

COMPUTER SECURITY AND ENCRYPTION

AN INTRODUCTION



S. R. CHAUHAN & S. JANGRA

COMPUTER SECURITY AND ENCRYPTION

LICENSE, DISCLAIMER OF LIABILITY, AND LIMITED WARRANTY

By purchasing or using this book (the “Work”), you agree that this license grants permission to use the contents contained herein, but does not give you the right of ownership to any of the textual content in the book or ownership to any of the information or products contained in it. *This license does not permit uploading of the Work onto the Internet or on a network (of any kind) without the written consent of the Publisher.* Duplication or dissemination of any text, code, simulations, images, etc. contained herein is limited to and subject to licensing terms for the respective products, and permission must be obtained from the Publisher or the owner of the content, etc., in order to reproduce or network any portion of the textual material (in any media) that is contained in the Work.

MERCURY LEARNING AND INFORMATION (“MLI” or “the Publisher”) and anyone involved in the creation, writing, or production of the companion disc, accompanying algorithms, code, or computer programs (“the software”), and any accompanying Web site or software of the Work, cannot and do not warrant the performance or results that might be obtained by using the contents of the Work. The author, developers, and the Publisher have used their best efforts to insure the accuracy and functionality of the textual material and/or programs contained in this package; we, however, make no warranty of any kind, express or implied, regarding the performance of these contents or programs. The Work is sold “as is” without warranty (except for defective materials used in manufacturing the book or due to faulty workmanship).

The author, developers, and the publisher of any accompanying content, and anyone involved in the composition, production, and manufacturing of this work will not be liable for damages of any kind arising out of the use of (or the inability to use) the algorithms, source code, computer programs, or textual material contained in this publication. This includes, but is not limited to, loss of revenue or profit, or other incidental, physical, or consequential damages arising out of the use of this Work.

The sole remedy in the event of a claim of any kind is expressly limited to replacement of the book, and only at the discretion of the Publisher. The use of “implied warranty” and certain “exclusions” vary from state to state, and might not apply to the purchaser of this product.

COMPUTER SECURITY AND ENCRYPTION

An Introduction

S. R. CHAUHAN

&

S. JANGRA, PhD



MERCURY LEARNING AND INFORMATION

Dulles, Virginia
Boston, Massachusetts
New Delhi

Copyright ©2020 by MERCURY LEARNING AND INFORMATION LLC. All rights reserved.
Reprinted and revised with permission.

Original title and copyright: *Computer Security and Encryption*.

Copyright ©2018 by University Science Press (An imprint of Laxmi Publications Pvt. Ltd. All rights reserved.)

This publication, portions of it, or any accompanying software may not be reproduced in any way, stored in a retrieval system of any type, or transmitted by any means, media, electronic display or mechanical display, including, but not limited to, photocopy, recording, Internet postings, or scanning, without prior permission in writing from the publisher.

Publisher: David Pallai
MERCURY LEARNING AND INFORMATION
22841 Quicksilver Drive
Dulles, VA 20166
info@merclearning.com
www.merclearning.com
1-800-232-0223

S. R. Chauhan & S. Jangra. *Computer Security and Encryption: An Introduction*.
ISBN: 978-1-68392-531-6

The publisher recognizes and respects all marks used by companies, manufacturers, and developers as a means to distinguish their products. All brand names and product names mentioned in this book are trademarks or service marks of their respective companies. Any omission or misuse (of any kind) of service marks or trademarks, etc. is not an attempt to infringe on the property of others.

Library of Congress Control Number: 2020939905
202122321 Printed on acid-free paper in the United States of America.

Our titles are available for adoption, license, or bulk purchase by institutions, corporations, etc. For additional information, please contact the Customer Service Dept. at 800-232-0223(toll free).

All of our titles are available in digital format at www.academiccOURSEware.com and other digital vendors. The sole obligation of MERCURY LEARNING AND INFORMATION to the purchaser is to replace the book, based on defective materials or faulty workmanship, but not based on the operation or functionality of the product.

