IP) is the fundamental network layer protocol for TCP/IP. Other commonly used protocols at the network layer are *Internet Control Message Protocol* (ICMP) and *Internet Group Management Protocol* (IGMP).
Data Link Layer. This layer handles communications on the physical network components. The best-known data link layer protocol is *Ethernet.*

Table 1. TCP/IP Layers

Security controls exist for network communications at each layer of the TCP/IP model. As previously explained, data is passed from the highest to the lowest layer, with each layer adding more information. Because of this, a security control at a higher layer cannot provide full protection for lower layers, because the lower layers perform functions of which the higher layers are not aware.

Controls at the network layer apply to all applications and are not application-specific. For example, all network communications between two hosts or networks can be protected at this layer without modifying any applications on the clients or the servers. In many environments, network layer controls such as IPsec provide a much better solution than transport or application layer controls because of the difficulties in adding controls to individual applications. Network layer controls also provide a way for network administrators to enforce certain security policies. Another advantage of network layer controls is that since IP information (e.g., IP addresses) is added at this layer, the controls can protect both the data within the packets and the IP information for each packet. However, network layer controls provide less control and flexibility for protecting specific applications than transport and application layer controls.

Internet Protocol Security (*IPsec*) has emerged as the most commonly used network layer security control for protecting communications. IPsec is a framework of open standards for ensuring private communications over IP networks. Depending on how IPsec is implemented and configured, it can provide any combination of the following types of protection:

- ***Confidentiality.*** IPsec can ensure that data cannot be read by unauthorized parties. This is accomplished by encrypting data using a cryptographic algorithm and a secret key—a value known only to the two parties exchanging data. The data can only be decrypted by someone who has the secret key.
- ***Integrity.*** IPsec can determine if data has been changed (intentionally or unintentionally) during transit. The integrity of data can be assured by generating a

message authentication code (MAC) value, which is a cryptographic checksum of the data. If the data is altered and the MAC is recalculated, the old and new MACs will differ.

- ***Peer Authentication***. Each IPsec endpoint confirms the identity of the other IPsec endpoint with which it wishes to communicate, ensuring that the network traffic and data is being sent from the expected host.
- ***Replay Protection***. The same data is not delivered multiple times, and data is not delivered grossly out of order. However, IPsec does not ensure that data is delivered in the exact order in which it is sent.
- ***Traffic Analysis Protection***. A person monitoring network traffic does not know which parties are communicating, how often communications are occurring, or how much data is being exchanged. However, the number of packets being exchanged can be counted. The IPsec protocols were developed within the IPsec Working Group of the Internet Engineering Task Force (IETF). They are defined in 2 types of documents: Request for Comment (RFC), which are accepted standards; and Internet-Drafts, which are working documents that may become RFCs.
- ***Access Control***. IPsec endpoints can perform filtering to ensure that only authorized IPsec users can access particular network resources. IPsec endpoints can also allow or block certain types of network traffic, such as allowing Web server access but denying file sharing.

II. IPSEC FUNDAMENTALS

II.1. Virtual Private Network (VPN)

A *Virtual Private Network* (*VPN*) is a group of two or more computer systems connected "securely" over a public network. VPNs can be installed between an individual machine and a private network (*remote user-to-site*) or between private networks (*site-to-site*). Security features differ from product to product, but most security experts[30] agree that VPNs should include encryption, strong authentication of remote users or hosts, and mechanisms for hiding or masking information about the private network topology from potential attackers on the public network.

There are three primary models for VPN architectures, as follows:

[30] European Centre for Medium-Range Weather Forecasts – "*IPsec Feasibility Study*", 2003;

- ***Gateway-to-gateway.*** This model protects communications between two specific networks, such as an organization's main office network and a branch office network, or two business partners' networks.
- ***Host-to-gateway.*** This model protects communications between one or more individual hosts and a specific network belonging to an organization. The host-to-gateway model is most often used to allow hosts on unsecured networks, such as travelling employees and telecommuters, to gain access to internal organizational services, such as the organization's e-mail and Web servers.
- ***Host-to-host.*** A host-to-host architecture protects communication between two specific computers. It is most often used when a small number of users need to use or administer a remote system that requires the use of inherently insecure protocols.

NIST's requirements and recommendations for the configuration of IPsec VPNs are:

- If any of the information that will traverse a VPN should not be seen by non-VPN users, then the VPN must provide confidentiality protection (encryption) for that information.
- A VPN must use a FIPS-approved encryption algorithm. *AES-CBC* (*AES* in *Cipher Block Chaining* mode) with a 128-bit key is highly recommended; *Triple DES* (*3DES-CBC*) is also acceptable[31].
- A VPN must always provide integrity protection.
- A VPN must use a FIPS-approved integrity protection algorithm[32]. HMAC-SHA-1 is highly recommended. HMAC-MD5 also provides integrity protection, but it is not a FIPS-approved algorithm.
- A VPN should provide replay protection.
- For IKEv1, *IKE Security Associations* (S*As*) should have a lifetime no greater than 24 hours (86400 seconds) and IPsec SAs should have a lifetime no greater than 8 hours (28800 seconds). For IKEv2, IKE SAs should be re-keyed after at most 24 hours and child SAs should be re-keyed after at most 8 hours.
- The *Diffie-Hellman* (*DH*) group used to establish the secret keying material for IKE and IPsec should be consistent with current security requirements. The larger DH groups will result in increased processing time.

[31] The *Data Encryption Standard* (*DES*) is also an encryption algorithm; since it has been successfully attacked, it should not be used;

[32] National Institute of Standards and Technology – "*FIPS 198-1 – The Keyed-Hash Message Authentication Code (HMAC)*", 2008, available at http://csrc.nist.gov/publications/fips/fips198-1/ FIPS-198-1_final.pdf.

II.1.1. Gateway-to-Gateway Architecture

IPsec-based VPNs are often used to provide secure network communications between two networks. This is typically done by deploying a VPN gateway onto each network and establishing a VPN connection between the two gateways. Traffic between the two networks that needs to be secured passes within the established VPN connection between the two VPN gateways. The VPN gateway may be a dedicated device that only performs VPN functions, or it may be part of another network device, such as a firewall or router. Figure 2 shows an example of an IPsec network architecture that uses the gateway-to-gateway model to provide a protected connection between the two networks.

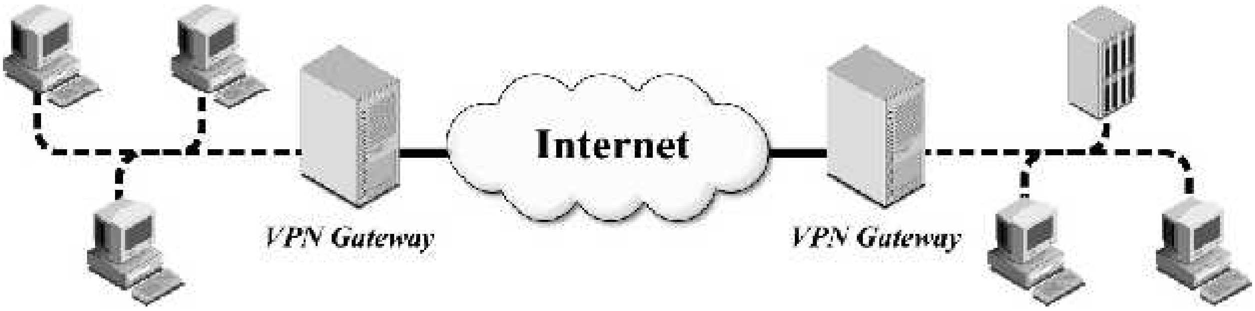


Figure 1. Gateway-to-Gateway Architecture example

This model is relatively simple to understand. To facilitate VPN connections, one of the VPN gateways issues a request to the other to establish an IPsec connection. The two VPN gateways exchange information with each other and create an IPsec connection. Routing on each network is configured so that as hosts on one network need to communicate with hosts on the other network, their network traffic is automatically routed through the IPsec connection, protecting it appropriately. A single IPsec connection establishing a tunnel between the gateways can support all communications between the two networks, or multiple IPsec connections can each protect different types or classes of traffic.

Figure 2 illustrates that a gateway-to-gateway VPN does not provide full protection for data throughout its transit. In fact, the gateway-to-gateway model only protects data between the two gateways, as denoted by the solid line. The dashed lines indicate that communications between VPN clients and their local gateway, and between the remote gateway and destination hosts (e.g., servers) are not protected.

The other VPN models provide protection for more of the transit path. The gateway-to-gateway model is most often used when connecting two secured networks, such as linking a branch office to headquarters over the Internet. Gateway-to-gateway VPNs often replace more costly private wide area network (WAN) circuits.

The gateway-to-gateway model is the easiest to implement, in terms of user and host management. Gateway-to-gateway VPNs are typically transparent to users, who do not need to perform separate authentication just to use the VPN. Also, the users' systems and the target hosts (e.g., servers) should not need to have any VPN client software installed, nor should they require any reconfiguration, to be able to use the VPN.

II.1.2. Host-to-Gateway Architecture

An increasingly common VPN model is the host-to-gateway model, which is most often used to provide secure remote access. The organization deploys a VPN gateway onto their network; each remote access user then establishes a VPN connection between the local computer (host) and the VPN gateway. As with the gateway-to-gateway model, the VPN gateway may be a dedicated device or part of another network device. Figure 3 shows an example of an IPsec host-to-gateway architecture that provides a protected connection for the remote user.

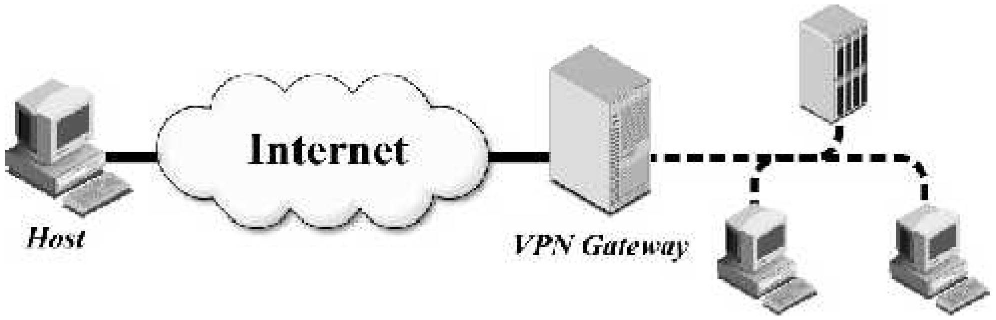


Figure 2. Host-to-Gateway Architecture example

In this model, IPsec connections are created as needed for each individual VPN user. Remote users' hosts have been configured to act as IPsec clients with the organization's IPsec gateway. When a remote user wishes to use computing resources through the VPN, the host initiates communications with the VPN gateway. The user is typically asked by the VPN gateway to authenticate before the connection can be established. The VPN gateway can perform the authentication itself or consult a dedicated authentication server. The client and gateway exchange information, and the IPsec connection is established. The user can now use the organization's computing resources, and the network traffic between the user's host and the VPN gateway will be protected by the IPsec connection. Traffic between the user and systems not controlled by the organization can also be routed through the VPN gateway; this allows IPsec protection to be applied to this traffic as well if desired.

As shown in Figure 3, the host-to-gateway VPN does not provide full protection for data throughout its transit. The dashed lines indicate that communications between the gateway and the destination hosts (e.g., servers) are not protected. The host-to-gateway model is most often used when connecting hosts on unsecured networks to resources on secured networks, such as linking travelling employees around the world to headquarters over the Internet. Host -to-gateway VPNs often replace dial-up modem pools. The host-to-gateway model is somewhat complex to implement and maintain in terms of user and host management. Host -to-gateway VPNs are typically not transparent to users because they must authenticate before using the VPN. Also, the users' hosts need to have VPN client software configured[33].

II.1.3. Host-to-Host Architecture

The least commonly used VPN architecture is the host-to-host model, which is typically used for special purpose needs, such as system administrators performing remote management of a single server. In this case, the organization configures the server to provide VPN services and the system administrators' hosts to act as VPN clients. The system administrators use the VPN client when needed to establish encrypted connections to the remote server. Figure 4 shows an example of an IPsec network architecture that uses the host-to-host model to provide a protected connection to a server for a user.

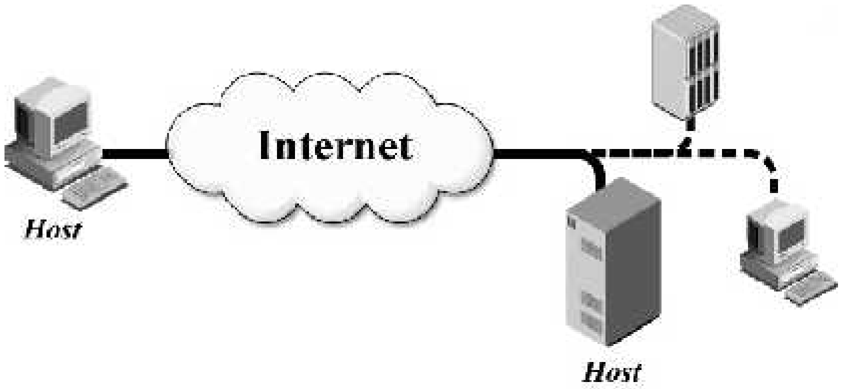


Figure 3. Host-to-Host Architecture example

In this model, IPsec connections are created as needed for each individual VPN user. Users' hosts have been configured to act as IPsec clients with the IPsec server. When a user wishes to use resources on the IPsec server, the user's host initiates communications with the IPsec server. The user is asked by the IPsec server to authenticate before the connection can be established. The client and server exchange information, and if the authentication is successful, the IPsec connection is established. The user can now use the server, and the network traffic between the user's host and the server will be protected by the IPsec connection.

[33] Most (but not all) PC operating systems have built-in VPN clients, so it may be necessary to install VPN clients on some hosts.

As shown in Figure 4, the host-to- host VPN is the only model that provides protection for data throughout its transit. This can be a problem, because network-based firewalls, intrusion detection systems, and other devices cannot be placed to inspect the decrypted data, which effectively circumvents certain layers of security. The host-to- host model is most often used when a small number of trusted users need to use or administer a remote system that requires the use of insecure protocols (e.g., a legacy system) and can be updated to provide VPN services. The host-to-host model is resource-intensive to implement and maintain in terms of user and host management. Host-to-host VPNs are not transparent to users because they must authenticate before using the VPN. Also, all user systems and servers that will participate in VPNs need to have VPN software installed and/or configured.

II.1.4. Model Comparison

Table 2 provides a brief comparison of the three VPN architecture models:

Feature	Gateway-to-Gateway	Host-to-Gateway	Host-to-Host
Provides protection between client and local gateway	No	N/A (client is VPN endpoint)	N/A (client is VPN endpoint)
Provides protection between VPN endpoints	Yes	Yes	Yes
Provides protection between remote gateway and remote server (behind gateway)	No	No	N/A (server is VPN endpoint)
Transparent to users	Yes	No	No
Transparent to users' systems	Yes	No	No
Transparent to servers	Yes	Yes	No

II.2. IPsec

IPsec is a collection of protocols that assist in protecting communications over IP networks[34]. IPsec protocols work together in various combinations to provide protection for communications. It can protect either the entire IP datagram or only the upper-layer protocols. IPsec is an end-to-end security protocol: all the functionality and intelligence of the VPN connection reside at the end points, either in a gateway or in the end-host. The service provider's

[34] RFC 2401, *Security Architecture for the Internet Protocol*, provides an overview of IPsec. The RFC is available for download at http://www.ietf.org/rfc/rfc2401.txt.

IP network is not aware of the existence of the IP VPN, as tunnelling technologies ensure the transport of application data by encapsulation. The source address and the destination address of these packets are the IP addresses of the end points of the tunnel. They are then routed as any normal IP packets through the shared IP network.

In the past, several IP tunnelling protocols have been deployed. Over the last years, however, IPsec has become the predominant IP tunnelling protocol and is currently the technology of choice when implementing site-to-site connectivity over a public network. IPsec was initially developed to ensure private communications over public IP networks. The protocol supports two main security functions:

- *Authentication: ensuring the authenticity and the integrity of the whole IP packet;*
- *Encryption*: *ensuring* the confidentiality of the payload.

Through IPsec it is possible to define a tunnel between two gateways. An IPsec gateway would typically be an access router or a firewall on which the IPsec protocol is implemented. IPsec gateways sit between the user's private network and the carrier's shared network.

IPsec tunnels are established dynamically and released when they are not in use. To establish an IPsec tunnel, two gateways must authenticate themselves and define which security algorithms and keys they will use for the tunnel. The entire original IP packet is encrypted and wrapped inside IPsec authentication and encryption headers. This becomes the payload of a new IP packet whose source and destination IP addresses are the public network IP addresses of the IPsec gateways. This ensures the logical separation between VPN traffic flows in a shared IP network. Traditional IP routing is then used between the tunnel end points.

IPsec achieves these objectives by using two traffic security protocols:

- the *Authentication Header* (*AH*)[35], which provides data integrity and the *Encapsulation Security Payload* (*ESP*), which provides data integrity and data confidentiality;

  AH can provide integrity protection for packet headers and data, but it cannot encrypt them. ESP can provide encryption and integrity protection for packets, but it cannot protect the outermost IP header, as AH can. However, this protection is not needed in most cases. Accordingly, ESP is used much more frequently than AH because of its encryption capabilities, as well as other operational advantages. For a VPN, which requires confidential communications, ESP is the natural choice.
- a cryptographic-key management protocol, the *Internet Key Exchange* (*IKE*), which is used to negotiate IPsec connection settings, authenticate endpoints to each other,

[35] AH is IP protocol number 51. The AH version 2 standard is defined in RFC 2402, *IP Authentication Header*, available at http://www.ietf.org/rfc/rfc2402.txt.

define the security parameters of IPsec-protected connections, negotiate secret keys and manage, update, and delete IPsec-protected communication channels.

- Optionally, *IP Payload Compression Protocol* (*IPComp*). IPsec can use IPComp to compress packet payloads before encrypting them

II.2.1. Authentication Header

AH has two modes: *transport and tunnel*. In tunnel mode, AH creates a new IP header for each packet; in transport mode, AH does not create a new IP header. In IPsec architectures that use a gateway, the true source or destination IP address for packets must be altered to be the gateway's IP address. Because transport mode cannot alter the original IP header or create a new IP header, transport mode is generally used in host-to-host architectures.

AH adds a header to each packet. Each AH header is composed of six fields:

- ***Next Header***. This field contains the IP protocol number for the next packet payload. In tunnel mode, the payload is an IP packet, so the Next Header value is set to 4 for IP-in-IP. In transport mode, the payload is usually a transport-layer protocol, often TCP (protocol number 6) or UDP (protocol number 17).
- ***Payload Length***. This field contains the length of the payload in 4-byte increments, minus 2.
- ***Reserved***. This value is reserved for future use, so it should be set to 0.
- ***Security Parameters Index*** (***SPI***). Each endpoint of each IPsec connection has an arbitrarily chosen SPI value, which acts as a unique identifier for the connection. The recipient uses the SPI value, along with the destination IP address and (optionally) the IPsec protocol type (in this case, AH), to determine which Security Association (SA) is being used. This tells the recipient which IPsec protocols and algorithms have been applied to the packet.
- ***Sequence Number***. Each packet is assigned a sequential sequence number, and only packets within a sliding window of sequence numbers are accepted. This provides protection against replay attacks because duplicate packets will use the same sequence number. This also helps to counter *denial of service* attacks because old packets that are replayed will have sequence numbers outside the window, and will be dropped immediately without performing any more processing.
- ***Authentication Information***. This field contains the MAC output. The recipient of the packet can recalculate the MAC to confirm that the packet has not been altered in transit.

II.2.2. Encapsulating Security Payload (ESP)

ESP[36] is the second core IPsec security protocol. In the initial version of IPsec, ESP provided only encryption for packet payload data. In the second version of IPsec, ESP became more flexible. It can perform authentication to provide integrity protection, although not for the outermost IP header. Also, ESP's encryption can be disabled through the *Null ESP Encryption Algorithm*. Therefore, in all but the oldest IPsec implementations, ESP can be used to provide only encryption, encryption and integrity protection, or only integrity protection.