CONTENTS

<i>Preface</i>	xix
Chapter 1: Security Concepts	1
1.1 Security Introduction	1
1.2 The Need for Security	2
1.3 Security Approaches	4
1.3.1 Security Models	4
1.3.2 Security Management Practices	4
1.4 Principles of Security	5
1.4.1 Confidentiality	6
1.4.2 Authentication	6
1.4.3 Integrity	7
1.4.4 Non-Repudiation	7
1.4.5 Access Control	8
1.4.6 Availability	8
1.5 Types of Attacks	9
1.5.1 Theoretical Concepts	9
1.5.2 The Practical Side of Attacks	12
1.5.3 Java Security	20
1.5.4 Specific Attacks	22
Exercises	24
Chapter 2: Public Key Cryptography and SSL	27
2.1 One-Way Functions Introduction	27
2.1.1 Motivation	28

2.2	One-Way Functions: Definitions	29
2.2.1	(Strong) One-Way Functions	29
2.3	Digital Signatures	30
2.4	Hash Functions	31
2.5	Centralized Certificates	32
2.6	Random Key Generation	34
2.7	Authentication Methods	35
2.8	Email Security	36
2.9	Challenge Handshake Authentication Protocol	37
2.10	Automatic Rekeying	38
2.11	Biometrics	39
2.12	Public Key Cryptography	43
2.13	Mutual Authentication	44
2.14	Multifactor Authentication	44
2.15	Elements of an Authentication System	45
2.16	Attacks	47
2.17	IP Security Encryption Router	50
2.18	Cryptography	53
2.19	Cryptosystems	53
2.20	Key-Based Methodology	54
2.21	Symmetric (Private) Methodology	54
2.22	Asymmetric (Public) Methodology	56
2.23	Key Distribution	59
2.24	Asymmetric Algorithms	61
2.25	Hash Functions vs. Key-Based Cryptosystems	62
	Exercises	63
Chapter 3:	World Wide Web Transaction Security	65
3.1	Internet Infrastructure	65
3.1.1	Internet	66
3.1.2	Internet Service Providers (ISPs)	66
3.1.3	Point of Presences (POPs)	66
3.1.4	Network Access Point (NAP)	66
3.1.5	Local Area Network (LAN)	67

3.2	Network Infrastructure	67
3.3	Basic Issues in Secret Key Management	68
3.3.1	Links	68
3.3.2	Routers	69
3.4	Addressing	70
3.5	System Security	71
3.6	Basic Issues in Internet Transaction Security	73
3.7	Network Information and Network Infrastructure Securities	74
3.8	Importance of Network Infrastructure Security	74
3.9	Internet Infrastructure Vulnerability	75
3.9.1	Solutions Usually Require Large Scale Modifications	75
3.9.2	Security and Performance Tradeoffs	76
3.9.3	Security is Only as Strong as the Weakest Link	76
3.9.4	Attacks Can Be Easily Launched and Are Difficult to Trace	76
3.10	Network Infrastructure Security—Switching	76
3.11	Switch Security is Important	78
3.12	How Switches Can Be Attacked	79
3.12.1	Mac Flooding	79
3.12.2	Content Addressable Memory Table	80
3.12.3	Mac Flooding Attacks	81
3.12.4	Mitigation	81
3.12.5	ARP Spoofing	83
3.13	ARP	83
3.14	The ARP Poisoning Process	85
3.15	Man-in-the-Middle Attack	86
3.15.1	DoS Attack	86
3.15.2	Hijacking	86
3.15.3	Spoofing WAN Traffic	86
3.16	Static ARP Entries	88
3.16.1	Detection	88
3.16.2	No Cache Update	88

3.17	STP Attacks	89
3.18	Topology Change (Bit 1)	89
3.18.1	Bridge ID	90
3.18.2	Port State	90
3.18.3	STP Timer	90
3.19	How STP Works	90
3.20	Topology Change	96
3.20.1	Failure to Receive the Hello Bpdus	96
3.21	STP Attack Scenarios	97
3.22	Root Claim and MITM	98
3.23	Affecting Network Performance	99
3.24	Countermeasures	100
3.24.1	BPDU Guard	100
3.25	Root Guard	101
3.26	VLAN Attacks	103
3.26.1	Easier Network Administration	103
3.26.2	Improved Bandwidth Usage	103
3.26.3	Blocking Broadcast Traffic	103
	Exercises	105
Chapter 4:	IP Security and Firewalls	107
4.1	Internet Firewalls	107
4.2	Protective Devices	109
4.2.1	Your Data	109
4.2.2	Resources	111
4.2.3	Reputation	111
4.3	Types of Attacks	113
4.3.1	Intrusion	113
4.3.2	Denial of Service	114
4.4	Network Taps	116
4.5	IP Security Firewall	118
4.6	Joy Riders	119

4.7	Vandals	120
4.8	Scorekeeper	120
4.9	Spies: Industrial and Otherwise	121
4.10	Irresponsible Mistakes and Accidents	122
4.11	Theoretical Attacks	123
4.12	Who Do You Trust?	124
4.12.1	No Security	125
4.13	Security Through Obscurity	125
4.14	Host Security	127
4.15	Network Security Model	128
4.15.1	No Security Model Can Do It All	129
4.15.2	Internet Firewalls	130
4.16	A Firewall Can Log Internet Activity Efficiently	133
4.17	A Firewall Limits Your Exposure	134
4.18	A Firewall Can't Protect Against Malicious Insiders	134
4.19	A Firewall Can't Protect Connections That Don't Go Through It	135
4.20	A Firewall Can't Protect Against New Threats	135
4.21	A Firewall Can't Fully Protect Against Viruses	135
4.22	A Firewall Can't Set Itself Up Correctly	137
4.22.1	What's Wrong with Firewalls?	138
4.23	Firewalls Interfere with the Internet	138
4.24	Firewalls Don't Deal with the Real Problem	139
4.24.1	Philosophical Arguments	140
4.25	Buying Versus Building	140
	Exercises	142
Chapter 5:	Public Key Certificates	143
5.1	Security Objectives	143
5.1.1	Security Issues when Connecting to the Internet	144
5.1.2	Protecting Confidential Information	145

5.2	Protecting Your Network: Maintaining Internal Network System Integrity	150
5.2.1	Network Packet Sniffers	151
5.2.2	IP Spoofing	152
5.2.3	Password Attacks	152
5.2.4	Denial-of-Service Attacks	152
5.2.5	Application Layer Attacks	153
5.3	Trusted, Untrusted, and Unknown Networks	154
5.3.1	Trusted Networks	154
5.3.2	Untrusted Networks	155
5.3.3	Unknown Networks	155
5.4	Establishing a Security Perimeter	155
5.5	Perimeter Networks	156
5.6	Developing Your Security Design	158
5.6.1	Know Your Enemy	158
5.6.2	Count the Cost	158
5.6.3	Identify Any Assumptions	158
5.6.4	Control Your Secrets	159
5.6.5	Human Factors	159
5.6.6	Know Your Weaknesses	159
5.6.7	Limit the Scope of Access	160
5.6.8	Understand Your Environment	160
5.6.9	Limit Your Trust	160
5.6.10	Remember Physical Security	160
5.6.11	Make Security Pervasive	160
5.7	Secure Sockets Layer	161
5.8	Email Security	163
5.9	Secure Email Protocols	164
5.9.1	Pretty Good Privacy (PGP)	165
5.9.2	Privacy-Enhanced Mail (PEM)	166
5.9.3	PGP Versus PEM	166
5.9.4	Secure MIME (S/MIME)	166
5.10	Web-Based Email Services	167

5.11	Certification Authority Hierarchies	168
5.12	Key Recovery and Escrowed Encryption	170
5.12.1	Key Recovery Methodologies	171
5.12.2	Key Recovery Entry	171
5.12.3	Key Escrow	172
5.13	Strong and Weak Cryptography	173
5.14	Security Alternatives for Web Forms	175
5.14.1	Web Security Considerations	175
5.15	Web Traffic Security Approaches	178
	Exercises	179
Chapter 6:	Security at the IP Layer	181
6.1	Cryptography	181
6.2	Stream Ciphers	183
6.3	Block Ciphers	184
6.3.1	Breaking Ciphers	185
6.4	Known Plaintext Attack	185
6.4.1	Chosen Plaintext Attack	186
6.5	Cryptanalysis	186
6.6	Brute Force	187
6.6.1	Social Engineering	187
6.6.2	Other Types of Attacks	187
6.7	Encryption	187
6.8	Symmetric Key Encryption	188
6.9	Data Encryption Standard (DES)	189
6.9.1	International Data Encryption Algorithm (IDEA)	190
6.9.2	CAST	190
6.9.3	Rivest Cipher #4 (RC4)	190
6.10	Asymmetric Key Encryption	190
6.11	Public Key Cryptosystems	192
6.11.1	Diffie-Hellman	192
6.11.2	Message Integrity	193

6.12	Secure Hash Algorithm-1 (SHA-1)	195
6.12.1	Authentication	196
6.13	Public Key Infrastructure	196
6.14	Secret Key Exchange	197
6.15	Web Security	201
6.15.1	Threats	202
6.15.2	Secure Naming	202
6.16	DNS Spoofing	203
6.16.1	Secure DNS	205
6.16.2	Self-Certifying Names	208
6.17	The Secure Sockets Layer	210
6.18	RSA Algorithm	214
	Exercises	219
Chapter 7:	Remote Access with Internet Protocol Security	221
7.1	Wireless Technologies	221
7.1.1	Types of Wireless Technology	222
7.2	Base Station	222
7.3	Technology of Offline Message Keys	223
7.4	Advanced Signaling Techniques Used to Mitigate Multipath	225
7.4.1	QAM with DFE	225
7.4.2	Spread Spectrum	226
7.4.3	FHSS	227
7.4.4	FDM	227
7.4.5	OFDM	228
7.4.6	VOFDM	229
7.5	Benefits of Using Wireless Solutions	230
7.6	Earth Curvature Calculation for Line-of-Sight Systems	230
7.7	Microwave Communication Links	232
7.7.1	What is Multipath?	232
7.7.2	Multipath in Non-LOS Environments	234