ESP has two modes: *transport* and *tunnel*. In *tunnel mode*, ESP creates a new IP header for each packet. The new IP header lists the endpoints of the ESP tunnel (such as two IPsec gateways) as the source and destination of the packet. Because of this, tunnel mode can be used with all three VPN architecture models. Tunnel mode can encrypt and/or protect the integrity of both the data and the original IP header for each packet as well. Encrypting the data protects it from being accessed or modified by unauthorized parties; encrypting the IP header conceals the nature of the communications, such as the actual source or destination of the packet. If authentication is being used for integrity protection, each packet will have an ESP Authentication section after the ESP trailer.

ESP tunnel mode is used far more frequently than ESP transport mode. In *transport mode*, ESP uses the original IP header instead of creating a new one and can only encrypt and/or protect the integrity of packet payloads and certain ESP components, but not IP headers. As with AH, ESP transport mode is generally only used in host-to-host architectures. Also, transport mode is incompatible with NAT. For example, in each TCP packet, the TCP checksum is calculated on both TCP and IP fields, including the source and destination addresses in the IP header.

ESP uses symmetric cryptography to provide encryption for IPsec packets. Accordingly, both endpoints of an IPsec connection protected by ESP encryption must use the same key to encrypt and decrypt the packets. When an endpoint encrypts data, it divides the data into small blocks (for the AES algorithm, 128 bits each), and then performs multiple sets of cryptographic operations (known as *rounds*) using the data blocks and key. Encryption algorithms that work in this way are known as *block cipher algorithms*. When the other endpoint receives the encrypted data, it performs decryption using the same key and a similar process, but with the steps reversed and the cryptographic operations altered. Examples of encryption algorithms used by ESP are *AES-Cipher Block Chaining* (*AES-CBC*), *AES* Counter *Mode* (*AES-CTR*), and *Triple DES* (*3DES*)[37].

[36] ESP is IP protocol number 50. The ESP version 2 standard is defined in RFC 2406, *IP Encapsulating Security Payload (ESP)*, available at http://www.ietf.org/rfc/rfc2406.txt.

[37] AES is described in FIPS 197, *Advanced Encryption Standard (AES)*, available at http://csrc.nist.gov/publications/ fips/fips197/fips-197.pdf.

ESP adds a header and a trailer around each packet's payload. Each ESP header is composed of two fields:

- *SPI*[38]. Each endpoint of each IPsec connection has an arbitrarily chosen SPI value, which acts as a unique identifier for the connection. The recipient uses the SPI value, along with the destination IP address and (optionally) the IPsec protocol type (in this case, ESP), to determine which SA is being used.
- *Sequence Number*. Each packet is assigned a sequential sequence number, and only packets within a sliding window of sequence numbers are accepted. This provides protection against replay attacks because duplicate packets will use the same sequence number. This also helps to thwart denial of service attacks because old packets that are replayed will have sequence numbers outside the window, and will be dropped immediately without performing any more processing.

The next part of the packet is the payload. It is composed of the *payload data*, which is encrypted, and the *initialization vector* (*IV*), which is not encrypted. The IV is used during encryption. Its value is different in every packet, so if two packets have the same content, the inclusion of the IV will cause the encryption of the two packets to have different results. This makes ESP less susceptible to cryptanalysis.

The third part of the packet is the ESP trailer, which contains at least two fields and may optionally include one more:

- *Padding*. An ESP packet may optionally contain padding, which is additional bytes of data that make the packet larger and are discarded by the packet's recipient. Because ESP uses block ciphers for encryption, padding may be needed so that the encrypted data is an integral multiple of the block size. Padding may also be needed to ensure that the ESP trailer ends on a multiple of 4 bytes. Additional padding may also be used to alter the size of each packet, concealing how many bytes of actual data the packet contains. This is helpful in deterring traffic analysis.
- *Padding Length*. This number indicates how many bytes long the padding is. The Padding Length field is mandatory.
- *Next Header*. In tunnel mode, the payload is an IP packet, so the Next Header value is set to 4 for IP-in-IP. In transport mode, the payload is usually a transport-layer protocol, often TCP (protocol number 6) or UDP (protocol number 17). Every ESP trailer contains a Next Header value.

If ESP integrity protection is enabled, the ESP trailer is followed by an Authentication Information field.

[38] SPI stands for *Security Parameter Index.* This is a 32 bit number which identifies the *Security Association*.

II.2.3. Tunnel mode vs. Transport mode

Both AH and ESP protocols operate in two modes: transport mode and tunnel mode. Each of these modes has its own applications:

- *Tunnel mode* is commonly used to encrypt traffic between secure IPsec gateways.
- *Transport mode* is used between end stations supporting IPsec or between an end station and a gateway, if the gateway is regarded as a host.

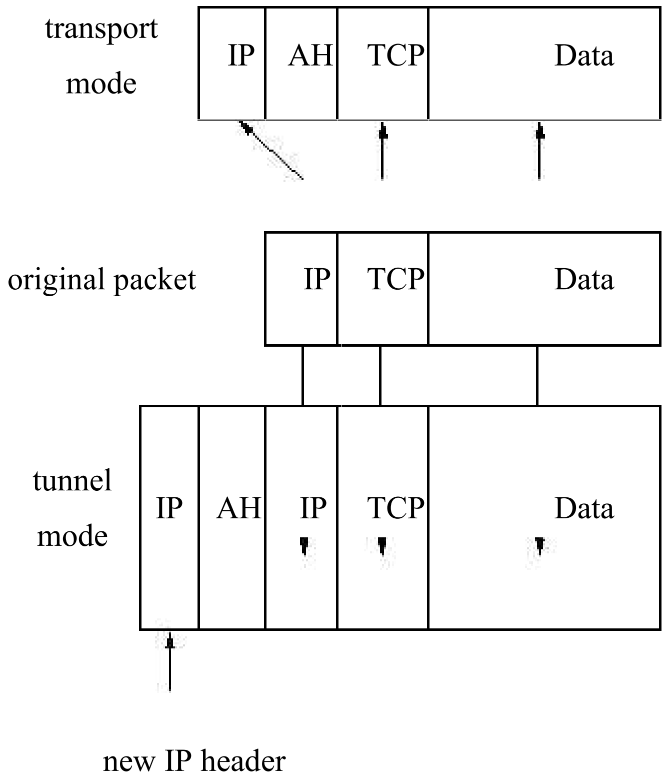


Figure 4. IPsec tunnel and transport mode

II.2.4. The Internet Key Exchange (IKE) protocol

The purpose of the Internet Key Exchange (IKE) protocol is to negotiate, create, and manage security associations[39]. The IKE protocol is very flexible and supports multiple authentication methods.

The IKE protocol functions in two phases:

- The first phase establishes an *Internet Security Association Key Management Security Association* (ISAKMP SA).
- In the second phase the ISAKMP SA is used to negotiate and setup the IPsec SAs.

The two peers must agree on a common authentication method through a negotiation process. The two main authentication protocols are:

[39] The IKE standard is defined in RFC 2409, *The Internet Key Exchange (IKE)*, available at http://www.ietf.org/rfc/rfc2409.txt. By default, IKE uses UDP port 500 for its communications.

- ***PreShared key*:**

  The same key is configured on each IPsec peer. IKE peers authenticate each other by computing and sending a keyed hash of data using the configured PreShared key. If the receiving peer is able to create the same hash independently using its own PreShared key, it knows that both peers must share the same secret, thus authenticating the other peer.

- ***RSA* (*Rivest, Shamir, Adleman*) *Signature*:**

  This uses a digital signature, where each device digitally signs a set of data and sends it to the other party. RSA signatures use a *CA* (*Certificate Authority*) to generate a unique digital certificate that is assigned to each peer for authentication. The digital certificate is similar in function to the PreShared key, but provides much stronger security.

PreShared keys are easy to implement but do not scale well, as each IPsec peer must be configured with the PreShared key of every other peer with which it will establish a session. In addition, PreShared keys are less secures and are configured in clear text format in some equipment, for example in a Cisco router.

II.2.5. Security Associations

A Security Association (SA) is an agreement between two peers engaging in a crypto exchange. This agreement includes the type and strength of the encryption algorithm used to protect the data. The SA includes the method and strength of the data authentication and the method of creating new keys for that data protection.

Each SA possesses a lifetime value for which an SA is considered valid. The lifetime value is measured in the both time (seconds) and volume (byte count) and is negotiated at SA creation. These two lifetime values are compared and agreement is reached on the lower of the two. Under normal circumstances, the lifetime value expires via time before the volume limit. Thus, if an interesting packet matches the SA within the final 120 seconds of the lifetime value of an active SA, the crypto re-key process is typically invoked. The crypto re-key process establishes another active SA before the existing SA is deleted. The result is a smooth transition with minimum packet loss to the new SA.

II.2.6. Data integrity and authenticity

Providing a way to check the integrity of information transmitted over or stored in an unreliable medium is a prime necessity in the world of open computing and communications. Mechanisms that provide such integrity checks based on a secret key are usually called message authentication codes (MACs). Typically, message authentication codes are used between two parties that share a secret key in order to authenticate information transmitted between these parties.

FIPS 198-1 defines the HMAC as a MAC that uses a cryptographic hash function in conjunction with a secret key. HMAC shall use an Approved cryptographic hash function[40]. HMAC uses the secret key for the calculation and verification of the MACs.

Data integrity is implemented by including a message *digest* (or *fingerprint*) of the data within the IPsec packets. Message digests are calculated using *hash* functions[41]. All IPsec capable devices should support hash functions *HMAC-MD5* and *HMAC-SHA*[42].

HMAC-MD5 and HMAC-SHA are based on MD5 and SHA combined with the additional crypto features of the HMAC algorithm. This is done to avoid tampering with the message digest itself. MD5 produces a 128-bit message digest and SHA produces a 160-bit message digest, therefore SHA is a more secure hash function than MD5. However, the HMAC-SHA and HMAC-MD5 variants used are truncated to the most significant 96 bits. *Truncation* has security advantages (less information on the hash available to the attacker) and disadvantages (less bits to predict for the attacker).

II.2.7. Data encryption

Data confidentiality is achieved in IPsec by the use of symmetric encryption algorithms and session keys[43]. The most commonly used algorithms are:

- ***ESP-NULL*** - No encryption applied.
- ***DES*** **(*Data Encryption Standard*)** - Provides encryption using a 56 bit key.
- ***3DES*** **(*Triple Data Encryption Standard*)** - Provides encryption using a 168 bit key.
- ***AES*** **(*Advanced Encryption Standard*)** - Provides encryption using 128, 192, and 256 key lengths.

An international VPN/IPsec via Internet must comply with the legislation of each country (encryption, size of the key, etc). Therefore, each site should be aware of national policy before using encryption.

[40] According to NIST – "*FIPS 180-3 – Secure Hash Standard (SHS)*", available at http://csrc.nist.gov/publications/fips/fips180-3/fips180-3_final.pdf;

[41] HMAC stands for *Hash Message Authentication Codes*. To derive this HMAC the IPsec protocols use hash algorithms like MD5 and SHA to calculate a hash based on a secret key and the contents of the IP datagram. This HMAC is then included in the IPsec protocol header and the receiver of the packet can check the HMAC if it has access to the secret key.

[42] As stated in the RFC (*Request For Comments*) 2401.

[43] According to RFC 2401, all IPsec devices should support at least ESP-NULL and DES. However, DES is considered a weak encryption algorithm due to its short key length, and as such, some vendors discourage its use and some others refuse to support it (i.e. FreeS/Wan).

II.2.8. Session key exchange

Diffie-Hellman (*DH*) is a public-key cryptography protocol. It allows two parties to establish a shared secret between them. DH is used within IKE to establish a shared secret that is used as a session key. The most common DH groups are:

- Group 1: Uses a 768 bit public key to establish a shared secret.
- Group 2: Uses a 1024 bit public key to establish a shared secret.

II.2.9. Tunnelling Protocols

Tunnelling protocols vary in the features they support, the problems they are designed to solve, and the amount of security they provide to the data being transported. The designs focus on the use of IPsec as a tunnelling protocol alone, and IPsec used in conjunction with *Generic Route Encapsulation* (GRE) and *Virtual Tunnel Interfaces* (VTI).

When used alone, IPsec provides a private, resilient network for IP unicast only, where support is not required for IP multicast, dynamic IGP routing protocols, or non IP protocols. When support for one or more of these features is required, IPsec should be used in conjunction with either GRE or VTI.

The p2p GRE over IPsec design allows for all three features described in the preceding paragraph, while a DMVPN design or a VTI design fulfils only the IP multicast and dynamic IGP routing protocol requirements.

Other possible tunnelling protocols include the following:

- *Secure Sockets Layer/Transport Layer Security* (SSL/TLS);
- *VPN* (WebVPN);
- *Point-to-Point Tunnelling Protocol* (PPTP);
- *Layer Two Tunnelling Protocol* (L2TP)[44].

III. TRUST MANAGEMENT FOR IPSEC

IPsec is the standard suite of protocols for network-layer confidentiality and authentication of Internet traffic. The IPsec protocols, however, do not address the policies for how protected traffic should be handled at security endpoints[45]. For many applications, security at the network layer has a number of advantages over security provided elsewhere in the protocol stack. The details of network semantics are usually hidden from applications, which therefore automatically and transparently take ad-vantage of whatever network-layer security services their environment

[44] These protocols are based on user- or client-to-gateway VPN connections, commonly called remote access solutions.

[45] Matt Blaze, John Ioannidis, Angelos D. Keromytis – "*Trust Management for IPsec*", 2003.

provides. More importantly, IPsec offers a remarkable flexibility not possible at higher- or lower-layer abstractions: security can be configured ***end-to-end*** (protecting traffic between two hosts), ***route-to-route*** (protecting traffic passing over a particular set of links), ***edge-to-edge*** (protecting traffic as it passes between "*trusted*" networks via an "*untrusted*" one, subsuming many of the current functions performed by network firewalls), or in any other configuration in which network nodes can be identified as appropriate security endpoints.

Despite this flexibility, IPsec does not itself address the problem of managing the policies governing the handling of traffic entering or leaving a host running the protocol. By itself, the IPsec protocol can protect packets from external tampering and eavesdropping, but does nothing to control which hosts are authorized for particular kinds of sessions or to exchange particular kinds of traffic. In many configurations, especially when network-layer security is used to build firewalls and virtual private networks, such policies may be necessarily quite complex. There is no standard interface or protocol for controlling IPsec tunnel creation, and most IPsec implementations provide only rudimentary, packet-filter-based and ACL-based policy mechanisms.

IPsec is based on the concept of ***datagram encapsulation***. Cryptographically protected network-layer packets are placed inside, as the payload of other network packets, making the encryption transparent to any intermediate nodes that must process packet headers for routing. Outgoing packets are encapsulated, encrypted, and authenticated (as appropriate) just before being sent to the network, and incoming packets are verified, decrypted, and decapsulated immediately upon receipt. Key management in such a protocol is straightforward in the simplest case. Two hosts can use any key-agreement protocol to negotiate keys with one another, and use those keys as part of the encapsulating and decapsulating packet transforms.

III.1. IPsec Packet Filters

It is of real importance to examine the security policy decisions an IPsec processor must make. The term "*policy*" refers specifically to the network-layer security policies that govern the flow of traffic among networks, hosts, and applications. It can be observed that policy must be enforced whenever packets arrive at or are about to leave a network security endpoint (which could be an end host, a gateway, a router, or a firewall).

IPsec "connections" are described in a data structure called a security association (SA). Encryption and authentication keys are contained in the SA at each end-point, and each IPsec-protected packet has an SA identifier that indexes the SA database of its destination host (not all SAs specify both encryption and authentication; authentication-only SAs are commonly used, and encryption-only SAs are possible but considered insecure).

When an incoming packet arrives from the network, the host first determines the processing it requires:

- If the packet is not protected, should it be accepted? This is essentially the “traditional” packet filtering problem, as performed, e.g. by network firewalls;
- If the packet is encapsulated under the security protocol:
  - Is there correct key material (contained in the specified SA) required to decapsulate it?
  - Should the resulting packet (after decapsulation) be accepted? A second stage of packet filtering occurs at this point. A packet may be successfully decapsulated and still not be acceptable (e.g., a decapsulated packet with an invalid source address, or a packet attempting delivery to some port not permitted by the receiver's policy).

A security endpoint makes similar decisions when an outgoing packet is ready to be sent:

- Is there a security association (SA) that should be applied to this packet? If there are several applicable SAs, which one should be selected?
- If there is no SA available, how should the packet be handled? It may be forwarded to some network interface, dropped, or queued until an SA is made available, possibly after triggering some automated key management mechanism such as IKE, the Internet Key Exchange protocol.

Observe that because these questions are asked on packet-by-packet basis, packet-based policy filtering must be performed, and any related security transforms applied, quickly enough to keep up with network data rates. This implies that in all but the slowest network environments there is insufficient time to process elaborate security languages, perform public key operations, traverse large tables, or resolve rule conflicts in any sophisticated manner.

IPsec implementations (and most other network-layer entities that enforce security policy, such as firewalls), therefore, employ simple, filter-based languages for configuring their packet-handling policies. In general, these languages specify routing rules for handling packets that match bit patterns in packet headers, based on such parameters as incoming and outgoing addresses and ports, services, packet options, etc.

IPsec policy control need not be limited to packet filtering, however. A great deal of flexibility is available in the control of when security associations are created and what packet filters are associated with them.

Most commonly however, in current implementations, the IPsec user or administrator is forced to provide “*all or nothing*” access, in which holders of a set of keys (or those certified by a particular authority) are allowed to create any kind of security association they wish, and others can do nothing at all.

A further issue with IPsec policy control is the need for two hosts to discover and negotiate the kind of traffic they are willing to exchange. When two hosts governed by their own policies want to communicate, they need some mechanism for determining what, if any, kinds of traffic the combined effects of one another's policies are permitted. Again, IPsec itself does not provide such a mechanism; when a host attempts to create an SA, it must know in advance that the policy on the remote host will accept it. The operation then either succeeds or fails. While this may be sufficient for small VPNs and other applications where both peers are under the same administrative control, it does not scale to larger-scale applications such as public servers.

IV. MOBILE MULTI-LAYERED IPSEC

Data confidentiality and integrity are two critical issues for wireless, mobile networks. These issues are of growing importance as wireless service providers attempt to increase wireless data traffic by providing mobile VPN services. The most widely accepted method for ensuring data confidentiality and integrity is to pass encrypted data end-to-end using a mechanism such as IPsec.

To achieve high throughput in wireless networks, smart forwarding and processing of packets in access routers are critical for overcoming the effects of the wireless links, especially highly variable delay and error rates. Several studies have shown that techniques such as smart scheduling with respect to the type of data being sent and regulation of TCP acknowledgment information, can greatly improve end-to-end performance in a wireless network. However, these services cannot be provided if data sessions are protected using end-to-end encryption as with IPsec, because the information needed by these algorithms resides inside the portion of the packet that is encrypted, and can therefore not be used by the access routers.

A previously proposed protocol, called *Multi-layered IPsec* (*ML-IPsec*) modifies IPsec in a way so that certain portions of the datagram may be exposed to intermediate network elements, enabling these elements to provide performance enhancements. However, the ML-IPsec is designed for static environments and does not examine mobility. A modified version of ML-IPsec to deal with mobility and suitable for wireless networks is the *Mobile ML-IPsec* (*MML-IPsec*).

Measurements in a wireless environment show that[46], depending on the mobility protocol chosen, integrated Mobile IP/ML-IPsec handoffs result in a pause of 56-105 milliseconds, of which only 31-85 milliseconds may be attributed to MML-IPsec. MML-IPsec only marginally

[46] Heesook Choi, Hui Song, Guohong Cao, Tom La Porta – "*Mobile Multi-Layered IPsec*", 2004;

reduced throughput compared to scenarios in which no encryption is used (9%), or those in which IPsec is used (4%), and when coupled with SNOOP[47], greatly increased throughput over scenarios using standard TCP end-to-end (50% on average), or using TCP over IPsec (165% on average). In conclusion, based on results, the MML-IPsec is a worthwhile protocol to pursue because it enables large performance improvements while providing end-to-end secure transfer of user data.