7.8	Elements of a Total Network Solution	235
7.8.1	Premises Networks	235
7.8.2	Access Networks	236
7.8.3	Core Networks	236
7.8.4	Network Management	236
7.8.5	Deployment	237
7.9	Billing and Management of Wireless Systems	238
7.9.1	Example Implementation	239
7.10	IP Wireless System Advantages	239
7.11	IP Wireless Services for Small and Medium Businesses	241
7.12	IP Point-to-Multipoint Architecture	242
7.13	IP Wireless Open Standards	245
7.14	IP Vector Orthogonal Frequency-Division Multiplexing	246
7.14.1	Channel Data Rate	247
7.14.2	Downstream and Upstream User Bandwidth Allocation	247
7.14.3	Duplexing Techniques	248
7.15	Multiple Access Technique	248
7.15.1	Unsolicited Grant Service	249
7.15.2	Real-Time Polling Service	249
7.15.3	Unsolicited Grant Service with Activation Detection	250
7.15.4	Non-Real-Time Polling Service	250
7.15.5	Best Effort Service	250
7.15.6	Committed Information Rate	250
7.15.7	Frame and Slot Format	251
7.16	Synchronization Technique (Frame and Slot)	251
7.17	Average Overall Delay Over Link	252
7.18	Power Control	252
7.19	Admission Control	253
7.20	Requirements for the Cell Radius	253

7.20.1	Requirement for Frequency Reuse	254
7.20.2	Radio Resource Management	255
7.20.3	Spectrum Management in a Cell	255
7.20.4	Load Balancing of CPES Within an Upstream Channel	255
7.20.5	Time-Slotted Upstream	255
7.21	Contention Resolution	256
7.21.1	Traffic Policing	256
7.22	Interface Specifications Based on the Generic Reference Model	257
7.23	Wireless Protocol Stack	258
7.24	System Performance Metrics	260
7.25	Supercell Network Design	261
7.26	Transport Layer Products	262
7.26.1	P2MP Transport Equipment Element—Customer Premises	263
7.26.2	Rooftop Unit	263
7.26.3	Basic Receiver	264
7.26.4	High-Gain Receiver	264
7.27	LMDS Environmental Considerations	264
7.28	WLAN Standards Comparison	265
	Exercises	266
Chapter 8:	Virtual Private Networks	267
8.1	Security Policy	267
8.2	IPSec Network Security	269
8.3	IPSec Protocols	272
8.3.1	Authentication Header (AH)	272
8.3.2	Encapsulated Security Payload (ESP)	273
8.3.3	IKE Protocol	274
8.4	NAT-Traversal	276
8.5	Virtual Private Network (VPN)	277
8.6	Gateway-to-Gateway Architecture	279

8.7	Host-to-Gateway Architecture	280
8.8	Model Comparison	282
8.9	TCP/IP Network Security Protocol	283
8.10	Node-to-Node Encryption	286
8.11	Site-to-Site Encryption	287
8.12	Where to Encrypt	288
8.13	Encryption Process	289
8.14	ESP Packet Fields	289
8.15	How ESP Works	291
8.16	ESP Version 3	292
8.17	Internet Key Exchange (IKE)	293
8.18	Phase One Exchange	293
8.19	Main Mode	293
8.20	Diffie-Hellman (DH) Group	295
8.21	Aggressive Mode	298
8.22	Phase Two Exchange	299
8.23	Informational Exchange	301
8.24	Group Exchange	302
8.25	IKE Version 2	302
8.26	IP Payload Compression Protocol (IPComp)	303
8.27	ESP in a Gateway-to-Gateway Architecture	304
8.28	ESP and IPComp in a Host-to-Gateway Architecture	306
8.29	ESP and AH in a Host-to-Host Architecture	307
	Exercises	309
Chapter 9:	The Security of Emerging Technologies	311
9.1	Security of Big Data Analytics	312
9.1.1	Big Data Analysis Can Transform Security Analytics	313
9.1.2	Big Data Analytics for Security Issues and Privacy Challenges	313
9.2	Security of Cloud Computing	314
9.2.1	Cloud Deployment Models	315