CONCLUSION

The IPsec standard provides a method to manage authentication and data protection between multiple crypto peers engaging in secure data transfer. IPsec includes the Internet Security Association and Key Management Protocol (ISAKMP) and two IPsec IP protocols: Encapsulating Security Protocol (ESP) and Authentication Header (AH).

IPsec uses symmetrical encryption algorithms for data protection. Symmetrical encryption algorithms are more efficient and easier to implement in hardware. These algorithms need a secure method of key exchange to ensure data protection. Internet Key Exchange (IKE) protocol provide this capability.

This solution requires a standards-based way to secure data from eavesdropping and modification. IPsec provides such a method. IPsec provides a choice of transform sets so that a user can choose the strength of their data protection. IPsec also has several Hashed Message Authentication Codes (HMAC) from which to choose, each giving different levels of protection for attacks such as man-in-the-middle, packet replay (anti-replay), and data integrity attacks.

REFERENCES

- CISCO Systems Inc. – "*IPsec VPN WAN Design Overview*", 2006;
- U.S. Department of Commerce, Technology Administration, National Institute of Standards and Technology – "*Special Publication 800-77*, *Guide to IPsec VPNs – Recommendations of National Institute of Standards and Technology*", 2005. Available from http://csrc.nist.gov/publications/nistpubs/800-77/sp800-77.pdf;
- http://en.wikipedia.org/wiki/IPsec, 03.06.2010;
- Niels Ferguson and Bruce Schneider – "*A Cryptographic Evaluation of IPsec*", 2000. Available from *www.schneier.com/paper-ipsec.pdf;*
- Ralf Spenneberg – "*IPsec HOWTO*", 2003. Available from http://www.ipsec-howto.org/ipsec-howto.pdf;

[47] A utility to capture and inspect network packets, included with the Solaris operating system.

- European Centre for Medium-Range Weather Forecasts – "*IPsec Feasibility Study*", 2003. Available from http://www.wmo.int/pages/prog/www/TEM/Guidance-doc/ IPSec-technote-EN.pdf;
- Mark C. Benvenuto, Angelos D. Keromytis – "*Easy VPN: IPsec Remote Access Made Easy*", 2003. Available from http://www1.cs.columbia.edu/~angelos/Papers/ easyvpn.pdf;
- Matt Blaze, John Ioannidis, Angelos D. Keromytis – "*Trust Management for IPsec*", 2003. Available from http://www.crypto.com/papers/knipsec.pdf;
- Joshua D. Guttman, Amy L. Herzog, F. Javier Thayer – "*Authentication and Confidentiality via IPsec*", 2000. Available from http://web.cs.wpi.edu/~guttman/ pubs/esorics-ipsec.pdf;
- Heesook Choi, Hui Song, Guohong Cao, Tom la Porta – "*Mobile Multi-Layered IPsec*", 2004. Available from http://www.cse.psu.edu/~tlp/paper/mml.pdf;
- Heng Yin, Haining Wang – "*Building an Application-aware IPsec Policy System*", 2004. Available from http://www.cs.wm.edu/~hnw/paper/usenix05.pdf;
- The Internet Society, Network Working Group – "*RFC 2401 – Security Architecture for the Internet Protocol*", 1998, available at http://www.ietf.org/ rfc/rfc2401.txt;
- The Internet Society, Network Working Group – "RFC 2402 – *IP Authentication Header*", 1998, available at http://www.ietf.org/rfc/rfc2402.txt;
- The Internet Society, Network Working Group – "RFC 2406 – *IP Encapsulating Security Payload*", 1998, available at http://www.ietf.org/rfc/rfc2406.txt;
- National Institute of Standards and Technology – "*FIPS 197 – Advanced Encryption Standard*", 2001, available at http://csrc.nist.gov/publications/fips/fips197/fips-197.pdf;
- The Internet Society, Network Working Group – "RFC 2409 – *The Internet Key Exchange (IKE)*", 1998, available at http://www.ietf.org/ rfc/rfc2409.txt;
- National Institute of Standards and Technology – "*FIPS 198-1 – The Keyed-Hash Message Authentication Code (HMAC)*", 2008, available at http://csrc.nist.gov/ publications/fips/fips198-1/ FIPS-198-1_final.pdf.

WIMAX SECURITY

CPT Sorin STOICA

Introduction

Security is an important topic in telecommunications. It is even more important when wireless systems are used because it is generally perceived that wireless system is easier to attack than wireline systems.

A lot has been written on the topic of WiMAX radio technology, but what about WiMAX security? Should users feel safe that their transmitted data is free from eavesdropping and manipulation? How does a WiMAX operator ensure that only authorized users access the network and that they use only the appropriate services?

This paper focuses on WiMAX and WiMAX security. WiMAX technology provides fixed, portable, or mobile non-line-of-sight (NLOS) wireless broadband access from a base station to a subscriber station. This capability may provide wide-area network coverage in the event of a wireline network failure.

WiMAX technology presents wireless broadband solutions to quickly establish network connections over a metropolitan area network (MAN). Primarily used in point-to-multipoint (PMP) mode, WiMAX allows non-line-of-sight (NLOS) wireless connectivity among various WiMAX-enabled customer premise equipment (CPE) (e.g., laptop computers, personal digital assistants [PDA]). As a result, WiMAX can support numerous communications applications (e.g., voice, video, text, and data services), thereby offering mobility and greater flexibility in communications capabilities. WiMAX is also capable of supporting backhaul technology, which would link multiple wireless local area networks (WLAN) to significantly extend a wireless footprint.

Government departments and agencies responding to a disaster can use WiMAX technology whenever a rapidly deployable network is necessary to provide temporary backup of communications. In addition, the long-range and high-speed attributes of WiMAX provide users the flexibility to perform a wide range of low latency applications to ensure higher quality of service, compared with other wireless transport technologies (e.g., Wireless Fidelity [Wi-Fi], cellular) that support lower data rates. This feature can be an advantage when coordinating emergency operations and route diversity missions.

As WiMAX becomes more popular and the standard evolves, WiMAX will be able to support new higher bandwidth applications. WiMAX was initially designed to operate over licensed spectrum; however, vendors are working toward attaining additional unlicensed spectrum to increase interoperability and WiMAX acceptance.

Defining and describing WIMAX

WiMAX (Worldwide Interoperability for Microwave Access) is a telecommunications protocol that provides fixed and fully mobile internet access. The current WiMAX revision provides up to 40 Mbit/s with the IEEE 802.16m update expected offer up to 1 Gbit/s fixed speeds. (WiMAX is based on the IEEE 802.16 standard, also called Broadband Wireless Access). The name "WiMAX" was created by the WiMAX Forum, which was formed in June 2001 to promote conformity and interoperability of the standard.[48]

WiMAX describes the technology based on the Institute of Electrical and Electronic Engineers (IEEE) 802.16 family of standards for metropolitan area networks (MAN). Originally known as WirelessMAN, the 802.16 standards were labeled WiMAX by the WiMAX Forum, an industry group composed of leading service providers and communications component and equipment companies. The WiMAX Forum works to promote and certify the compatibility and interoperability of equipment that conforms to the IEEE 802.16 standards, and defines WiMAX as “a standards-based technology enabling the delivery of last mile broadband wireless access (BWA) as an alternative to cable, digital subscriber line (DSL), or T1/E1 service”.

The original 802.16 standard, published in April 2002, specified fixed point-to-multipoint (PMP) broadband wireless systems operating in the 10–66 gigahertz (GHz) licensed spectrum. The 802.16a amendment added NLOS support for the 2–11 GHz frequency range by using multipath techniques to send and receive a signal. This amendment was updated in July 2004, with the release of 802.16-2004 (also known as 802.16d), to align the standard with aspects of the European Telecommunications Standards Institute (ETSI) HIPERMAN standard as well as lay down conformance and test specifications. Finally, WiMAX addressed mobility with the 802.16e extension published in December 2005.

WiMAX is a rather new technology and is intended to provide line-of-sight (LOS) service coverage of 30 miles from a base station to a subscriber station, or customer premise equipment (CPE), and advertises data rates up to 70 megabits per second (Mbps), with actual rates expected to be about 40 Mbps. In NLOS environments, the service range will reduce to about 6 miles for fixed and portable access applications. There is enough bandwidth in WiMAX networks to simultaneously support hundreds of businesses with T1 speed connectivity and thousands of residences with DSL speed connectivity. Furthermore, mobile WiMAX offers better building

[48] http://en.wikipedia.org/wiki/Wimax

penetration and improved security measures than fixed WiMAX and is expected to provide 15 Mbps of capacity.

WiMAX is gaining recognition because of its ability to support a wide range of applications, including broadband Internet access, T1/E1 connectivity, Voice over Internet Protocol (VoIP), Internet Protocol Television (IPTV), backhaul for Wi-Fi hotspots and cellular towers, mobile emergency response services, and wireless backhaul as a replacement for fiber optic cable. These capabilities give WiMAX the potential to serve as a reliable backup communications system or to replace a number of existing telecommunications infrastructures.

In addition, WiMAX can be used in conjunction with other emerging wireless technologies, such as Wi-Fi and Mesh Networking (type of networking wherein each node in the network may act as an independent router, regardless of whether it is connected to another network or not)[49], to extend the coverage area of the network and to provide high-speed mobile data and telecommunications services. For example, WiMAX can be used as a backhaul technology to connect multiple Wi-Fi hotspots with each other and to other parts of the Internet. It is likely that WiMAX and Wi-Fi will become complementary rather than competing technologies because Wi-Fi is designed for local area networks (LAN), whereas WiMAX is designed for broader MANs. In the very near future, it is expected that end user devices (e.g., laptops and personal digital assistants [PDA]) will be available that adhere to both the WiMAX and Wi-Fi standards to deliver wireless connectivity capabilities.

WiMAX can also be implemented in a wireless mesh network, which is a decentralized, reliable, resilient, and relatively inexpensive type of Internet infrastructure, to provide connectivity to external servers. Mesh networks consist of several nodes that act as repeaters to transmit data from nearby nodes to users located far away, resulting in networks that span large distances. Next table provides a high-level comparison of Wi-Fi, WiMAX, and Wireless Mesh technologies.

High-level Comparison of Emerging Wireless Broadband Technologies

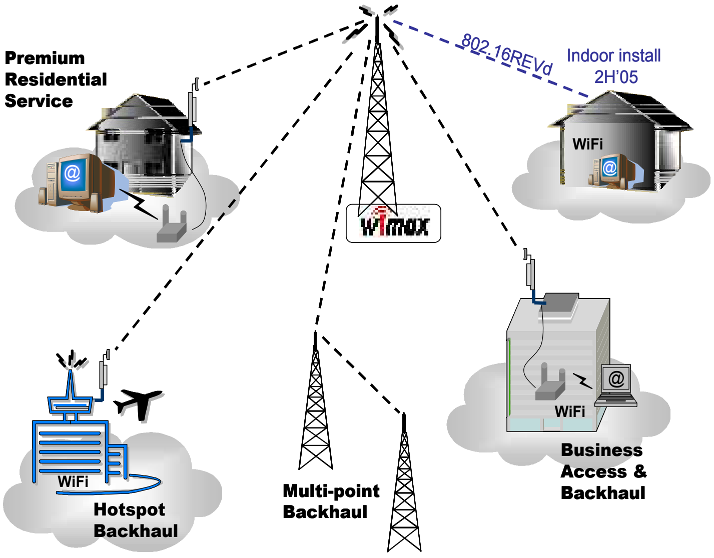
Technolog y	**Wi-Fi**	**WiMAX**	**Wireless Mesh**
Features	• Supports WLANs (e.g., indoor, office, campus environment) • Uses PMP mode, with	• Supports MANs • Has P2Pand PMP capabilities • Can operate in LOS	• Supports peer-to-peer communications, with each mobile

[49] http://en.wikipedia.org/wiki/Mesh_networking

	each client connected to an access point (AP); point-to-point (P2P) mode, with mobile users connected directly to each other • Can operate in LOS and NLOS situations • Supports fixed, portable, and mobile communications	and NLOS situations • Supports fixed, portable, and mobile communications • Is typically used as a backhaul to connect multiple Wi-Fi hotspots to external networks	user acting as a client and AP • Is self-organizing, self-healing, and auto-configuring • Typically uses wireless technologies in the unlicensed band, including Wi-Fi

Functionality

WiMAX operates similarly to Wi-Fi technology but at higher speeds, over larger distances, and accommodates more wireless users. As illustrated in the figure below, a typical WiMAX system consists of a WiMAX-enabled base station or tower, and a subscriber station or receiver. WiMAX towers are implemented by service providers and can provide a wireless service footprint as large as 2,500 square miles, similar in concept to cellular communications towers.



This capability provides broadband wireless access for users in remote rural areas, which can be difficult to reach with wires used by traditional telephone and cable companies. Initially, WiMAX receivers and antennas will consist of a small box or Personal Computer Memory Card International Association (PCMCIA) card, and eventually will be developed into portable devices that will be comparable to Wi-Fi-enabled products (e.g., laptops, telephones, PDA) on the market today.

Typical WiMAX System

Depending on the scenario, WiMAX towers can be deployed in either P2P or PMP architectures, resulting in variations in throughput based on the number of subscribers. The WiMAX base station can connect to the network backbone using a high-bandwidth, wired line (e.g., T3 line), or it can connect to another tower using a LOS microwave link, often referred to as backhaul. Using an external antenna, base stations can send and receive high-speed data and voice to/from subscriber equipment, thereby eliminating extensive and expensive wireline infrastructures and providing flexible communications solutions.

By using different frequencies, WiMAX can offer two primary forms of wireless services, LOS and NLOS. In LOS mode, a fixed dish antenna is pointed directly at the WiMAX base station from a rooftop or window. The transmissions are stronger and more stable because higher frequencies in the 10–66 GHz range can used, in which case, there is less interference and more bandwidth. On the other hand, NLOS service uses the 2–11 GHz range (similar to Wi-Fi) to transmit data because lower-wavelength transmissions are subject to fewer disruptions from physical obstructions. This is an improvement from earlier wireless technologies (e.g., local multipoint distribution system [LMDS] and multichannel multipoint distribution system [MMDS]), which were unable to provide NLOS service.

Technical Features

The 802.16-2004 and 802.16e WiMAX specifications have been developed from the IEEE standards committees to provide wireless capabilities that include various technical features, which are common across all WiMAX technologies.

For 802.16-2004 WiMAX, the IEEE selected the OFDM signaling format because of its NLOS performance, which permits significant equalizer design simplification to support operation in multipath propagation environments. WiMAX is able to support individual channel rates from 2 Mbps to 30 Mbps, and even up to 70 Mbps. WiMAX defines interoperable system profiles targeted for common licensed and unlicensed bands used around the world. This enables 802.16-based equipment to be used in diverse spectrum allocations around the world and provides broadband two-way terrestrial wireless services in both licensed and unlicensed bands from 2–11 GHz and 10–66 GHz [2].

Quality of Service (QoS) in the 802.16 MAC layer takes one of five forms: unsolicited grant service (UGS), real-time polling service (rtPS), extended real-time polling service (ErtPS), non-real-time polling service (nrtPS), and best effort (BE). The standard defines service flows for individual connections and service flow classes for groups of connections in order to define QoS levels. QoS architectures are designed to support the mapping of QoS markings (e.g., Differentiated Services Code Point [DSCP], IPv6 flow labels) to the bandwidth grant-request mechanisms.

802.16-2004 also considers optional sub-channelization in the uplink. This feature is useful when a power-limited device (e.g., laptop) is considered as the subscriber station in an indoor portable or mobile environment. The security sub-layer is part of the MAC layer and addresses authentication, establishment of keys, and encryption. A security association (SA) defines keys and encryption algorithms for each connection. The standard defines use of the Data Encryption Standard (DES) and Advanced Encryption Standard (AES). Depending on link conditions, the radio automatically adjusts the modulation level (i.e., combination of modulation scheme, code rate, and guard interval) in order to optimize performance.

802.16e incorporates SOFDMA technology, which uses finer granularity of sub-channelization and offers improved NLOS coverage and mobile performance. 802.16e QoS incorporates improved mobility to support the mapping of traditional QoS markings (e.g., DSCP, Multi-Protocol Label Switching [MPLS] flow labels, etc.) to 802.16 scheduling services. Other QoS features include Energy Savings Mechanisms for Handheld Support. This feature includes several power savings modes, including sleep mode and listening mode, to save power on handheld, battery-powered devices. Improved security in 802.16e requires mutual authentication between base station and subscriber station as opposed to the unilateral authentication supported in fixed WiMAX. 802.16e also supports device and user level authentication. Next table summarizes WiMAX's technical features as described above.

Summary of WiMAX's Technical Features

Technical Feature	Primary WiMAX Technology	
	802.16-2004	802.16e
Security	DES, AES	DES, AES
Frequency Band/ Channel Modulation	< 11 GHz OFDM (256 subcarriers), 64 QAM, 16 QAM, QPSK	< 6 GHz SOFDMA/OFDMA (128–2048), QPSK, 16QAM, 64QAM

Technical Feature	Primary WiMAX Technology	
	802.16-2004	802.16e
Range and Coverage *(commonly advertised)*	3–5 miles; Maximum range 30 miles based on tower height, antenna gain, transmit power, LOS versus NLOS, etc.	1–3 miles
Data Rate *(commonly advertised)*	Up to ~30 Mbps at 10 MHz time division duplexing (TDD)	Up to ~30 Mbps at 10 MHz TDD
Quality of Service (QoS)	Defined service flows and service classes with bandwidth request and grant mechanisms	Defined service flows and service classes with bandwidth request and grant mechanisms
Mobility	Fixed	Nomadic, regional roaming, mobile
Channel Bandwidths	Selectable channel bandwidths between 1.25 and 20 MHz with up to 16 logical sub-channels	Selectable channel bandwidths between 1.25 and 20 MHz with up to 16 logical sub-channels

WIMAX Security

Security is an important concern for the network operator and the network user. The network operator wants to know that the users and the devices connected to their network are who they say they are (to prevent malicious attacks, user spoofing), that they are accessing services that they are authorized to access and that the network users pay for the services they have used. The network users want to ensure that their privacy is protected, that the integrity of the data they send and receive is not compromised, that they can access the services they have subscribed to and that they are not over charged for those services. In fact, the expectations of the network operator and the network user are not contradictory but complimentary. Any well designed network needs to deliver these perfectly reasonable expectations which can only be achieved by the equipment vendors, system integrators and network operators working together and making the right design choices. I have summarized these security expectations by dividing the needs of network users and network operators.

Network users want: *privacy (Protect from eavesdropping),

*data integrity (Protect user data from being tampered in transit), *access to services (User has the correct credentials),

*correct accounting (Accuracy and efficiency of accounting), and

Network operators want:

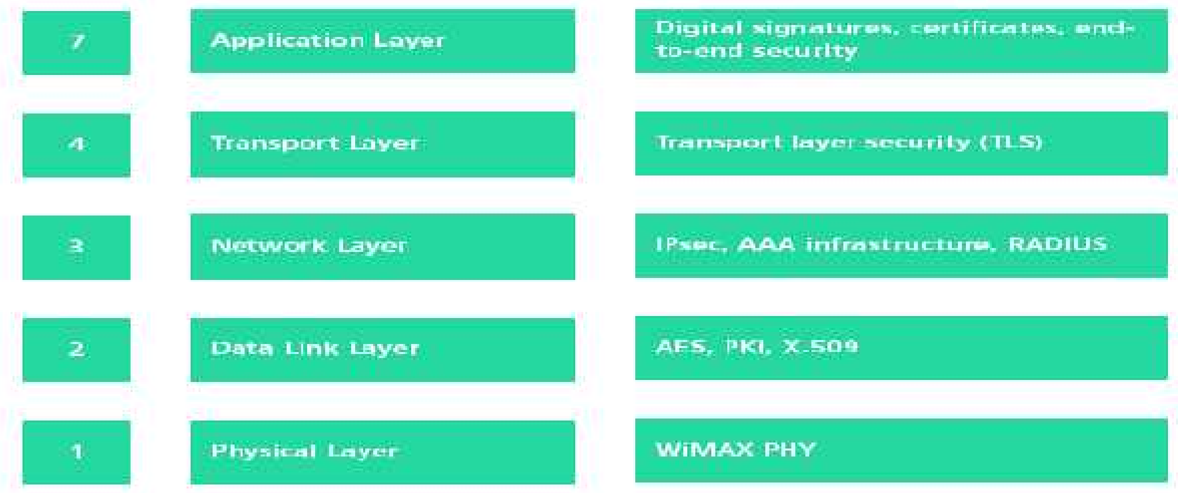
*user authentication (Is the user who he says he is?),

*device authentication (Is the device the correct device?),

*authorization (Is the user authorized to receive a particular service?),

*access control (Only authorized users have access to services).