9.2.2	The Three Layers of the Cloud Computing Services Model (Software, Platform, or Infrastructure (SPI) Model)	316
9.2.3	Security Concerns and Challenges of Cloud Computing	317
9.2.4	Cloud Security as a Consumer Service	317
9.3	Security of the Internet of Things (IoT)	318
9.3.1	Evolution of the IoT	318
9.3.2	Building Blocks of the Internet of Things (IoT)	319
9.3.4	IoT Layer Models	320
9.3.5	Applications of the IoT	321
9.3.6	New Challenges Created by the IoT	323
9.3.7	Security Requirements of the IoT	323
9.3.8	IoT Attacks	324
9.3.9	Hybrid Encryption Technique	325
9.3.10	Hybrid Encryption Algorithm Based on DES and DSA	326
9.3.11	Advance Encryption Standard (AES)	327
9.3.12	Requirements for Lightweight Cryptography	328
9.3.13	Lightweight Cryptography in the IoT	328
9.3.14	Prevention of Attacks on the IoT	329
9.4	Security of the Smart Grid	329
9.4.1	Smart Grid Challenges	330
9.4.2	Smart Grid Layers	330
9.4.3	Information Security Risks and Demands of Smart Grids	330
9.4.4	Smart Grid Security Objectives	332
9.4.5	The Smart Grid System: Three Major Systems	332
9.4.6	Types of Security Attacks that can Compromise the Smart Grid Security	332
9.4.7	Cybersecurity Attacks on a Smart Grid	333

9.5	Security of SCADA Control Systems	333
9.5.1	Components of SCADA Systems	334
9.5.2	SCADA System Layers	334
9.5.3	Requirements and Features for the Security of Control Systems	335
9.5.4	Categories of Security Threats to Modern SCADA Systems	336
9.6	Security of Wireless Sensor Networks (WSNs)	336
9.6.1	WSN Layers	337
9.6.2	Security Requirements in WSNs	337
9.6.3	WSN Attack Categories	338
9.6.4	Security Protocols in WSNs	339
	Exercises	342
	Index	345

PREFACE

Cryptography and system security may be the fastest growing technologies in our culture today because of the rapid growth of cybercrime. This book describes various aspects of cryptography and system security, with a particular emphasis on the use of rigorous security models and practices in the design. The first portion of the book presents the overall system security and provides a general overview of the features such as object models and inter-object communications. The objective of this portion is to provide an understanding of the cryptography underpinnings on which the rest of the book is based. The whole text has been divided into nine chapters:

Chapter 1. This chapter attempts to provide answers to the basic questions, the principles of any security mechanism, for the security and security models, Denial of Service (DoS) and type of active attacks, Virus, Worms, Trojan Horse, Java Applets, Java Security, and Applet and Active X Controls.

Chapter 2. In this chapter, we provide One-Way Functions, Digital Signature, Authentication Method, Hash Function, Digital Certificates, Challenge Handshake Authentication Protocol, Biometrics and Mutual Authentication.

Chapter 3. This chapter provides discussions of Internet Service Providers (ISP), the Network Access Point (NAP), Routers, Addressing, ATM, Ethernet, Fiber Distributed Data Interface (FDDI), Multi-Protocol Label Switching (MPLS), Point-to-Point Protocols (PPP) and High-level Data Link Control (HDLC).

Chapter 4. This chapter covers Firewalls and their purpose. In this chapter we present Protective Device, Denial of Service, Spies (Industrial and Otherwise), Network Taps, Host Security, How A Firewall Can Log Internet Activity Efficiently, Buying Versus Building, and Why A Firewall Can't Fully Protect Against Viruses.

Chapter 5. This chapter provides answers to understanding the types of attacks that may be used by hackers to undermine network security; understand the types of vulnerabilities that may be present in your network; learning to classify the different types of networks and users that may interact with your own; and evaluate their risk factors; learn to evaluate your network topology and requirements; developing a suitable security policy for implementation; and becoming familiar with the tools available for protecting confidential information and your network.

Chapter 6. In this chapter, exchange of information and commerce to be secured on any network; securing information on a network; plaintext to produce a stream of secrets, block ciphers, plaintext attack, Public Key Cryptosystems, Symmetric Key Encryption, Data Encryption Standard (DBS), and Secret Key Exchange have been discussed.

Chapter 7. This chapter provides information and commerce; identify different types of wireless technologies; identify different wireless solutions; introduce quadrature amplitude modulation; explain wireless systems, and discuss the benefits of using wireless technologies for communications.

Chapter 8. IPsec is a framework of open standards developed by the Internet Engineering Task Force (IETF). IPsec provides security for transmission of sensitive information over unprotected networks – such as the Internet comparison of IPsec to Cisco Encryption Technology, IPsec protocols, IKE Protocol, Internet Engineering Task Force (IETF), NAT-Traversal, Gateway-to-Gateway Architecture, Virtual Private Networking (VPN), Encryption Processes, End-To-End Encryption, and ESP v.3.

Chapter 9. This chapter discusses the security protocols and processes involved with emerging technologies. The chapter covers topics such as Big Data Analytics, Cloud Computing, Internet of Things (IoT), Smart Grid, Supervisory control and data acquisition (SCADA) Control Systems, Wireless Sensor Network (WSN).

SECURITY CONCEPTS

Chapter Goals

- The principles of any security mechanism
- The need for the security
- Security models
- Denial-of-Service (DoS) and other types of active attacks
- Viruses, Worms, Trojan horses, and Java Applets
- Java Security
- Applet and ActiveX controls

1.1 SECURITY INTRODUCTION

This is a book on network and Internet security. As such, before we embark on our journey of understanding the various concepts and technical issues related to security, it is essential to know what we are trying to protect. What are the dangers of using computers, computer networks, and the biggest network of them all, the Internet? What are the likely pitfalls? What happens if we do not implement the right security policies, frameworks, and technology? This chapter attempts to provide answers to these basic questions.

We start with a discussion of the fundamental point: Why is security required in the first place? People sometimes say that security is like statistics: what it reveals is trivial, what it conceals is vital. In other words, the right security infrastructure opens up just enough doors. We will discuss a few real-life incidents that should prove beyond any doubt that security is important. Now

that critical business and other types of transactions are being conducted over the Internet to such a large extent, inadequate or improper security mechanisms can destroy a business or play havoc with people's lives.

We will also discuss the key principles of security. These principles will help us identify the various areas that are crucial while determining the security threats and possible solutions. Electronic documents and messages are now considered the equivalent to paper documents in terms of their legal validity, and we will examine the implications of this new view of information.

1.2 THE NEED FOR SECURITY

Most of the first computer applications had no or at best, very little, security. This lack of security continued for a number of years until the importance of data was truly realized. Until then, computer data was considered useful, but not something that needed to be protected. When computer applications were developed to handle financial and personal data, the real need for security was felt like never before. People realized that the data on computers is an extremely important aspect of modern life, and various areas in security began to gain importance. Two typical examples of security mechanisms are as follows:

- Provide a user id and password to every user, and use that information to authenticate a user
- Encode information stored in the databases in some fashion so that it is not visible to users who do not have the right permissions.

Organizations employed their own mechanisms to provide security. As technology improved, the communication infrastructure became extremely mature, and newer applications were developed to meet various user demands and needs. Soon, people realized that the basic security measures were not enough.

The Internet is used globally, and there were many examples of what could happen if there was insufficient security built into the applications developed for the Internet. Figure 1.1 shows such an example of what can happen when you use your credit card for making purchases over the Internet. From the user's computer, the user's details, such as the user id, order details, such as the order id and item id, and payment details, such as the credit card information, travel across the Internet to the merchant's server. The merchant's server stores these details in its database.

There are various security holes in this process. First, an intruder can capture the credit card details as they travel from the client to the server. If we somehow protect this transit from an intruder's attack, it still does not

solve our problem. Once the merchant receives the credit card details and validates them to process the order and obtain payments, the merchant stores the credit card details in its database. An attacker can simply access this database and gain access to all the credit card numbers stored therein! One Russian attacker (called Maxim) managed to hack a merchant's Internet site and obtain 300,000 credit card numbers from its database. He then attempted extortion by demanding protection money (\$100,000) from the merchant. The merchant refused to oblige. Following this, the attacker published about 25,000 of the credit card numbers on the Internet. Some banks reissued all the credit cards at a cost of \$20 per card, and others warned their customers about unusual entries in their statements.

Such attacks can obviously lead to great losses, both in terms of finances and goodwill. Generally, it takes about \$20 to replace a credit card. Therefore, if a bank has to replace 300,000 such cards, the total cost of such an attack is about \$6 million. Had the merchant in the example employed proper security measures, he would have saved money and bother.