Security is handled at multiple layers of the network, each layer handling a complimentary aspect of security. Security functions can be mapped to different layers of the OSI 7-layer model as seen in the next figure.



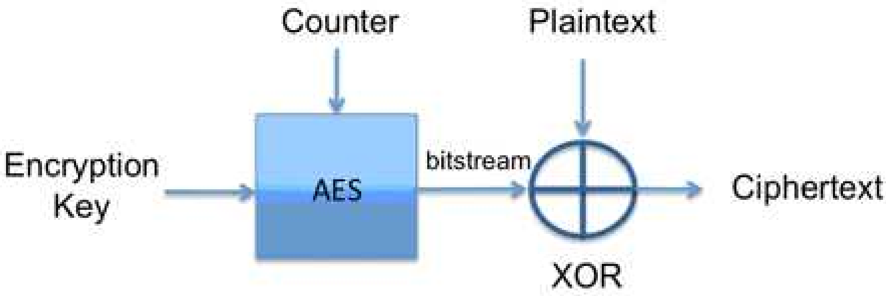
7	Application Layer	Digital signatures, certificates, end-to-end security
4	Transport Layer	Transport layer security (TLS)
3	Network Layer	IPsec, AAA infrastructure, RADIUS
2	Data Link Layer	AES, PKI, X.509
1	Physical Layer	WiMAX PHY

Security functions at various network layers

Data privacy and integrity

Encryption is a mechanism that protects data confidentiality and integrity. Encryption takes plain text (i.e., your data) and mixes that information using a complex mathematical algorithm to produce ciphertext. The ciphertext is then transmitted over the wireless network and cannot be understood by an eavesdropper.

WiMAX uses the Advanced Encryption Standard (AES) to produce ciphertext. AES takes an encryption key and a counter as input to produce a bitstream. The bitstream is then exclusive OR'd with the plaintext to produce the ciphertext (see next Figure).



AES Encryption

The receiver of the ciphertext simply reverses the process to recover the plaintext. In order for this process to work, the transmitter and the receiver must share the same encryption key.[50]

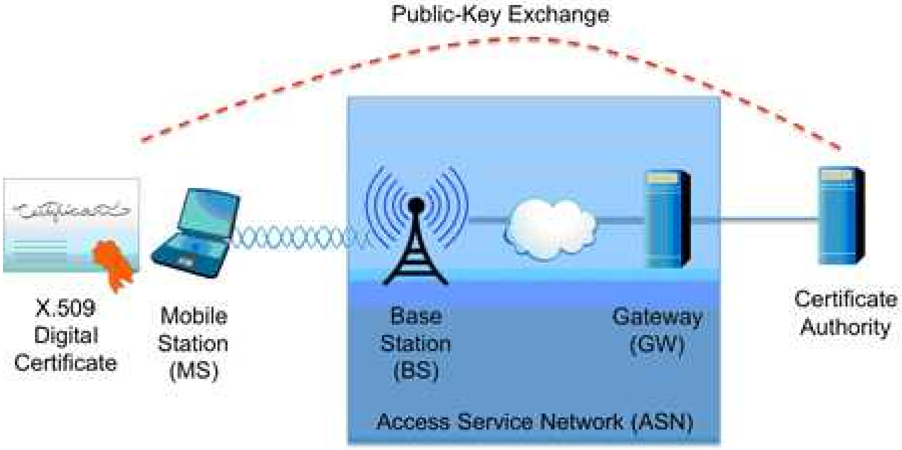
The WiMAX 802.16e-2005 standard uses the Privacy and Key Management Protocol version 2 (PKMv2) for securely transferring keying material between the base station and the mobile station. The PKMv2 mechanism validates user identity and establishes an authorization key (AK). The AK is very important because it is used to derive the encryption key described in the previous section. The Private Key Management (PKM) protocol is inherent in the Data-Over-Cable Service Interface Specification (DOCSIS) and Baseline Privacy Interface Plus (BPI+) specification. PKM is based on a series of Security Associations (SA), which are cryptographic techniques and associated keys. During the initialization, the subscriber station uses at least one SA, and each connection (except for the basic and primary management connections) is either dynamically mapped or mapped at start-up.

PKMv2 supports the use of the Rivest-Shamir-Adlerman (RSA) public key cryptography exchange. The RSA public key exchange requires that the mobile station establish identity using either a manufacturer-issued X.509 digital certificate or an operator-issued credential such as a subscriber identity module (SIM) card.

The X.509 digital certificate contains the mobile station's Public-Key (PK) and its MAC address. Traffic is encrypted by using a 56-bit (or greater) Data Encryption Standard (DES), and keys are exchanged using Triple DES (3DES). The PKM protocol messages are authenticated with the Hashed Message Authentication Code (HMAC) protocol, using secure Hashing Algorithm (SHA-1). Message authentication, when required, relies on the PKM protocol. The mobile station transfers the X.509 digital certificate to the WiMAX network, which then

[50] WiMAX security By Paul DeBeasi,SearchMobileComputing.com,
http://searchmobilecomputing.techtarget.com/tip/WiMAX-security

forwards the certificate to a certificate authority (figure below). The certificate authority validates the certificate, thus validating the user identity.



Public Key Infrastructure

Once the user identity is validated, the WiMAX network uses the public key to create the authorization key, and sends the authorization key to the mobile station. The mobile station and the base station use the authorization key to derive an identical encryption key that is used with the AES algorithm.

The features provided for Mobile WiMAX security include Extensible Authentication Protocol (EAP) based authentication, Advanced Encryption Standard-Change Configuration Management (AES-CCM) based authenticated encryption, and Cipher-based Message Authentication Code (CMAC) and HMAC-based control message protection schemes. User credentialing mechanisms include Subscriber Identity Module/Universal Subscriber Identity Module (SIM/USIM) cards, Smart Cards, Digital Certificates, and Username/Password schemes based on the relevant EAP methods for the credential type. Support exists for mutual device/user authentication, flexible key management protocol, strong traffic encryption, control and management plane message protection, and security protocol optimizations for fast handovers.

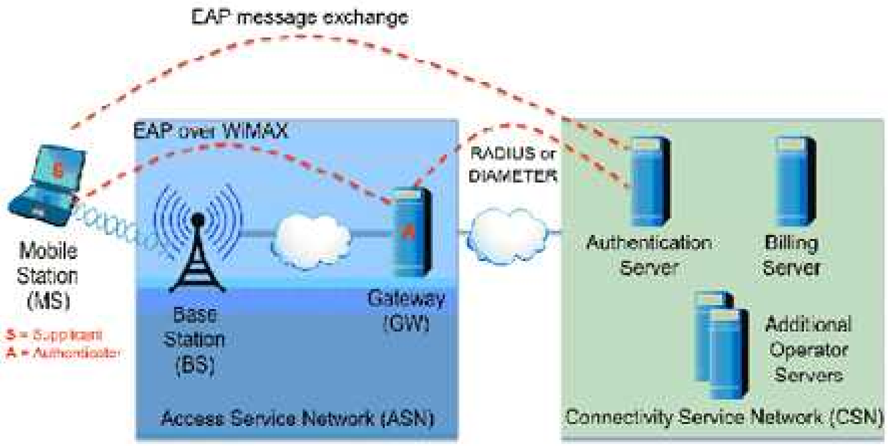
PKM Version 2 (PKMv2) is the basis of Mobile WiMAX security as defined in 802.16e. This protocol manages the Media Access Control (MAC) security using PKM-request/response (REQ/RSP) messages. PKM EAP authentication, Traffic Encryption Control, Handover Key Exchange, and Multicast/Broadcast security messages are all defined in this protocol. Mobile WiMAX supports Device and User Authentication using Internet Engineering Task Force (IETF) EAP protocol by providing support for credentials that are SIM/USIM-based, or Digital Certificate or Username/Password-based. Corresponding EAP-SIM, EAP-Authentication and Key Agreement (AKA), EAP-Transport Layer Security (TLS), or EAP-Microsoft Challenge Handshake Authentication Protocol Version 2 (MSCHAPv2) authentication methods are supported through the EAP protocol. Key deriving methods are the only EAP methods supported.

To support fast handover in WiMAX, a three-way handshake scheme is used to optimize the re-authentication mechanisms. This mechanism is also useful to prevent any man-in-the-middle attacks.

Authentication

Authentication is the process of validating a user identity and often includes validating which services a user may access. The authentication process typically involves a supplicant (that resides in the mobile station), an authenticator (that may reside in the base station or a gateway), and an authentication server (see next figure).

WiMAX uses the Extensible Authentication Protocol (EAP) to perform user authentication and access control. EAP is actually an authentication framework that requires the use of "EAP methods" to perform the actual work of authentication. The network operator may choose an EAP method such as EAP-TLS (Transport Layer Security), or EAP-TTLS MS-CHAP v2 (Tunneled TLS with Microsoft Challenge-Handshake Authentication Protocol version 2). The messages defined by the EAP method are sent from the mobile station to an authenticator. The authenticator then forwards the messages to the authentication server using either the RADIUS (Remote Authentication Dial-In User Service - a client/server protocol and software that enables remote access servers to communicate with a central server to authenticate dial-in users and authorize their access to the requested system or service. RADIUS allows a company to maintain user profiles in a central database that all remote servers can share. It provides better security, allowing a company to set up a policy that can be applied at a single administered network point) or DIAMETER (type of computer networking protocol for authentication, authorization and accounting) protocols.



EAP-based authentication

The EAP exchanges validate the user, ensure appropriate access control, and may also start the billing process. Enterprise network managers use a very similar process to authenticate users on a Wi-Fi network.

The Extensible Authentication Protocol (EAP) is a protocol for wireless networks that expands on authentication methods used by the Point-to-Point Protocol (PPP), a protocol often used when connecting a computer to the Internet. EAP can support multiple authentication mechanisms, such as token cards, smart cards, certificates, one-time passwords, and public key encryption authentication.

Here's how it works: in communications using EAP, a user requests connection to a wireless network through an access point (a station that transmits and receives data, sometimes known as a transceiver). The access point requests identification (ID) data from the user and transmits that data to an authentication server. The authentication server asks the access point for proof of the validity of the ID. After the access point obtains that verification from the user and sends it back to the authentication server, the user is connected to the network as requested.

Authentication comes in two forms:

- unilateral authentication where the BS authenticates the MS and
- mutual authentication where the BS authenticates the MS and the MS authenticates the BS.

Every WiMAX implementation must have unilateral authentication. Experience has shown that mutual authentication is also extremely useful to have. Authentication is achieved using a public key interchange protocol which ensures not only authentication but also the establishment of encryption keys. In public key interchange schemes each participant must have a private key and a public key. The Public key is known widely whereas the private key is kept secret. WiMAX 802.16e-2005 standard defines a Privacy Key Management (PKM) protocol which allows for three types of authentication:

a) RSA based authentication - X.509 digital certificates together with RSA encryption,

b) EAP based authentication (optional),

c) RSA based authentication followed by EAP authentication.

PKM authentication protocol establishes a shared secret key called Authorization Key (AK) between the MS and the BS. Once a shared AK is established between the BS and the MS, Key Encryption Key (KEK) is derived from it. KEK is then used to encrypt subsequent PKM exchanges of Traffic Encryption Key (TEK). In the RSA based authentication, a BS authenticates the MS by virtue of its unique X.509 digital certificate which has been issued by the MS manufacturer. The X.509 certificate contains the MS's Public Key (PK) and its MAC address. When requesting an AK, the MS sends its digital certificate to the BS which validates the certificate and then uses the verified PK to encrypt an AK which is then sent back to the MS. All MSs that use RSA authentication have factory installed private/public key pairs (or an algorithm to generate the keys dynamically) together with factory installed X.509 certificates. In the case of EAP based authentication the MS is authenticated either through a unique operator issued credential, such as a SIM or though an X.509 certificate as described above. The choice of authentication method depends on the operator's choice of type of EAP as follows:

- EAP-AKA (Authentication and Key Agreement) for SIM based authentication,
- EAP-TLS for X.509 based authentication,
- EAP-TTLS for MS-CHAPv2 (Microsoft-Challenge Handshake Authentication Protocol).

The BS associates the MS's authenticated identity to a paying subscriber and hence to the services the subscriber is authorized to access. Thus, through the exchange of AK, the BS determines the authenticated identity of the MS and the services it is authorized to access.

Security Association

A Security Association (SA) is defined as the set of security information shared between a BS and one or more of the MSs connected to that BS in order to support secure communications across the WiMAX access network. Three types of SA have been defined, primary, static and dynamic. Each MS establishes a primary SA during the MS initialization phase. Static SAs are provided within the BS. Dynamic SAs are created and destroyed in real time in response to the creation and termination of service flows. Each MS can have several service flows on the go and can therefore have several dynamic SAs. The BS makes sure that the assigned SAs are compatible with the service types the MS is authorised to access.

Authorization

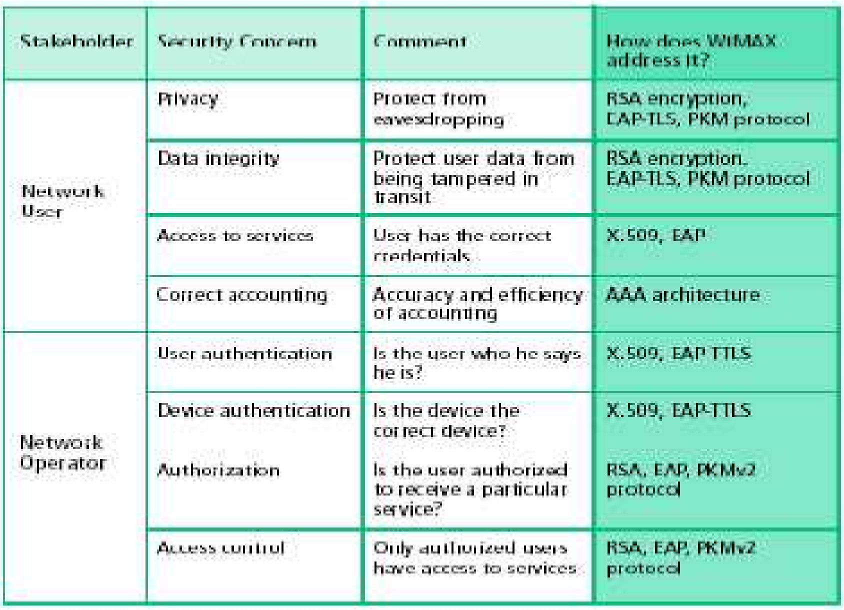
Following authentication, MS requests authorization from the BS. This is a request for an AK as well as for an SA identity (SAID). The Authorization Request includes MS's X.509 certificate, encryption algorithms and cryptographic ID. In response, the BS carries out the necessary validation (by interacting with an AAA server in the network) and sends back an Authorization reply which contains the AK encrypted with the MS's public key, a lifetime key and an SAID. After the initial authorization, the AAA via the BS periodically reauthorizes the MS.

Traffic Encryption

As we have seen above, the authentication and authorization process results in the assignment of and Authorization Key, which is 160 bits long. The Key Encryption Key is derived directly from the AK and is 128 bits long. The KEK is not used for encrypting traffic data; for this we require the Traffic Encryption Key which is generated as a random number in the BS using the TEK encryption algorithm where KEK is used as the encryption key. TEK is then used for encrypting the data traffic.

To support traffic encryption, AES-CCM is used as the cipher mechanism for protecting all the user data over the Mobile WiMAX MAC interface. The keys used for driving the cipher are generated from the EAP authentication. A Traffic Encryption State machine that has a periodic key refresh mechanism enables sustained transition of keys to further improve protection. Control data is protected using AES-based CMAC, or Message Digest 5 (MD5) based HMAC schemes.

The table below summarises how the mobile WiMAX standard addresses the security requirements mentioned in the beginning of this chapter.



Stakeholder	Security Concern	Comment	How does WiMAX address it?
Network User	Privacy	Protect from eavesdropping	RSA encryption, EAP-TLS, PKM protocol
	Data integrity	Protect user data from being tampered in transit	RSA encryption. EAP-TLS, PKM protocol
	Access to services	User has the correct credentials	X.509, EAP
	Correct accounting	Accuracy and efficiency of accounting	AAA architecture
Network Operator	User authentication	Is the user who he says he is?	X.509, EAP TTLS
	Device authentication Authorization	Is the device the correct device? Is the user authorized to receive a particular service?	X.509, EAP-TTLS RSA, EAP, PKMv2 protocol
	Access control	Only authorized users have access to services	RSA, EAP, PKMv2 protocol

How WiMAX standard addresses security expectations

WiMAX Security and Deployment

A terrorist attack could destroy wired connections to ISPs. Instead, the government or the military could construct WiMAX stations within heavily armoured bunkers. With redundant transmitters constructed, no single attack could take down the entire network. Officials using the network could continue communicating throughout an attack despite best attempts to thwart communications.

WiMAX base stations may be purchased by city governments to attract businesses to for instance financial districts. The base stations would provide wireless internet access, similar to how Wi-Fi hotspots are provided, but covering a much larger area. Businesses may also install WiMAX base stations and proceed to charge users for access. On the other hand, businesses may also provide access to their WiMAX network free of charge as an incentive to purchase their products, as is currently done for cafes and coffee shops with smaller Wi-Fi networks. The cost of purchasing access to a WiMAX network would be in theory less than currently wired ISPs, because the cost due to laying wire is absent.

VoIP (Voice over Internet Protocol) allows users to make phone calls over the internet using their computer. If WiMAX is widely deployed in a city, then users could bypass their phone providers and place VoIP phone calls without charge (except for the price of WiMAX access). This reality increases the competitiveness of VoIP and should cause telephone and cell phone providers to change their price structures in a manner that should benefit consumers.[51]

WiMAX Advantages and Disadvantages

Advantages of WiMAX

Network connections based on the WiMAX platform offer many benefits to clients, especially in the areas of ease of deployment and cost. Currently, service providers can take up to 3 months to provision a T1/E1 network for a business customer; however, service could be provided in a matter of days and at a fraction of the cost with wireless broadband technology. WiMAX technology no longer requires the use of wires to set up a LAN. Thus, WiMAX networks are easier and quicker to install compared with traditional wired LANs. This feature can be useful in emergency situations where communications must be established in a short period of time. Costs related to excess cabling and labour are also significantly reduced with the use of WiMAX.

In addition, WiMAX technology is standards based and offers wider coverage and higher capacity, thereby providing greater flexibility in available communications services. Where Wi-Fi technology typically offers service coverage within smaller footprint environments (e.g., office, conference hall, small campus), WiMAX technology extends service to broader metropolitan areas (e.g., city-wide), with each WiMAX device capable of providing greater coverage compared with Wi-Fi devices. This is made possible with advanced techniques such as beam-forming and multiple-input, multiple-output (MIMO) technology, which improve NLOS performance. This capability offers users situated in locations not accessible by wires with an

[51] 802.11 (WiFi) and 802.16 (WiMax) wireless networks, ITM 440 Introduction to Networks and the Internet, Professor Kevin Vacarro, Student Steve Talbot

alternative method for easily connecting to a nearby network. WiMAX also provides for higher capacities in the areas of data rates and actual throughputs with flexible channel bandwidths, which is important for latency sensitive services such as voice and video. Moreover, a standards-based technology means greater interoperability. Interoperable equipment lets users purchase WiMAX Certified equipment from more than one vendor without worrying about compliance issues.

Now that WiMAX is capable of supporting mobile access, users possessing portable WiMAX devices can stay attached to the wireless network while roaming. The fact that WiMAX supports mobility introduces additional voice and video applications that could not be previously supported. Applications such as VoIP and vehicular communications require low latency in order to provide high-quality service, which is now possible through WiMAX networks. WiMAX is also complementary to Wi-Fi, and service providers could use WiMAX equipment to connect Wi-Fi hotspots in order to expand a network. This significantly simplifies coordination of critical field operations necessary during an emergency crisis.

Disadvantages of WiMAX

Although WiMAX is growing in popularity, a few limitations must be taken into account before initiating deployment. Presently, the major drawback is the operating radio frequency (RF) spectrum that WiMAX has defined. Mobile WiMAX specifies additional operating frequencies in the 2.3 and 2.5 GHz bands, which increases spectrum availability, but users still face the issue of using licensed spectrum. Vendors are working toward obtaining unlicensed spectrum (e.g., 2.4 GHz) as well, but controlling service quality is still a problem because other users operating on the same band can lead to interference (e.g., Wi-Fi) and cause a reduction in data throughput. Devices such as microwaves and cordless telephones cause interference as well because they operate on the same frequency bands.

One misconception associated with WiMAX is that it is a technology that will enable wireless communications at rates of 70 Mbps over 30 miles. It is important to note that a trade off must be made between data rates and coverage range depending on the needs of a particular scenario. LOS is required for long distance connections (e.g., 30 miles), and the data rate will decrease as the distance between base station and subscriber station increases. In addition, certain conditions, such as the terrain, weather, or other obstructions, can act to reduce the maximum range of the system.

Finally, WiMAX, particularly the mobile version, may face competition from IEEE 802.20 mobile broadband technology, which targets high-speed, wireless, IP-based connectivity to devices such as cellular telephones, laptops, and PDAs. WiMAX and IEEE 802.20 are considered two different technology approaches that are targeted at distinct markets. IEEE 802.20, however, is still in the very early stages of standards development and is not considered a threat to the development of WiMAX. Additionally, because IEEE 802.20 does not currently have broad industry support like the WiMAX Forum, interoperability in 802.20 is currently questionable and therefore much further away.

CONCLUSIONS

In conclusion WiMAX provides robust user authentication, access control, data privacy and data integrity using sophisticated authentication and encryption technology. WiMAX users should feel confident that their transmitted data is free from eavesdropping or manipulation and that only authorized users can access WiMAX services.[52]

X.509 hence provides enough security against major threats to services. However, lacking in base stations and service provider authentication leaves a big loop whole in the authentication mechanism used by WiMax privacy and key management (PKM), this expose WiMax subscribers to different confidentiality and availability attacks by unauthorized users. To protect from such vulnerability, amendments where made in 802.16 standards. Extensible Authentication Protocol (EAP) was introduced in 802.16e.

In terms of encryption, the **i**ntroduction of 802.16e and support for the AES provide the strong support of confidentially of data traffic. 802.16 has same issue just like 802.11 management frames are not encrypted, that increases the chances for attackers or intruders to collect information about subscribers and network.

Regarding the availability**,** WiMAX uses licensed Radio Frequency (RF) spectrum, which provides protection from unauthorized access to some extent. There are tools available which are easy to use for jamming the spectrum from all planned WiMAX deployments. Along with physical layer denial of service attacks, intruders can use legacy management frames to disconnect the current legitimate network connections.

To conclude with WiMAX threats we can say that with all efforts to make WiMAX a very secure technology, there are still several potential attacks which are threats to WiMAX usability few of them are DOS attacks, Rogue Base Stations, Man in the middle attacks, spoofing of management frames etc. The actual testing of WiMAX security will come into play when WiMAX provider begins wide scale network deployment. This will give attackers good chance and better exposure to manipulate the network security, and access CPE equipment. Until the proper deployment of WiMAX technology, security issues are just speculations.

[52] http://searchmobilecomputing.techtarget.com/tip/WiMAX-security?mboxConv=searchCIO_RegActivate_Submit&

REFERENCES

1. ***WiMAX*** - NCS RDP Whitepaper: Wi-Fi Retrieved June 10, 2010 from *www.ncs.gov/rdp/.../RDP%20WiMAX%20White%20Paper.doc* /.
2. Ohrtman, Frank, & WMX Systems, LLC. (2006). *WIMAX: A SIMPLE EXPLANATION TO A COMPLEX SUBJECT.* Retrieved June 10, 2010, from http://www.wimax.com/education/wimax/information.
3. Fujitsu Microelectronics America Inc. (2004). *RF Spectrum Use in WiMAX*. Retrieved June 10, 2010, from http://www.analogzone.com/nett1129.pdf.
4. Eklund, C., Marks, R., Stanwood, K., & Wang, S. (2002). *IEEE Standard 802.16: A Technical Overview of the WirelessMAN Air Interface for Broadband Wireless Access*. Retrieved June 13, 2010, from http://www.ieee802.org/16/docs/02/C80216-02_05.pdf.
5. http://www.wifinotes.com/wimax/wimax-security.html, Retrieved June 13, 2010.
6. http://en.wikipedia.org/wiki/WiMAX
7. WIMAX Security, by Paul DeBeasi Retrieved June 13, 2010 from http://searchmobile computing.techtarget.com/tip/WiMAX-security?mboxConv=searchCIO_RegActivate _Submit&
8. Mobile WiMAX: The *4G* Revolution Has Begun - Mobile WiMAX Retrieved June 13, 2010 from *www.sprint.com/.../Mobile_WiMAX_The_4G_Revolution_ Has_Begun_ Jan2010.pdf*

INFORMATION SECURITY POLICY AND POLICY ON APPROPRIATE USE OF COMPUTERS AND NETWORK SYSTEMS AT THE UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

MAJ Dinu Emilian TUNDREA

INTRODUCTION

Storage of university data on computers and transfer across the network eases use and expands our functionality. Commensurate with that expansion is the need for the appropriate security measures. Security is not distinct from the functionality.

The Information Security Policy (Policy) recognizes that not all communities within the University are the same and that data are used differently by various units within the University. The principles of academic freedom and free exchange of ideas apply to this policy, and this policy is not intended to limit or restrict those principles. These policies apply to all units within the University.

Each unit within the University should apply this policy to meet their information security needs. The Policy is written to incorporate current technological advances. The technology installed at some units may limit immediate compliance with the Policy. Instances of non-compliance must be reviewed and approved by the chief information officer or the equivalent officer(s).

Throughout the document the term *must* and *should* are used carefully. "Musts" are not negotiable; "shoulds" are goals for the university. The terms *data* and *information* are used interchangeably in the document.

The terms *system* and *network* administrator are used in this document. These terms are generic and pertain to any person who performs those duties, not just those with that title or primary job duty. Many students, faculty and staff member are the system administrators for their own machines.

1. INFORMATION SECURITY POLICY

1.1. PURPOSE OF THIS POLICY

By information security we mean protection of the University's data, applications, networks, and computer systems from unauthorized access, alteration, or destruction

The purpose of the information security policy is:

- To establish a University-wide approach to information security.
- To prescribe mechanisms that help identify and prevent the compromise of information security and the misuse of University data, applications, networks and computer systems.
- To define mechanisms that protect the reputation of the University and allow the University to satisfy its legal and ethical responsibilities with regard to its networks' and computer systems' connectivity to worldwide networks.
- To prescribe an effective mechanism for responding to external complaints and queries about real or perceived non-compliance with this policy.

1.2. RESPONSIBILITY

The chair of the University Technology Management Team (UTMT) is responsible for implementing the policy. UTMT, chaired by the Vice President for Administration, is a coordinating group comprised of chief information officers from the three campuses, the university administration, and the hospital.

UTMT must see to it that:

- The information security policy is updated on a regular basis and published as appropriate.
- Appropriate training is provided to data owners, data custodians, network and system administrators, and users.
- Each unit appoints a person to be responsible for security implementation, incident response, periodic user access reviews, and education of information security policies including, for example, information about virus infection risks.

Members of UTMT are each responsible for establishing procedures to implement these policies within their areas of responsibility, and for monitoring compliance.

1.3. GENERAL POLICY

Required Policies

- The University will use a layered approach of overlapping controls, monitoring and authentication to ensure overall security of the University’s data, network and system resources.
- Security reviews of servers, firewalls, routers and monitoring platforms must be conducted on a regular basis. These reviews must include monitoring access logs and results of intrusion detection software, where it has been installed.

Recommended Practices

- Vulnerability and risk assessment tests of external network connections should be conducted on a regular basis. At a minimum, testing should be performed annually, but the sensitivity of the information secured may require that these tests be done more often.
- Education should be implemented to ensure that users understand data sensitivity issues, levels of confidentiality, and the mechanisms to protect the data. This should be tailored to the role of the individual, network administrator, system administrator, data custodian, and users.
- Violation of the Information Security Policy may result in disciplinary actions as authorized by the University in accordance with University and campus disciplinary policies, procedures, and codes of conduct.

1.4. DATA CLASSIFICATION POLICY

It is essential that all University data be protected. There are however gradations that require different levels of security. All data should be reviewed on a periodic basis and classified according to its use, sensitivity, and importance. They have specified three classes below:

- **High Risk:** Information assets for which there are legal requirements for preventing disclosure or financial penalties for disclosure. Data covered by federal and state legislation, such as FERPA, HIPAA or the Data Protection Act, are in this class. Payroll, personnel, and financial information are also in this class because of privacy requirements.

This policy recognizes that other data may need to be treated as high risk because it would cause severe damage to the University if disclosed or modified. The data owner should make this determination. It is the data owner’s responsibility to implement the necessary security requirements.

- **Confidential:** Data that would not expose the University to loss if disclosed, but that the data owner feels should be protected to prevent unauthorized disclosure. It is the data owner’s responsibility to implement the necessary security requirements.
- **Public:** Information that may be freely disseminated

All information resources should be categorized and protected according to the requirements set for each classification. The data classification and its corresponding level of protection should be consistent when the data is replicated and as it flows through the University.

- Data owners must determine the data classification and must ensure that the data custodian is protecting the data in a manner appropriate to its classification.

- No University-owned system or network subnet can have a connection to the Internet without the means to protect the information on those systems consistent with its confidentiality classification.
- Data custodians are responsible for creating data repositories and data transfer procedures which protect data in the manner appropriate to its classification.
- High risk data must be encrypted during transmission over insecure channels.
- Confidential data should be encrypted during transmission over insecure channels.
- All appropriate data should be backed up, and the backups tested periodically, as part of a documented, regular process.
- Backups of data must be handled with the same security precautions as the data itself. When systems are disposed of, or repurposed, data must be certified deleted or disks destroyed consistent with industry best practices for the security level of the data.

1.5. ACCESS CONTROL POLICY

- Data must have sufficient granularity to allow the appropriate authorized access. There is a delicate balance between protecting the data and permitting access to those who need to use the data for authorized purposes. This balance should be recognized.
- Where possible and financially feasible, more than one person must have full rights to any university owned server storing or transmitting high risk data. The campuses and University Administration (UA) must have a standard policy that applies to user access rights. This will suffice for most instances. Data owners or custodians may enact more restrictive policies for end-user access to their data.
- Access to the network and servers and systems should be achieved by individual and unique logins, and should require authentication. Authentication includes the use of passwords, smart cards, biometrics, or other recognized forms of authentication.
- As stated in the current campus policies on appropriate and acceptable use, users must not share usernames and passwords, nor should they be written down or recorded in unencrypted electronic files or documents. When limited access to university-related documents or files is required specifically and solely for the proper operation of University units and where available technical alternatives are not feasible, exceptions are allowed under an articulated unit policy that is available to all affected unit personnel. Each such policy must be reviewed by the unit executive officer and submitted to the CIO for approval. All users must secure their username or account, password, and system access from unauthorized use.

- All users of systems that contain high risk or confidential data must have a strong password- the definition of which will be established and documented by UTMT after consultation with the community. Empowered accounts, such as administrator, root or supervisor accounts, must be changed frequently, consistent with guidelines established by UTMT.
- Passwords must not be placed in emails unless they have been encrypted.
- Default passwords on all systems must be changed after installation. All administrator or root accounts must be given a password that conforms to the password selection criteria when a system is installed, rebuilt, or reconfigured.
- Logins and passwords should not be coded into programs or queries unless they are encrypted or otherwise secure.
- Users are responsible for safe handling and storage of all University authentication devices. Authentication tokens (such as a SecureID card) should not be stored with a computer that will be used to access the University's network or system resources. If an authentication device is lost or stolen, the loss must be immediately reported to the appropriate individual in the issuing unit so that the device can be disabled.
- Terminated employee access must be reviewed and adjusted as found necessary. Terminated employees should have their accounts disabled upon transfer or termination. Since there could be delays in reporting changes in user responsibilities, periodic user access reviews should be conducted by the unit security person.
- Transferred employee access must be reviewed and adjusted as found necessary.
- Monitoring must be implemented on all systems including recording logon attempts and failures, successful logons and date and time of logon and logoff.
- Activities performed as administrator or superuser must be logged where it is feasible to do so.
- Personnel who have administrative system access should use other less powerful accounts for performing non-administrative tasks. There should be a documented procedure for reviewing system logs.

1.6. VIRUS PREVENTION POLICY

- The willful introduction of computer viruses or disruptive/destructive programs into the University environment is prohibited, and violators may be subject to prosecution.
- All desktop systems that connect to the network must be protected with an approved, licensed anti-virus software product that it is kept updated according to the vendor's recommendations.

- All servers and workstations that connect to the network and that are vulnerable to virus or worm attack must be protected with an approved, licensed anti-virus software product that it is kept updated according to the vendor's recommendations.
- Headers of all incoming data including electronic mail must be scanned for viruses by the email server where such products exist and are financially feasible to implement. Outgoing electronic mail should be scanned where such capabilities exist.
- Where feasible, system or network administrators should inform users when a virus has been detected.
- Virus scanning logs must be maintained whenever email is centrally scanned for viruses.

1.7. INTRUSION DETECTION POLICY

- Intruder detection must be implemented on all servers and workstations containing data classified as high risk.
- Operating system and application software logging processes must be enabled on all host and server systems. Where possible, alarm and alert functions, as well as logging and monitoring systems must be enabled.
- Server, firewall, and critical system logs should be reviewed frequently. Where possible, automated review should be enabled and alerts should be transmitted to the administrator when a serious security intrusion is detected.
- Intrusion tools should be installed where appropriate and checked on a regular basis.

1.8. INTERNET SECURITY POLICY

- All connections to the Internet must go through a properly secured connection point to ensure the network is protected when the data is classified high risk.
- All connections to the Internet should go through a properly secured connection point to ensure the network is protected when the data is classified confidential.

1.9. SYSTEM SECURITY POLICY

- All systems connected to the Internet should have a vendor supported version of the operating system installed.
- All systems connected to the Internet must be current with security patches.
- System integrity checks of host and server systems housing high risk University data should be performed.

1.10. ACCEPTABLE USE POLICY

Each Campus and UA must have a policy on appropriate and acceptable use that includes these requirements:

- University computer resources must be used in a manner that complies with University policies and State and Federal laws and regulations. It is against University policy to install or run software requiring a license on any University computer without a valid license.
- Use of the University's computing and networking infrastructure by University employees unrelated to their University positions must be limited in both time and resources and must not interfere in any way with University functions or the employee's duties. It is the responsibility of employees to consult their supervisors, if they have any questions in this respect.
- Uses that interfere with the proper functioning or the ability of others to make use of the University's networks, computer systems, applications and data resources are not permitted.
- Use of University computer resources for personal profit is not permitted except as addressed under other University policies.
- Decryption of passwords is not permitted, except by authorized staff performing security reviews or investigations. Use of network sniffers shall be restricted to system administrators who must use such tools to solve network problems. Auditors or security officers in the performance of their duties may also use them. They must not be used to monitor or track any individual's network activity except under special authorization as defined by campus policy that protects the privacy of information in electronic form.

1.11. EXCEPTIONS

In certain cases, compliance with specific policy requirements may not be immediately possible. Reasons include, but are not limited to, the following:

- Required commercial or other software in use is not currently able to support the required features;
- Legacy systems are in use which do not comply, but near-term future systems will, and are planned for;
- Costs for reasonable compliance are disproportionate relative to the potential damage.

In such cases, units must develop a written explanation of the compliance issue and a plan for coming into compliance with the University's Information Security Policy in a reasonable amount of time. Explanations and plans must be submitted to the campus CIO or the equivalent officer(s).

2. POLICY ON APPROPRIATE USE OF COMPUTERS AND NETWORK SYSTEMS

2.1. PURPOSE OF THIS POLICY

The University of Illinois at Urbana-Champaign provides extensive computing and network communications services. These services, known collectively as UIUCnet, are part of the campus infrastructure, and their purpose is to support the University's teaching, research, and service missions. Unless explicitly noted, these policies apply to all computing and network communications equipment in all units.

This document addresses issues specific to University of Illinois computing and network usage. Sections 1 through 5 articulate policies regarding individual users of computers and networks; sections 6 and following summarize administrative protocols for computing and network administrators and should not be construed as creating additional rights for individual users. Other university and campus policies that address specific activities and behaviors, some of which are cited later in this policy, continue to apply to computing and network use. Individuals using campus computing and networking services should be particularly aware of policies that apply to discrimination, harassment, the use of copyrighted materials, and those that apply to the appropriate use of university resources. Many major policy documents can be found on the web site. Computing and network communications are changing rapidly both in terms of technology and application, and the University reserves the right to amend this policy at any time.

All members of the campus community are given notice of this policy by virtue of its publication, and are subject to it on the same basis. Ignorance of this policy does not relieve anyone of his or her responsibilities under it.

2.2. UNDERLYING PRINCIPLES

a) The principles of *academic freedom* apply in full to electronic communications.

b) The use of computing and network services provided by the campus is subject to all applicable state and federal laws, as well as general University and campus policies.

c) All standards of behavior, courtesy, and etiquette that govern vocal and written communications also extend to electronic communications.

d) When the Office of the Chief Information Officer (CIO$^)$ becomes aware of any use of UIUCnet that violates provisions of University policy, presents a security risk, or degrades services to others, it may suspend or terminate network access and use and/or notify appropriate disciplinary and/or legal authorities. Where possible, the Office of the CIO will provide prior notification of actions that affect network use and access. In all cases, the CIO or designee will provide timely notification of the reasons for said actions and will document the appeal process available to those affected. The responsibilities of the Office of the CIO include:

i) The choice of protocols supported by the network,

ii) The definition of campus standards necessary for efficient operation of the network and for the security of transmitted data and networked computers,

iii) Application of network management policies adopted by the campus to ensure inter-operability of departmental local area networks (LANs),

iv) Monitoring the overall system to ensure the reliability, robustness and security of the campus network infrastructure, and

v) Serving as the campus representative to the Internet community and ensuring that the campus is a responsible member of that community.

2.3. COMPUTERS AND NETWORK SYSTEMS: UIUCnet DEFINED

The term *UIUCnet* is used here to denote the campus computer and data communications infrastructure at the University of Illinois at Urbana-Champaign. It includes the campus backbone and local area networks, all equipment connected to those networks (independent of ownership), and all equipment registered to any domain name owned by the University.