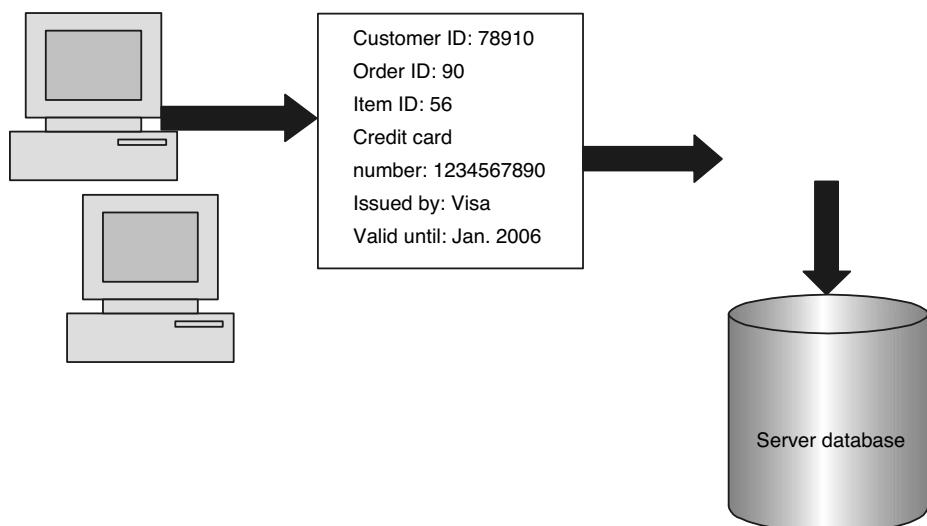


FIGURE 1.1 Example of information traveling from a client to a server over the Internet

Of course, this is just one example. More cases have been reported, and the need for proper security is increasing with every attack. In another example, in 1999, a Swedish hacker broke into Microsoft's Hotmail Website and created a mirror site. This site allowed anyone to enter any Hotmail user's email id and read the user's emails.

Also in 1999, two independent surveys were conducted to invite people's opinions about the losses that occur due to successful attacks on security. One survey pegged the losses at an average of \$256,296 per incident, and another found the average was \$759,380 per incident. In the following year, this figure rose to \$972,857.

1.3 SECURITY APPROACHES

The last twenty years have witnessed a major development in a formal methods used to improve security protocols. The design of such protocols has so far been largely an empirical and ad-hoc procedure, giving rise to various approaches. These methods are now being systematically applied to develop a system that is both efficient and can comply with strong security requirements.

1.3.1 Security Models

An organization can take several approaches to implement its security model:

- **No security:** In this simplest case, the approach could be a decision to implement no security at all.
- **Security through obscurity:** In this model, a system is secure simply because nobody knows about its existence and content.
- **Host security:** In this scheme, the security for each host is enforced individually. This is a very safe approach, but the trouble is that it cannot scale well. The complexity and the diversity of modern sites/organizations make the task even harder.
- **Network security:** Host security is difficult to achieve as organizations grow and become more diverse. In this technique, the focus is to control network access to various hosts and their services, rather than the individual host's security. This is an efficient and scalable model.

1.3.2 Security Management Practices

Good security management practices always include a security policy. Putting a security policy in place is actually quite difficult. A good security policy and its proper implementation go a long way in ensuring adequate security management practices. A good security policy generally takes care of four key aspects.

- **Affordability:** How much money and effort does this security implementation cost?
- **Functionality:** What is the mechanism of providing security?
- **Cultural Issues:** Does the policy take into consideration people's expectations, working style, and beliefs?
- **Legality:** Does the policy meet the legal requirements?

Once a security policy is in place, the following points should be ensured:

- a. Include an explanation of the policy to all concerned.
- b. Outline everybody's responsibilities.
- c. Use simple language in all communications.
- d. Accountability should be established.
- e. Provide for exceptions and periodic reviews.

1.4 PRINCIPLES OF SECURITY

Having discussed some of the attacks that have occurred in real life, let us now classify the principles related to security. This will help us understand the attacks better and think about the possible solutions.

Let us assume that a person, A, wants to send a check worth \$100 to another person, B. Normally, what are the factors that A and B think of in such a case? A will write the check for \$100, put it inside an envelope, and send it to B.

- A wants to ensure that no one expects B will get the envelope, and even if someone else gets it, she does not want anyone to know about the details of the check. This is the principle of *confidentiality*.
- A and B would like to make sure that no one can tamper with the contents of the check (such as its amount, date, signature, or name of the payee). This is the principle of *integrity*.
- B would like to be assured that the check has indeed come from A, and not from someone else posing as A (as it could be a fake check). This is the principle of *authentication*.
- What will happen tomorrow if B deposits the check into her account, the money is transferred from A's account, and then A claims to have not written/sent the check? The court of law will use A's signature to disallow A to refute this claim and settle the dispute. This is the principle of *non-repudiation*.

These are the four chief principles of security. There are two more, *access control* and *availability*, which are not related to the particular message, but are linked to the overall system as a whole.