2.4. PROPER AND AUTHORIZED USE OF UIUCnet

The Office of the CIO, through Campus Information Technologies and Educational Services (CITES), is charged with ensuring the integrity of UIUCnet computers and communications. CITES takes active steps to ensure the physical integrity of the infrastructure, including routine monitoring of performance and reliability. While CITES does not routinely monitor appropriate use of UIUCnet by individuals, the CITES Security Office will respond to complaints or other notifications of inappropriate use. Units that provide access to UIUCnet are responsible for ensuring that use is limited to legitimate users and is consistent with University policies and contractual obligations that govern the software and services offered on UIUCnet. Use of UIUCnet is a privilege, not a right, and such use may be suspended or terminated at the direction of the CITES Security Office when, in its judgment, this policy has been violated by the user. The CIO or designee will provide timely notification of the reasons for suspension or termination and will document the appropriate appeals process.

a) **Purpose of UIUCnet, the campus computing and communications infrastructure**: UIUCnet exists to support the educational, research, and public service missions of the University, and its use should be limited to those purposes.

b) **Authorized Users**: The document *Authorized Users at the University of Illinois at Urbana-Champaign*, defines who may use UIUCnet, the types of accounts available, and the duration of access.

c) **Appropriate Use of UIUCnet**: No individual may use UIUCnet resources for commercial or profit-making purposes or other purposes that interfere with the mission of the University. As with all University computing and network facilities, UIUCnet may not be used for improper or illegal purposes, such as unauthorized use of licensed software, intentional efforts to breach security, or the transmission of computer viruses.

    i) **Ownership of Network Identifiers**: University-supplied network identifiers (network IDs), University identification numbers, and computer sign-ons are the property of the University. The University may revoke these identifiers or sign-ons at any time.

    ii) **Responsibility to Maintain Privacy of Passwords**: Passwords associated with an individual's network IDs and computer sign-ons should not be shared without authorization. Compromised passwords may affect not only the individual, but also other users on campus or on the Internet.

    iii) **Proper Identity Required**: Electronic mail and other forms of electronic communication must carry the proper identity of the sender at all times. Information servers (e.g., Web servers) must display the email address and identity of the unit or person responsible for maintaining the information.

    iv) **Appropriate Use of Bandwidth**: As described in Section 10 below, bandwidth both within campus and connecting to the Internet is a shared, finite resource. Users of UIUCnet must make reasonable efforts to use this resource in ways that do not negatively affect others. Units may set guidelines on bandwidth utilization for purposes of resource allocation.

d) **Use by Faculty and Staff**: Use of UIUCnet by faculty and staff is also governed by the University's Standards and Ethics Handbook.

    i) **Passwords and University Units**: Faculty and staff, including student employees, must not under most circumstances share their passwords with others, even with supervisors. However, when limited access to university-related documents or files is required specifically and solely for the proper operation of University units and where available technical alternatives are not feasible, exceptions are allowed under an articulated unit policy that is available to all affected unit personnel. Each such policy must be reviewed by the unit executive officer and submitted to the CIO for approval.

ii) **Use Unrelated to University Positions**: Use by University employees unrelated to their University positions must be limited in both time and resources and must not interfere in any way with University functions or the employee's duties. It is the responsibility of employees to consult their supervisors, if they have any questions in this respect.

e) **Use by Students**: Use of UIUCnet by students is governed by the Code of Policies and Regulations Applying to All Students.

i) **Responsibility for Passwords**: Students must not share their passwords with others, even with friends. Students are responsible for ensuring that their computers are secure from unauthorized use. When working as employees, students are covered under section d) above.

f) **Use by Non-University Users**: Non-University individuals and organizations may not use UIUCnet, except as specified by written University contract. It is the responsibility of the contracting unit to ensure that content and usage of UIUCnet adhere to all general University policies and that resources are provided in a secure manner. For purposes of this policy, a contracting organization shall be deemed to be a unit of the University, and designated officials of the organization may exercise the responsibilities of University administrators as described in this policy, except that the contracting organization may not exercise or supercede the authority of the CIO.

i) **Limited to University-related activities**: Legitimate non-University users may use their University-provided accounts and Internet access only in conjunction with their authorized University-related activities.

ii) **Authorized Organizations**: UIUCnet resources may be used in support of organizations identified in Authorized Users at the University of Illinois at Urbana-Champaign. While it is appropriate for the home pages of these organizations to provide some information about external organizations, clubs, commercial entities, etc., UIUCnet-connected equipment may not be the primary repository for that information.

iii) **University-sponsored External Entities**: Any University program that, in the interest of collaboration, wishes to provide an external entity with Internet access or to host non-University materials on a UIUCnet-connected server must first consult with CITES about alternatives and secure approval from the Office of the CIO.

iv) **Network Services Limited to Authorized Users at the University of Illinois at Urbana-Champaign**: Except as indicated here, unless permission has been granted in an Allied Agency agreement or otherwise obtained in writing from the CIO or designee, a system connected to UIUCnet must not be used to provide network services or access to any person or organization not identified in the University of Illinois at Urbana-Champaign Authorized Users.

2.5. PROTECTION OF INFORMATION IN ELECTRONIC MEDIA

2.5.1 Status of Information in Electronic Media

Information and data maintained in electronic media on University computer systems are protected by the same laws and policies, and are subject to the same limitations, as information and communications in other media. Before storing or sending confidential or personal information, campus users should understand that most materials on University systems are, *by definition*, public records. As such, they are subject to laws and policies that may compel the University to disclose them. The privacy of materials kept in electronic data storage and electronic mail is neither a right nor is it guaranteed.

2.5.2 Examination of Contents of Electronic Messages and Files

Unless required by law or by authorized administrative approval to do otherwise, campus and unit-level system administrators will not examine the contents of electronic messages and files and will make every reasonable effort to protect them from unauthorized inspection, subject to the following:

a) **Contents of Email**: The contents of electronic messages might be seen by a system administrator in the course of routine maintenance or in order to dispose of undeliverable messages. In addition, electronic mail systems store messages in files (e.g., the file containing a user's inbound mail.) These files are copied in the course of system backups, and these backup copies may be kept long after original messages were deleted.

b) **System Files and Logs**: In the course of resolving system performance or security problems, system administrators may examine the contents of files that control the flow of tasks through the system or that grant unauthenticated access to other systems. This includes systems logs that document some activities of users.

c) **File and Directory Names**: File names and directory names are treated as public information and are not protected.

2.5.3 Process for Requesting Disclosure of Contents of Messages and Files

a) **Requesting Disclosure**: Requests for disclosure must be made in writing through regular reporting channels, consistent with the guidelines below. Requests for disclosure are made to the campus Chief Information Officer (CIO), who is assigned the responsibility for implementing this policy and ensuring that the scope of the disclosure is limited to a legitimate University purpose. The CIO carries out these responsibilities in consultation with Legal Counsel and other appropriate offices. The CIO may designate an individual to act on his or her behalf in fulfilling these responsibilities. All authorizations by the CIO or designee will include specifications for the form and timing of notification to the person whose information is accessed or disclosed.

b) **Action While a Request is Pending**: While a request consistent with this process is pending or under consideration, the requesting unit executive officer may ask computer system administrators to take reasonable, necessary steps to maintain, store, or otherwise prevent the deletion or modification of the information being sought. This must be done in such a way as to maintain the privacy of said information until the requested disclosure is reviewed. The Office of the CIO may be able to advise units on appropriate procedures.

c) **Notification of Affected Individual(s)**: When the CIO or a designated authorized unit administrator provides access to, and disclosure of, email messages and/or file content under provisions of external laws, regulations or applications of this University policy, the requesting administrator will normally notify in *advance* the individual(s) whose information is to be released, indicating the information to be released and the law, regulation or policy that governs the release. If individuals are not notified in advance, the CIO will be responsible for determining when notification is appropriate and for ensuring that such notification is carried out. Circumstances in which notification may be delayed include, but are not limited to, (1) the presentation by legal bodies of subpoenas or other instruments prohibiting advance notification, (2) situations where the safety of individuals is involved, or (3) investigations or inquiries conducted under published University policies.

d) **Conditions for Disclosure**: In the absence of legally compelled access or disclosure, the CIO is authorized to grant access to a user's file contents or electronic mail messages, or to give copies of them to any third party *within* the University only if *all* the guidelines below are met:

    i) The access or disclosure is requested in writing through regular University reporting channels, including the unit executive officer of the individual whose information is being disclosed and the next administrator in that reporting chain.

    ii) The reason for the requested disclosure serves a legitimate University purpose.

iii) The disclosure is not invasive of legitimate privacy interests or unreasonable under the circumstances, e.g., in light of alternative means of acquiring the information or achieving the requester's purpose.

iv) The nature and scope of the disclosure is submitted in writing to and approved by the CIO. This request is normally submitted by the approving executive officer indicated above.

v) The affected individuals are notified in a timely manner in writing of any access or disclosure, consistent with section 5.3c above.

2.5.4 Review of Disclosure:

UIUCnet users whose information is accessed or disclosed under the above provisions should use existing University complaint and/or grievance procedures when concerned about the application of this policy.

2.6. RESPONSIBILITIES IN MANAGING UIUCnet

This section outlines responsibilities in managing UIUCnet that may affect units and individuals.

a) **Network Design**: CITES will work with any unit to develop or modify a network to meet unit needs. Needs directly related to the University's education, research, or public service missions have first claim on resources. Networks serving the Residence Halls, Certified Student Housing and other housing are unusual because of their high density and unique environment. Consistent with University policy, CITES may require special restrictions on their use, if needed to protect the quality of service to those who share these networks.

b) **News Services**: CITES offers news services, which may include news groups originating on- and off-campus, including commercially provided groups of interest to the campus community. CITES will not review in advance or censor the content of the groups provided. However, if the presence of a news group is likely to impact campus services adversely or if a posting is in violation of law, CITES may elect not to distribute that news group or may elect to remove that posting. Examples might include news postings that violate copyright law or news groups that generate burdensome amounts of traffic on the network or campus computers.

i) Which news groups are provided and the retention period for news postings will be at the discretion of CITES or the unit providing the service, based on input from the campus community and the availability of resources.

ii) Professors have the right to moderate their class newsgroups.

iii) A posting that falsifies the identity of the individual making the post constitutes impersonation and thus violates University policy, as well as potentially violating laws and regulations. Members of the community using University resources are expected to identify themselves using their campus network ID.

2.7. NETWORK DESIGN

CITES is responsible for the design or approval of departmental local area networks (LANs) that are connected to the campus network and their connections to the campus backbone. The following subsections document policies and procedures relevant to these areas. The term LAN as used here refers to the routers, switches, repeaters, cabling and patch panels, but excludes servers and other computers.

a) **UIUCnet Address Space**: Only CITES -approved domains may be operated within UIUCnet address space. Publicly accessible Domain Name Servers must be approved by CITES before they are placed in service.

b) **Responsibility for Telecommunications Wiring**: CITES is responsible for the telecommunications wiring system on the University of Illinois at Urbana-Champaign campus. If portions of this system are used in the construction of a LAN, all such use must conform to campus standards.

c) **Local Network Policies**: Network administrators and the owners of local networks may develop their own network policies, as long as they are not in conflict with campus or University policies. Unit-level policies may not restrict access to campus services, except where specific security concerns require it, and may not contravene policies stated here.

d) **Responsibility of Units**: Units are responsible for the uses of their local area networks and servers. In particular, units are responsible for ensuring that materials published electronically or otherwise placed on their servers are relevant and appropriate to the unit's mission.

e) **Licensing and other Restrictions**: Some servers connected to UIUCnet provide services or software that are restricted by licensing agreements to use by University students, faculty, and staff. Some licenses may further limit use to campus, one or more colleges or particular units. Servers must be configured so that restricted services or software are accessible only to those who are eligible.

f) **LAN Administrators**: Each LAN must have at least one designated administrator who is responsible for its administration and management, and whom CITES may contact if it detects a problem.

2.8. NETWORK SECURITY

The security functions of commonly used desktops, servers, and communications technologies are often vulnerable, allowing unauthorized access to or viewing of system resources. A security violation on one machine may threaten security of other systems on the network, allowing unauthorized users to disrupt or damage interconnected systems. Because of this, each individual and unit has certain responsibilities to ensure that their systems are reasonably secure. This section describes security-related roles and responsibilities. It also describes circumstances under which UIUCnet user data can be collected and examined by an individual managing a LAN, server, or system.

a) **Responsibilities of Network Administrators**: It is the responsibility of every network administrator to have expertise sufficient to maintain appropriate levels of security and system integrity on local LANs. The CITES Security Office will document best practices and procedures for maintaining network security and integrity, in consultation with the campus community and peers nationally. CITES provides training, consulting, and general support to network administrators.

b) **Ensuring Integrity of UIUCnet**: In the event that the CITES Security Office judges a LAN to present an immediate risk to the integrity of UIUCnet equipment, software, or data, or presents a risk to the external network (resulting in potential liability for the University), it may terminate or restrict the LAN's network connection without notice. If there is no immediate risk, the CIO Security Office will bring the matter to the attention of the LAN's network administrator. If unable to resolve the problem at this level, the Security Office will contact the unit executive officer or the next level administrator. In addition, if an individual system administrator of a multi-user system determines that an account presents an immediate security risk, he or she may inactivate the computer account without prior notice. The administrator must contact the CITES Security Office in a timely manner to report and discuss the situation. In the course of ensuring the integrity of UIUCnet and local LANs, CITES staff and system administrators, respectively, may use tools, monitoring hardware and software, and log information as indicated here:

    i) **Security Tools**: The CITES Security Office may use tools designed to locate security flaws in equipment connected to the campus network and will take appropriate steps to protect the privacy of data (as provided by this policy) in the process. When the CITES Security Office documents risks to security or network integrity, units are responsible for immediate responses to mitigate or remove the risk. Whether so notified or not, units are responsible for appropriate security with respect to equipment within their LAN, and the faculty within each unit shall be responsible for computer security policy consistent with campus guidelines.

ii) **Network Monitoring Tools**: In order to solve network problems, campus and unit system administrators may employ software or hardware devices from time to time that capture contents of packets traversing the network, including email, Web, and other services. These monitoring tools will be used to monitor and improve the performance or integrity of the network. They will not be used to monitor or track any individual's network activity except under the special authorizations provided for under Section 6.

iii) **System Log Files**: Managers of campus or unit systems and network services may log connections to their machines and services made via dialup or UIUCnet. The information recorded may include the source and destination for a connection, and session start and end times. Operators of multi-user systems may keep logs of activities on their systems. The logs may include login name, timestamps and commands issued. Network administrators may **not** monitor individual users' data or files except under special authorizations provided for under Section 6.

2.9. BANDWIDTH GUIDELINES

UIUCnet and its connections to the Internet are a shared, finite resource. While every effort is made to provide adequate bandwidth for University purposes, bandwidth may not be available for every use.

a) **New Applications**: Extensive use of new applications that require very large amounts of bandwidth on the campus backbone must be discussed with CITES beforehand, so that appropriate planning can take place.

b) **Degrading Network Performance**: If use of a computing or network service by a project or individual seriously degrades network service to others, CITES will try to help the project or individual obtain the needed service in a way that does not seriously impact others. If a network upgrade is required, the unit or user may be asked to pay all or part of the cost.

c) **Responsibilities of Network Administrators**: Network administrators are responsible for monitoring and managing traffic on their LANs to protect the quality of service from adverse impact by users whose applications require substantial bandwidth or other network resources.

CONCLUSION

Using an organization network is a privilege, not a right, so the users must obey the organization rules. The same procedures are used in our organization too, and the users have the same privileges and obligations.

REFERENCES

University CIO Policy page:
http://www.cio.illinois.edu/policies/index.html
University Academic Staff Handbook:
http://www.ahr.uiuc.edu/ahrhandbook/default.htm
Campus Policy and Procedure Manuals at *Looking Up Campus Information* (LUCI) at:
http://www.fs.uiuc.edu/luci/
Code of Policies and Regulations Applying to All Students:
http://www.uiuc.edu/admin_manual/code/code_contents.html

TETRA SECURITY

CPT Ioana MARTIN

INTRODUCTION

Public Safety and Security (PSS) service responders provide society all the indispensable services like police, fire and other emergency. Each individual in our society has the growing expectation of, if not the right to, this kind of services.

In turn, society expects that its government will expend the necessary resources to aid those in emergency need. The provision of emergency services extends beyond the social contract and invokes a moral obligation to protect life, welfare, and property.

Within the public safety organizations, users have a wide range of special requirements that commercial systems often find financially unfeasible to provide. System capabilities include such features as pre-emptive emergency call, Direct Mode, dynamic system reconfiguration, alarm/control, status messages, tracking and monitoring the status of each radio user, monitoring and recording all traffic and a variety of purpose-built mobile data services from satellite positioning to management of resources, queries to confidential databases and sharing of drawings and images etc. The terminals used in public safety networks also often have unique requirement. They need to be extremely robust and often must be able to operate in hostile environments, including high heat, extreme cold, dust, rain and water and not to forget under extreme ambient noise levels. They need very high audio output and must be extremely reliable, being able to withstand shock, vibration and other rugged treatment. They may need to be able to operate in explosive atmospheres. Public safety must have full access to and control of their communications facilities. The command centre or centralized dispatch room must have control of which agencies and personnel are permitted access to the system and to communicate with which agencies, be able to set up talk groups of personnel by work groups, departments and agencies, and must be able to establish the priority levels of all responders, all on a dynamic, moment by moment basis. Such control is critical for public safety operations. It is already because of these functionality reasons that most public safety organisations and ministries in charge of public safety operations conclude that a dedicated radio network based on dedicated Professional Mobile radio technology – such as the TETRA standard – is the best way to build the mobile communication capability for them. But there are more issues to be taken into account.

Unlike commercial carrier networks, designed for general public's use, mission critical networks are created specifically for public safety operational situations. There is an "always available" lifeline to responders providing up-to-date information at all times. In many crisis situations, commercial services simply do not cope. When disaster hits, the public reaches for cellular phones - resulting in overload of networks which often fall over completely. Public safety users simply cannot afford to be without communications, especially during a crisis. Mission critical network are built to 99,999% availability on key components, and the base sites will continue local operation, even if connection to switching centers should be lost completely. Commercial networks do not have that capability. Also power supply is essential different: commercial networks has no or very limited (1-2 hours) backup where dedicated public safety networks can have battery backup and generators to have power continuity for several days. Priority access and the ability to override other users on the system are essential for public safety users. These users cannot wait in queue with non-public safety users on commercial systems in order to complete mission critical calls.

In addition, public safety communications managers must be able to dynamically assign priority among users. Different emergency scenarios will require changing levels of priority for users. Because commercial systems quickly become saturated and channels become unavailable during times of disaster and emergency, public safety users rely on public safety systems to provide the quick access needed for their mission. For example in current systems, users can typically access a channel and make a connection in approximately one-third of a second or less and this would be to a large group of users. The business model for commercial wireless infrastructure is very different from that of public safety agencies. Revenues from commercial carriers are derived from individual users. Therefore, they must build out in areas where there is the largest demand.

However Public Safety systems must be designed for potential incidents that could occur anywhere outside these areas. Systems must also provide good coverage into underground and remote and often inaccessible areas of buildings not required by commercial systems. Other special requirements include communications between aircrafts, marine equipment and the terrestrial system. Public Safety systems also require direct mode operation and vehicular repeaters to supplement coverage.

Maintaining local radio coverage around the radio site in the event of loss of transmission network is something that only the dedicated mission critical networks can provide, this is known as base station fall-back mode.

There are many radio communications technologies in existence around the world[53]. Some are standardised, some are proprietary. Each has its own advantages in its own different applications. Each has security designed to an appropriate level for the chosen markets. However, TETRA is – on one hand - the most secure standardised technology in the world – and on the other hand - available for commercial purchase from multiple manufacturers. TETRA offers more layers of security, and higher security than any other radio communications technology designed for mass commercial deployment[54].

I. TETRA overview

I.1. TETRA standard

TETRA is an open standard developed by the European Telecommunications Standards Institute (ETSI). The main purpose of the TETRA standard was to define a series of open interfaces, as well as services and facilities, in sufficient detail to enable independent manufacturers to develop infrastructure and terminal products that would fully interoperate with each other as well as meet the needs of traditional Professional Mobile Radio (PMR) and Public Access Mobile Radio (PAMR) user organizations.

The initial responsibility of ETSI Project TETRA (now known as ETSI Technical Committee (TC) TETRA) was to deliver as set of standards, under a mandate from the European Commission, for a Digital Trunked PMR communications system that could be deployed in Western Europe. As well as producing these mandatory ETSI deliverables (now completed), TC TETRA's responsibility was, and still is, to make sure that the portfolio of standards continue to be developed in accordance with user needs and priorities.

Standard originated in 1989 as Mobile Digital Trunked Radio System (MDTRS), later renamed to Trans European Trunked Radio, and is called TETRA since 1997. TETRA is the only existing digital PMR standard defined by the European Telecommunications Standard Institute (ETSI).

Among the standard's many features are voice and extensive data communications services. Networks based on the TETRA standard will provide cost-effective, spectrum-efficient and secure communications with advance capabilities for the mobile and fixed elements of companies and organizations. As a standard, TETRA should be regarded as complementary to

[53] During the last decade several standards for Mobile telecommunications have been developed: Global Standard for Mobile Communications (GSM), Digital Enhanced Cordless Telecommunications (DECT), Terrestrial Trunked Radio (TETRA), Universal Mobile Telecommunication System (UMTS)

[54] Wireless Public Safety Communications Network - planning considerations, TETRA Association, June 2008

GSM and DECT. In comparison with GSM as currently implemented, TETRA provides faster call set-up, higher data rates, group calls and direct mode.

The technology solutions chosen to meet user requirements contained in the TETRA standards have been, and continue to be, developed primarily by well know and respected manufacturers[55] who have been serving the PMR market with products and services for several decades.

Although the prime responsibility of ETSI is to develop standards for Europe, many of its standards are also adopted world-wide, as evidenced by the uptake of GSM, the first wireless technology standard to be developed by ETSI. Similarly, TETRA has already been deployed in many regions and nations outside Europe, resulting in TETRA becoming a truly global standard.

There is no doubt that a proprietary technology solution can be brought to market in less time than a solution conforming to a recognized open standard. However, large user organizations, especially those in the public sector, have recognized that some proprietary solutions can meet their needs but the 'tie in' to a single supplier can have significant disadvantages. Even though there are some disadvantages, the main advantages and benefits of adopting an open standard are:

- *Economies of scale provided by a large harmonized market served by several independent manufacturers and suppliers competing for the same business resulting in competitively priced solutions*
- *Second source security if existing suppliers exit the market*
- *Evolution (instead of revolution) of the technology standard ensuring longevity and good return on investment for both users and suppliers*
- *Choice of manufacturers for new products keeping prices down*
- *Greater choice of products for specialized applications*
- *Greater responsiveness to future needs by existing suppliers because of competition*

Because there are several independent manufacturers of both TETRA network infrastructure and radio terminals all the benefits of standardization listed also apply to the TETRA market.

TETRA is an evolving standard and it has been developed in Releases (phases) known as TETRA Release 1 and TETRA Release 2. Even though both TETRA Releases have been completed, work continues within ETSI Technical Committee (TC) TETRA to further enhance the standard thus satisfying new user requirements as well as gleaning the benefits of new

[55] Details of manufacturers can be viewed on the member's page of the TETRA Association by clicking on the left hand menu under core products

technology innovations. Outside of Europe the ETSI TETRA Standard has been formerly adopted in China and South Korea.

I.2. TETRA market

Since the first generation of networks were deployed in 1997, hundreds of TETRA networks have been deployed across the world. Even though a considerable number of these networks are deployed in Europe, a rapid uptake is occurring in the regions of Asia, Middle East and South America. Although all PMR market segments are already being served by TETRA, the largest market is that of public safety, where the trend is for the deployment of nationwide networks shared by all public safety organizations for reasons of economics (sharing), autonomy of operation for routine communications and the ability to fully interoperate with other services during emergency situations and disasters.

Romania is one of the countries in which TETRA-system is in use *nation-wide* by the public sector, STS (Special Telecommunications Service) being one of the customers.

The transportation market is the next fastest growing market, especially for Mass Rapid Transport systems and major Airports. Interestingly, TETRA is also used by the military for non-tactical operations, a market application not originally anticipated for TETRA.

The success and market uptake of TETRA has attracted many independent manufacturers and suppliers of TETRA products and services, thereby providing users with healthy competition, second source security and wide choice of radio terminal equipment for specific applications. The success of TETRA has also created a strong base of application developers who are able to provide a wide variety of applications for use with TETRA.

Recognizing that important market requirements outside the responsibility of ETSI needed to be addressed to ensure the success of TETRA, a number of organizations formed the TETRA MoU (Memorandum of Understanding) Association in December 1994. Since it has been established, the TETRA Association has grown significantly and now provides a forum which acts on behalf of its members, being user organisations, manufacturers, application providers, integrators, operators, test houses, regulators, consultants, etc. The main objectives of the TETRA Association are to promote the TETRA standard and to ensure multi-vendor equipment interoperability.

I.3. Technology Benefits

The core technologies used in the TETRA standard, such as Digital, Trunking and Time Division Multiple Access (TDMA) also provide a number of inherent advantages and benefits as follows:

Digital

Nowadays, practically everything electronic uses digital technology and wireless communications are no exception. Even though analogue FM PMR communications will remain a viable option for several years, digital radio provides relative advantages and disadvantages in the important performance areas of:

- *Voice Quality*
- *RF Coverage*
- *Non-Voice Services*
- *Security*
- *Cost*

Trunking

Trunking techniques have been used for many years in switched telephone networks. The first trunked mobile radio communication systems were deployed as early the 70's in North America with proprietary signaling protocols and shortly afterwards in Europe using analogue MPT1327 technology. The main benefit of trunking is normally seen as spectrum efficiency, or more radio users per RF channel compared with a conventional radio channel for a given Grade of Service (GoS), brought about by the automatic and dynamic assignment of a small number of communication channels shared amongst a relatively large number of users.

Because trunking systems support more radio users than conventional systems, national administrations actively support the deployment of trunking systems as this helps reduce pressure on meeting PMR spectrum demands. However, from a radio users operational point of view, spectrum efficiency does not really mean anything. What users want is to solve all the operational problems associated with conventional PMR, yet still retain the simplicity of conventional open channel 'all informed net' operation. The fundamental element of trunking that solves these conventional PMR problems is the use of a control channel. Table 1 below lists the operational problems of conventional PMR and how the use of trunking solves these problems.

Conventional PMR Problem	**Trunking Solution**
Contention	All call requests are handled on the control channel for immediate call processing or in order of queue priority if the system is busy.
Manual Switching of Channels	Automatic cell handover takes away the need for manual channel selection
Inefficient Channel Utilization	The automatic and dynamic assignment of a small number of

	communication channels shared amongst a relatively large number of users ensures an equal grade of service for all radio users on the system.
Lack of Privacy	The dynamic and random allocation of channels makes it more difficult for a casual eavesdropper to monitor conversations.
Radio User Abuse	Abuse is minimized as the identity of all radio users and the time and duration of messages are known and can therefore be easily traced to the abuser.

Table 1: Conventional PMR problems solved by Trunking

It is important to note that the operational simplicity of conventional PMR ‘all informed net’ talk group communications is still retained by employing fast call set-up “Push To Talk” (PTT) operation on radio terminals.

Additional Services and Facilities

As the control channel acts as a signaling communications link between the Trunking Controller and all mobile radio terminals operating on the system, the Trunking Controller knows the status of the system at any moment in time as well as its historic usage, which is stored in its memory. For example, the Trunking Controller knows:

- *The individual and group identity of all radio units registered on the system*
- *The individual identity and time radio units registered on the system*
- *The individual identity and time radio units de-registered from the system*
- *The individual and group identity, time and duration of all messages*

With additional intelligence in both the radio terminals and the trunking controller the advantages and benefits of trunking can be increased. For example, the length of the control channel signaling messages can be increased by a set amount to accommodate a variety of new services and facilities. Also, the trunking controller can be programmed to handle calls in a variety of ways as required by the operator of the system.

Time Division Multiple Access (TDMA)

A four time slot TDMA technology was adopted in TETRA as it offered the optimum solution to balance the cost of equipment with that of supporting the services and facilities required by user organizations for a medium to high capacity network providing single site local RF coverage and/or multiple site wide area RF coverage.

RF Spectrum efficiency is a combination of three main factors being the occupied bandwidth per communication channel, the frequency re-use factor determined by the Carrier to Interference protection ratio C/I in dB’s and the trunking technology used. As previously mentioned TETRA utilizes the latest in trunking technology. Also, the TDMA technology used in TETRA provides 4 independent communications channels in a 25 kHz RF bandwidth Channel, making it twice as efficient in occupied bandwidth terms as a traditional 12.5 kHz RF bandwidth FDMA channel. Although FDMA technologies tend to have a better C/I performance than TDMA TETRA, the overall spectrum efficiency advantage lies with TETRA, especially for medium to high capacity networks.

Because of using TDMA technology, the cost and equipment space at base station sites can be significantly reduced compared with traditional FDMA technology trunking solutions. Another advantage of TDMA technology is that it enables new services and facilities to be supported with minimum cost. Some examples are:

Higher Data Rates

The 'laws of physics' limits the maximum data rate in a given RF channel bandwidth. Assuming the same modulation scheme, the wider the channel bandwidth the higher the data rate. Because TDMA uses wider channels than FDMA, the combined data rate on a single RF carrier is greater.

Improved Data Throughput in Poor RF Signal Conditions. The net data rate in TDMA is better than FDMA in poor RF propagation conditions. This is because Automatic Repeat Requests (ARQ's) are required when received data is corrupted as a result of RF fading. As TDMA terminal devices effectively operate in full duplex ARQ's can be sent efficiently after each time slot transmission instead of waiting until the end of each voice transmission, as is usually the case with FDMA.

Bandwidth on Demand

In TDMA any number of time slots up to the maximum limit of the technology being employed can be combined to increase data throughput as required for specific applications.

Concurrent Voice and Data

Because of the TDMA time slot structure it is possible to assign one time slot to support voice and the next time slot to support data in a two slot transmission from radio terminals. This capability effectively allows a single radio terminal to concurrently transmit or receive voice and data at the same time.

Full duplex Voice Communications

TDMA technology inherently supports full duplex communications. Although full duplex voice communications can be supported on FDMA systems the need for duplex operation requires RF screening between the transmitter and receiver and also a duplexer to allow single antenna working. Because of this, duplex FDMA radio terminals are usually bulkier and more costly to produce than TDMA terminals, which do not need RF screening or antenna duplexers.

II. TETRA SECURITY FUNCTIONS

TETRA contains a wealth of security functions designed to protect users' information. This information can consist of the users' speech and data traffic and also other information that relates to the identities and operations of the users themselves. When describing these TETRA security functions it is important to make a distinction between the different categories of functions and their specific application. In TETRA the following categories can be identified:

- **Security mechanisms -** independent self-contained functions that aim to achieve a specific security objective such as confidentiality of information or authentication of mobile terminals. Security mechanisms are the main building blocks for a security system.
- **Security management features -** are used to control, manage and operate the individual security mechanisms. They form the heart of the security and should guarantee that the security features are integrated into a consistent security system. Furthermore they are used to realize interoperability of the security mechanisms over different networks. Key management is the most essential security management function.
- **Standard cryptographic algorithms -** are standardized system specific mathematical functions that are used, normally in combination with parameters called "cryptographic keys", to provide an adequate security level for the security mechanisms and the security management features. Standardized cryptographic algorithms are offered in TETRA to support interoperability between different TETRA systems.
- **Lawful interception mechanisms -** are used within communication systems to provide the lawfully required access to information and communication, with the aim to fulfill national regulatory requirements. It is essential that such functions do not undermine the regular security of the system. Therefore these functions should be controlled through the security management.

II.1. Security mechanisms

II.1.1. Mutual authentication over the air interface

The TETRA standard supports the mutual authentication of a Mobile Station (MS) and the network, which is in TETRA normally referred to as the Switching and Management Infrastructure (SwMI). This makes it possible for a TETRA system to control the access to it and for an MS to check if a network can be trusted.

In TETRA, as in most other secure systems, the authentication process provides a firm basis for the overall security. It can be used for the following purposes:

- To ensure correct billing in Public Access systems;
- To control the access of the MS to the network and its services;

- To derive a unique session encryption key, the Derived Cipher Key (DCK) which is linked to the authentication, and which is then used to provide confidentiality of information transfer;
- To create a secure distribution channel for sensitive information such as other encryption keys;
- To control the disabling and enabling of an MS/SIM is a secure way;
- To ensure that TETRA MSs are connected to a legitimate TETRA system.

This mutual authentication security mechanism is available for Voice and Data. In Direct Mode Operation (DMO) an explicit authentication mechanism is not available (MSs do not share their authentication keys with each other). In this case the use of Static Cipher Keys (SCKs) can however provide implicit mutual authentication. There is a single standardized authentication algorithm set.

Mutual authentication is done on the basis of an authentication key K, which is unique for every MS or SIM if the latter is used. The K is both stored in the MS/SIM and in the network. Normally a specific network element is used to store the Authentication keys. This is called the Authentication Centre (AUC).

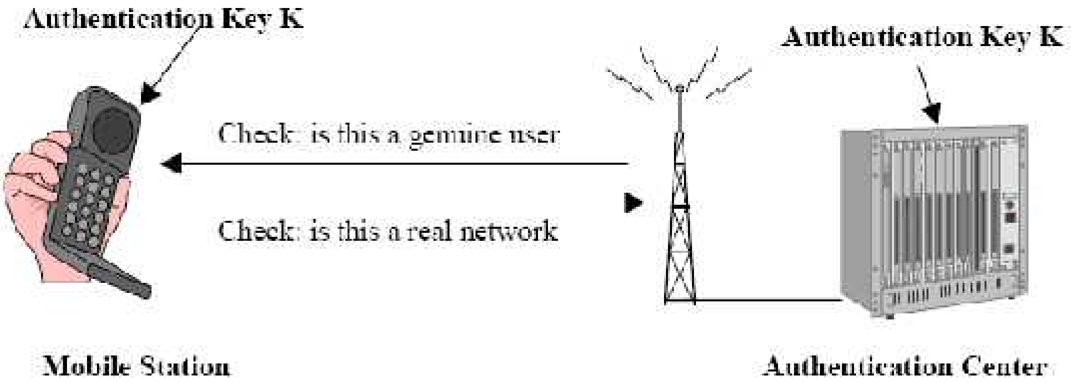


Figure 1: Mutual authentication

II.1.2. Encryption

The air interface is very vulnerable to eavesdropping and so modern mobile wireless communication systems need to have some form of air interface security. This air interface security is intended to secure the connection between MSs and the network. Air interface security is an effective means to provide security in a mobile network and some essential security functions can only be realized by air interface security. In most cases it is sufficient to rely on air interface security and take no further security measures. However, in TETRA systems needing a very high level of security, additional security may be required to protect information transmitted from one MS to another not only over the air interface but also within the network. In this case end-to-end security provides an efficient solution.

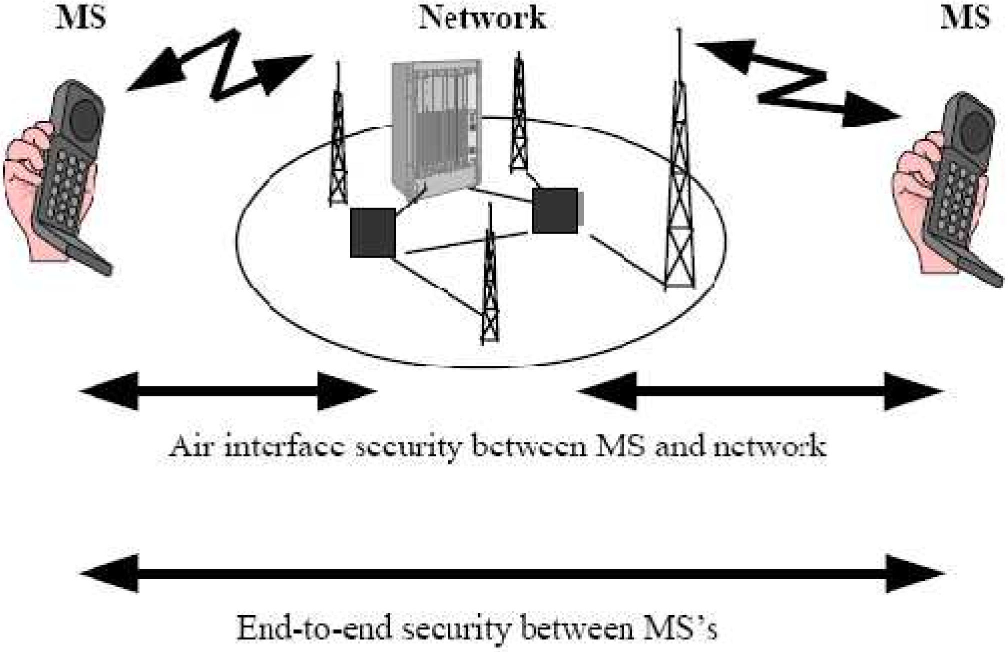

MS
Network
MS
Air interface security between MS and network
End-to-end security between MS's

Figure 2: Air interface security versus end-to-end security

II.1.2.1. Air interface encryption

User traffic and signaling information can be encrypted over the air interface between the MS and the SwMI, both for individual and group communications. The Air interface encryption mechanism is available for Voice and Data in Trunked Mode Operation and in Direct Mode Operation. The use of several encryption algorithms, both standard and proprietary, is supported.

Traffic encryption protects user speech and data. Signaling encryption provides protection from traffic analysis, and prevents an eavesdropper from discovering who is operating in a particular area, or who is calling who.

II.1.2.2. End-to-end encryption

The TETRA end-to-end service can be realized in any number of ways. This means that a user may easily tailor an end-to-end encryption system to their particular requirements. This flexibility is essential for a standard like TETRA that will be implemented in many forms for different user groups.

Public Safety organizations will have specific (high) national security requirements for their implementation of end-to-end encryption, which will be different from the requirements of Military user groups, which have even greater security requirements. All such organizations need to be able to specify an end-to-end encryption system according to their own requirements. It can also be expected that commercial user groups will have a need for secure end-to-end encryption systems.

Whereas the TETRA standard leaves the implementation of End to End encryption relatively open, it is important to realize that there are benefits in having standardized solutions. A standardized solution means that end users, even those who have particular requirements over the cryptography used, do not need to specify the rest of the end-to-end system (including the Key Management). This has led to the production of TETRA Association Security and Fraud Prevention Group (SFPG) Recommendation 02. This Recommendation fully specifies all that is required for an end-to-end service other than the detail of the cryptographic algorithms. These are treated as black-box functions.

In order to provide a complete solution for the general user, the Recommendation concludes with Appendices showing how these cryptographic functions can be realized by using sample implementations of publicly available algorithms. The first sample implementation used the International Data Encryption Algorithm (IDEA), which was a very well respected algorithm at the time, and an agreement was set up to allow reasonable use of the IPR. However more recently due to TETRA market demand a second sample implementation has been made using the Advanced Encryption Standard (AES), which is becoming widely adopted by many government users in Europe and elsewhere. AES has the advantages of being a newer design and of being IPR free.

Although these algorithms are described as sample solutions, in practice the choice of well respected public domain algorithms that have stood the test of publicly available cryptanalysis means that these solutions are completely acceptable for the majority of potential users of TETRA End to End encryption. The advantage of their adoption is the availability of MSs and key management solutions from multiple manufacturers.

II.1.3. Anonymity

The TETRA standard incorporates a mechanism for encrypting users' individual and group identities before transmitting these across the air interface. It is possible to make this encryption dynamic in the sense that an identity is encrypted in a different way on different occasions. This provides anonymity for the end users, and protection from traffic analysis.

Again, this mechanism is available for Voice and Data in Trunked Mode Operation and in Direct Mode Operation.

II.1.4. Secure enabling and disabling of terminals

TETRA supports different options for a direct secure disabling or enabling of either:

- the MS equipment, based on the Terminal Equipment Identity (TEI);
- the MS subscription, based on the Individual TETRA Subscriber Identity (ITSI);
- both the MS equipment and the MS subscription.

The purpose of providing separate mechanisms for equipment and subscription allows a practical means of disabling an MS even if the implementation places the ITSI on a separate SIM card, which is inserted into a Mobile Equipment to make a complete MS. The mechanisms allow the system operator to choose either the ITSI or the equipment, or both together.

If the TEI is disabled the MS's equipment cannot be used any more, even if another ITSI is inserted into the MS. If the ITSI is disabled an MS's equipment can still be used in combination with another (enabled) ITSI, whereas the ITSI cannot be used in any MS anymore. In addition the disabling can be either temporary (which leaves the possibility to enable again over the air) or permanent (which is irreversible). In systems demanding a high security, disabling and enabling should only take place after mutual authentication has been performed. If this is not the case the feature (especially disabling) can obviously be used to attack the system.

II.2. Security management features

The mere fact that security functions are integrated in a system does not automatically imply that a system is fully secure. However, what is normally achieved is that the security risks are "condensed", that is they are concentrated to specific elements in the system, which can be adequately controlled.

This control is one of the tasks of the security management. Another task of security management is to guarantee that the security mechanisms are used in the proper way and that the different mechanisms are integrated in an appropriate way to achieve an overall secure system. Security management is also responsible for realising the secure interoperability between different (TETRA) systems.

The form into which the security is condensed is normally that of "keys". A key is a piece of secret information that is used, often in combination with cryptographic algorithms, to provide the actual security for a security mechanism. Often the keys form the interface between security management and the security features. Security management is responsible for dealing with the keys in a secure way. Though security management is partly an issue for the implementation, in communication systems like TETRA it is possible to specify certain management features, which support the security management. In addition the TETRA Association SFPG has produced Recommendations intended to support the management of security (especially key management).

Adequate security management is just as important as the actual security mechanisms. In TETRA key management, functionality and flexibility are key words. A large number of features have been integrated to support the key management.

II.2.1. Authentication Key

The authentication key K is used for mutual authentication between an MS and the SwMI. There are three possible methods for generating K which are outlined below.

Method 1 – Generation of K from an Authentication Code (AC)

In this case the user types in an Authentication code via the keyboard of the handset. The digits of the AC are represented as a string of bits. This bit string is translated using an algorithm to the key K.

The AC is normally not stored in the handset. In the Network (Authentication Centre) either the K or the AC is stored. In the latter case the K is derived form the AC every time this is needed. This method is used if it is needed to identify the user of a handset, but not the handset. It should be noted that the AC would normally have much less then 128 information bits. Therefore this method for generation of K should only be used in exceptional cases, e.g. if there is a need for user authentication only or if a key needs to be generated immediately and there is no possibility to use a User Authentication Key (UAK - see below).

Method 2 - Generation of K from an User Authentication Key (UAK)

The User Authentication Key is an unpredictable (random) value of any desirable length (usually 128 bits). The K is derived from the UAK using an algorithm. The UAK or (normally) the K is stored in the handset (or SIM) and the network (Authentication Centre). If the UAK is stored then every time the K has to be derived from it. This method is used if it is needed to identify the handset. It will be the most common method of key generation in TETRA systems.

Method 3 - Generation of K from an Authentication Code (AC) and an User Authentication Key (UAK)

In this case the K is derived from an AC entered by the user via the keyboard of the handset and a UAK stored in the handset. The derivation of K from AC and UAK is done via an algorithm. In the network either only the resulting K is stored, or both the AC and UAK are stored. This method is used if it is necessary to identify both the user and the handset.

II.2.2. Keys for air interface encryption

There are several sorts of encryption keys. Some keys may be derived or transferred as part of the authentication procedure, some keys can be sent to MSs using Over The Air Re-keying (OTAR) or some may be preloaded in the MSs. There are keys with long term and short term key lifetimes. Special mechanisms are included to protect the keys with a long lifetime.

- The **Derived Cipher Key (DCK)** is derived during the authentication procedure. It can be used to encrypt the link between the network and the MS on an individual basis. Thus it can also provide an extended implicit authentication during the call, and can be used for encryption of uplink communications (i.e. the communication from the MS to the network) as well as downlink communications from network to an individual MS.

- The **Common Cipher Key (CCK)** is generated by the SwMI and distributed, encrypted with the DCK, to each MS. It is efficient to use this key for encryption of messages that are directed to groups of MSs spread across one or more Location Areas (LAs). When the CCK is distributed to an MS over the air interface using OTAR it is encrypted with the DCK of this MS.

- The **Group Cipher Key (GCK)** is linked to a specific closed user group. It is generated by the SwMI and distributed to the MSs of a group (e.g. by pre-provisioning of the MS, on a Smart card, or by using OTAR (see below)). Within a Location Area the GCK is always used in a modified form. It is combined with the CCK in a specific algorithm to obtain the Modified Group Cipher Key (MGCK). The MGCK is used to encrypt the closed user group messages for groups of MSs. When the GCK is distributed to an MS over the air interface using OTAR it is encrypted with a session encryption key derived from the Authentication Key for this MS, or with a Group Session Key.

- The **Static Cipher Key (SCK),** finally, is a predetermined key, which can be used without prior authentication. It is “static” in the sense that it is a fixed key that is not changed by another security function (e.g. by an authentication exchange) until it is replaced. TETRA supports the use of up to thirty-two (32) SCKs in an MS, per network. They can be distributed similarly to the GCKs. Their use is largely implementation dependent but they can be used for encryption in Direct Mode Operation (where they may also provide explicit authentication) and in certain TETRA systems also for encryption for group and individual communications. The SCK may also be used in a system that normally uses DCKs and CCKs as an alternative to those keys in fallback conditions. When an SCK is distributed to an MS over the air interface using OTAR it is encrypted with a session encryption key derived from the Authentication Key for this MS.

When used in DMO, SCKs may be grouped in a way that allows several SCKs to be associated with the same talkgroup(s). This allows an MS to have a current SCK defined for transmission, but to allow reception on one of the others. This allows a practical key management mechanism to be constructed, where one MS may be commanded to start using a new SCK for transmission before the changeover message has reached another MS.

II.2.3. Over The Air Re-keying (OTAR)

As indicated above there is a possibility to distribute or update CCKs, GCKs and SCKs using a Over The Air Re-keying (OTAR) mechanism. This mechanism makes it possible to send air interface encryption keys in a secure way from the SwMI over the air directly to an MS and can be applied provided that an authentication key K is available for the MS. The OTAR messages for an individual MS are encrypted using session encryption keys that are derived from the authentication key for that MS. Alternatively, a Group Session Key for OTAR may be used to distribute keys to groups of MSs at the same time.

A similar OTAR mechanism is also available for the management of end-to-end encryption keys. This is usually referred to as Over The Air Keying (OTAK) to distinguish it from the air interface service.

II.2.4. Transfer of authentication information between networks

If a TETRA MS roams to a TETRA network other than its “home” network, this “visited” TETRA network will need to obtain authentication information from the “home“ network of this MS in order to be able to perform mutual authentication and generate and/or distribute encryption keys. The transfer of authentication information between networks is in principle supported in three ways. The most straightforward method is to simply transfer the authentication key K to the visited network. For security reasons this is however not advisable. A second option is to transfer certain information that can be used for one single authentication procedure. This is basically the same method as is applied in GSM and can be implemented in a very secure way. However this is only practical where the MS cannot mutually authenticate the SwMI – otherwise the visited SwMI would have to interrogate the home SwMI for a response each time the MS invoked this mutual authentication. A third alternative is therefore supported. This allows a home network to transfer a set of session authentication keys for an MS, which can be used for repeated authentications, to a visited network without revealing the original authentication key of the MS. This option combines security and efficiency and permits mutual authentication to take place at a realistic pace.

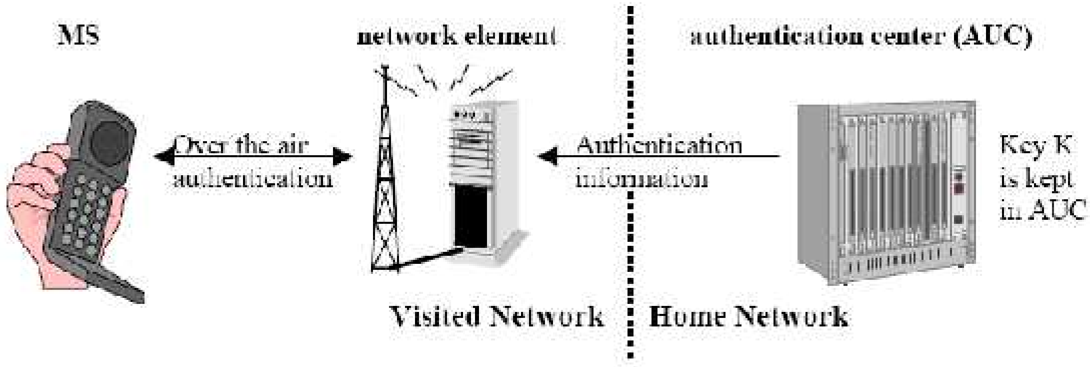

MS
network element
authentication center (AUC)
Over the air authentication
Authentication information
Key K is kept in AUC
Visited Network
Home Network

Figure 3: Authentication in a visited network without disclosing the authentication key

II.3. The standard TETRA cryptographic algorithms

The TETRA standard offers a number of standard cryptographic algorithms which all have their own specific purpose. This section explains this purpose and the use of these standard algorithms.

II.3.1 Air interface encryption algorithms

TETRA users can specify their own air interface encryption algorithm. However, for reasons of easy interoperability in multi vendor systems, a number of air interface encryption algorithms have been specified as part of the TETRA standard. Several requirements have been taken into account when specifying these standard algorithms. The most important of these are the need for diversity and export control regulations.

Need for diversity

There will be a wide range of TETRA networks and applications. Not all users want to ‘share’ their standard encryption algorithms with all other TETRA users. For example, the European Public Safety Organisations (associated with the European Schengen organisation) require their own standard air interface encryption algorithm.

Export control regulation

Equipment that includes encryption algorithms is likely to be subject to specific export controls in addition to any other functional controls. The encryption related controls are slowly being relaxed. Such controls are country specific, but 33 major industrial countries derive their national controls from a commonly agreed policy. This policy is published under the banner of the Wassenaar Arrangement[56].

[56] http://www.wassenaar.org

Four standard encryption algorithms are currently available for use in TETRA systems. These have been developed by ETSI's Security Algorithm Group of Experts (SAGE) to two different criteria. These are explained below.

TEA2 and TEA3: Restricted Export Algorithms

These algorithms are controlled items under the 1998 Wassenaar Arrangement rules. The algorithms have been primarily designed for use by Public Safety Organisations. The former algorithm (TEA2) has been assigned for use by Public Safety Organisations in Schengen and related countries.

TEA1 and TEA4: Readily Exportable Algorithms

TEA1 (as the numbering implies) was an early design. TEA4 reflects the more relaxed controls of the 1998 Wassenaar Arrangement.

The standard TETRA Encryption Algorithms are available to TETRA users and manufacturers. They are distributed by a custodian. In case of the TEA1, TEA3 and TEA4 the custodian is ETSI[57]. The TEA2 is distributed by the Dutch Police IT organisation.

II.3.2 Air interface authentication and key management algorithms

There is also a set of standard air interface authentication and key management algorithms, designed to allow easy interoperability in multi-vendor systems, which have been specified as part of the TETRA standard.

The requirements on diversity and export control regulations do not exist in the case of authentication and key management algorithms. Therefore, only a single set of standard air interface authentication and key management algorithms has been specified. This algorithm set is called the TAA1. Its specification is distributed by its custodian, which is also ETSI.

II.3.3. End-to-end encryption algorithms

SFPG Recommendation 02, which describes a standard End to End encryption implementation is written around four black-box cryptographic functions designated E1 to E4. Those users with the necessary expertise may define how these are realized using algorithm(s) of their own choice. The only constraint is that the algorithm(s) have to fit within the broad parameters of functions E1 to E4. For those users who are content to follow a public standard, the recommendation includes Appendices which shows how these cryptographic functions can be realized using the IDEA or AES algorithm. So, the body of the recommendation together with the appendix forms the complete specification for a standard TETRA end-to-end encrypted voice service. The IPR for IDEA is owned by MediaCrypt AG, who should be approached for

[57] see http://www.etsi.org , section algorithms and codes

licensing information. AES has the advantages of being a newer design and of being IPR free and is becoming widely adopted by many government users in Europe and elsewhere.

II.4. Lawful interception mechanisms

In most European countries there is an obligation on operators of public (and sometimes private) telecommunication networks to provide lawful interception facilities to the responsible national authorities. Since a standardized solution is much more cost efficient than proprietary implementations on a case by case basis, it was decided to provide support for lawful interception within the TETRA standard. A subgroup of the TETRA security group has specified the requirements for a Lawful Interception Interface to support the mechanisms for lawful interception. The detailed implementation of this interface might differ on a country to country basis.

III. FUTURE PROSPECTS

The TETRA industry is now investigating the evolutionary path through which TETRA network operators can offer broadband wireless services to mission critical users. In the short term, the proposal is for operators to enable inter-connection between TETRA and complementary networks (e.g. broadband networks based on WiMAX or Wi-Fi). This requires users to have separate radios for each network and operators to provide gateways to support the required degree of interconnection between these networks. In the longer term, there is a proposal to evolve TETRA towards a “Broadband TETRA” solution.

As users became interested in developing richer applications, the TETRA Association responded by working with ETSI to develop the TEDS (TETRA Enhanced Data Service) standard, a wideband data solution which enhances TETRA with a much higher capacity and throughput for data. Today, there is a high concentration of mature data applications within the capabilities of narrowband and wideband radio bearers. Indeed, public safety agencies and commercial organizations are leveraging the capabilities of SDS (short data service), PD (Packet Data) and MSPD to support applications ranging from telemetry to incident management. Significantly, the bandwidth requirements of most of these applications are within the capabilities of TEDS.

A TEDS capable network, be it a TEDS upgraded TETRA network or a separate overlay data network, will yield significantly more spectrum efficient data traffic.

One TEDS carrier will support as much as 40 times as much data as a single-slot carrier, and 10 times as much as a multi-slot carrier. TEDS will multiply the over-the-air SDS capacity of a base station by tenfold.

This means that the base stations can handle the traffic even at peak times such as during incidents when the movements of a very large number of operatives need to be tracked in the area of one or two base station sites.

A TEDS capable network can support high volumes of transactional narrowband data and provide high speed data services, such as transmitting photographic images, electronic reports from the field, and person identification information.

TETRA MoU has published studies, focusing on the usability of TETRA vs UMTS[58] or TETRA vs GSM-ASCI[59] for the PSS segment. The issue of possible use of public mobile networks and GSM derivatives for Public Safety radio communication has been addressed in various studies during recent years. The conclusions made in the studies vary radically from in favour to against. The readers are encouraged to make their own judgment.

The conclusion is that, even in the future, specialized technologies like TETRA will survive. The main reasons are security arguments, since dedicated networks always offer a better quality of service for mission- critical users than shared networks. Additionally, technical requirements most probably weigh stronger than economic arguments when choosing a solution for this segment.

Compared to a dedicated solution, the amount of network elements is much higher in a commercial network, which makes potential upgrades very expensive and risky. Solutions purely based on mainstream technologies will be able to attract customer groups with lower quality of service requirements. These technologies are able to provide very basic group call functionalities at a very competitive price level, since the modifications to the core network are relatively small.

However, for the mission- critical PSS segment, there will still be a demand for dedicated networks with specialized functionalities. It remains to be seen whether the PSS market is large enough for telecommunication equipment manufacturers to focus on a specific solution. The tendencies during the last years have shown that most probably only a few infrastructure and mobile manufacturers will remain. In order to keep the market size as large and homogeneous as possible, it is very likely that several manufacturers will choose the same technologies for their solution. Whether in the future this will be a standardized solution or a de-facto standard is difficult to estimate for the time being.

Radio IP Software[60] chose the 2009 TETRA World Congress in Munich to demonstrate for the first time, next-gen Mobile VPN technology advancements, namely “Concurrent VPN”. For TETRA users, it is important to understand that a mission-critical mobile VPN makes it possible

[58] TETRA MoU Association; TETRA or UMTS – let the user decide; 2001

[59] TETRA MoU Association; TETRA or GSM-ASCI network for Public Safety – let the users decide; 2004

[60] TETRAnews, information from the TETRA Association issue 3/2009

for mobile workforces to securely combine a TETRA voice network with IP data applications, utilizing a laptop or a PDA for access from the field. Mobile VPNs typically provide security features such as user authentication, device authentication and data encryption. Mobile VPNs improve connectivity such that application sessions remain uninterrupted even when network connections temporarily drop due to signal loss. The software seamlessly roams between one network to another, allowing mobile workforces to combine their PMR, or TETRA infrastructure with other cellular or Broadband IP networks to take better advantage of emerging bandwidth-hungry applications and extending coverage where needed. IT Administrators are able to set policies to better manage which network will carry a specific application's data based on priority, application need, time of day or cost of access. However, existing roaming capabilities have their limitations. Mobile workforces are able to roam from one network to another, but are not able to access both at the same time – a real problem for emergency responders because mission-critical data applications, such as computer-aided dispatch (CAD) due to their very nature should generally be ensured to transmit over the most reliable wireless network available. Bandwidth intensive applications with less severe security and resilience requirements (for example, video streaming) should be able to concurrently transfer data over 3G or broadband networks without impacting or being impacted by applications that are transmitting over mission-critical PMR systems like TETRA at the very same time. With Concurrent VPN technology, due to be widely available in 2010, users will soon be able to transmit over multiple networks simultaneously, thus maximizing their infrastructure resources and saving time and money. Concurrent VPN bridges the best features of multiple networks and creates what amounts to a single "network of networks" to achieve a high level of performance, reliability and trusted access.

CONCLUSIONS

To use a technology in a mission critical environment, the technology must be designed to work in that environment from the start. In particular, the security design must be adequate and appropriate. TETRA has been designed specifically to meet the challenges of the unique mission critical environment, providing essential functions such as group call communications and Direct Mode Operation and providing the specific security and security management functions to allow the complete environment to remain secure.

No public cellular technology has been designed for this environment, which leads to dangerous compromises and loss of security when attempts are made to adapt such technologies to this demanding environment. Even if a commercial network was designed to meet the needs – functional, operational, resilience, QoS, etc. of PSS users – most governments would still want to ensure that ownership of the operator would be under their control (selling the shares to unwanted parties is not considered acceptable in many countries). In addition they may require continued guaranteed financial viability of the operator and rights to take management control of the operator if needed. Conflicts of interest between public safety and general public use could not be permitted. Most existing specialized operators of that type have these types of legal constraints. Many Governments furthermore feel they need to retain control over the radio spectrum to achieve that level of control. Governments will also require longevity of supply and maintenance – a dedicated technology with multiple vendor support is the only way of achieving this. Public systems are great for communication from/to the public.

TETRA is a purposebuilt high capacity solution and the tool for the Emergency Services Radio Communication. During disasters or other crisis situations the Emergency Services needs the maximum performance and reliability from the system they use. This is exactly the situation when commercial networks become stressed beyond their limits.

REFERENCES

1. *TEDS: Enabling the Next Evolution of Mission Critical Data Applications*, Motorola white paper
2. *TETRA or GSM-ASCI network for Public Safety*, TETRA MoU Association, May 2004
3. Carmine Rizzo (ETSI) and Charles Brookson (BIS UK), *Security for ICT - the Work of ETSI,* third edition - December 2009
4. *Wireless Public Safety Communications Network - planning considerations,* TETRA Association, June 2008
5. Simon Riesen*, The usage of mainstream technologies for public safety and security networks,* Helsinki University of Technology, October 2003
6. http://www.tetramou.com/
7. http://www.telecomasia.net/
8. http://www.tetra-association.com
9. http://www.etsi.org
10. http://www.seminarprojects.com/Thread-terrestrial-trunked-radio-tetra#

ALPHABETICAL INDEX OF AUTHORS