1.4.1 Confidentiality

The principle of *confidentiality* specifies that only the sender and the intended recipient(s) should be able to access the content of a message. Confidentiality gets compromised if an unauthorized person is unable to access a message. An example of compromising the confidentiality of a message is shown in Figure 1.2. The user of a computer A sends the message to the user of a computer B. (From here onwards, we use “A” to mean the user A, and “B” to mean user B, although we just show the computers of the users A and B). Another user, C, gets access to this message, which is not desired, and therefore, defeats the purpose of confidentiality. An example of this could be a confidential email message sent by A and B. This type of attack is called *interception*.

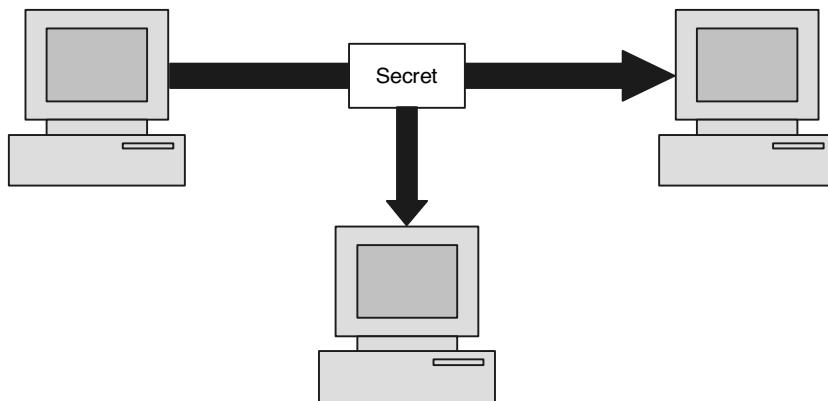


FIGURE 1.2 The loss of confidentiality: Interception causes the loss of the message’s confidentiality.

1.4.2 Authentication

Authentication mechanisms help establish proof of identities. The authentication process ensures that the origin of an electronic message or document is correctly identified. For instance, suppose that user C sends an electronic document over the Internet to user B. However, the trouble is that user C is posing as user A when she sent this document to user B. However, would user B know that the message has come from user C, who is posing as user A? A real-life example of this would be the case of a user C, posing as user A, sending a funds transfer request (from A's account to C's account) to bank B.

The bank will happily transfer the funds from A's account to C's account, and it would think that user A has requested the funds transfer. This concept is shown in Figure 1.3. This type of attack is called *fabrication*.

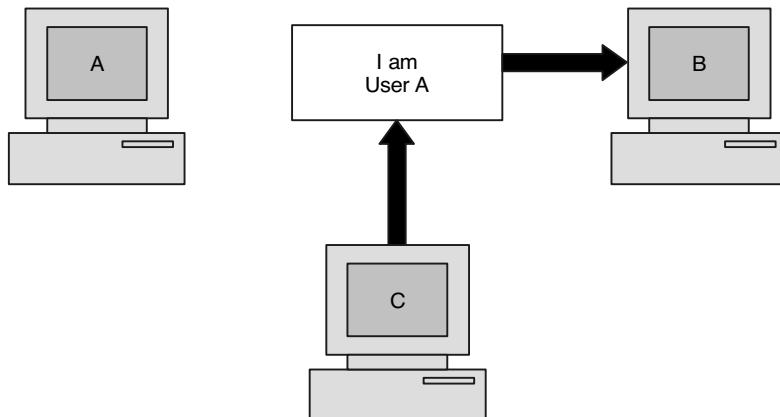


FIGURE 1.3 The absence of authentication

1.4.3 Integrity

When the contents of a message are changed after the sender sends it, but before it reaches the intended recipient, we say that the integrity of the message is lost. For example, suppose you write a cheque for \$100 to pay for some goods. However, when you see your next account statement, you are startled to see that the cheque resulted in a payment of \$1,000. This type of case is about the loss of the message's integrity. Conceptually, this is shown in Figure 1.4. User C tampers with a message originally sent by user A, which was actually destined for user B. User C somehow manages to access it, change its contents, and send the changed message to user B. User B has no way of knowing that the contents of the message were changed after user A sent it. User A also does not know about this change. This type of attack is called *modification*.

1.4.4 Non-Repudiation

There are situations where a user sends a message, and then says that she never sent that message. For instance, user A could send a funds transfer request to bank B over the Internet. After the bank performs the funds transfer as per A's instructions, A could claim that she never sent the funds transfer instruction to the bank. Thus, A repudiates, or denies, her funds transfer

instruction. The principle of *non-repudiation* defeats the possibility of denying something after having done it.

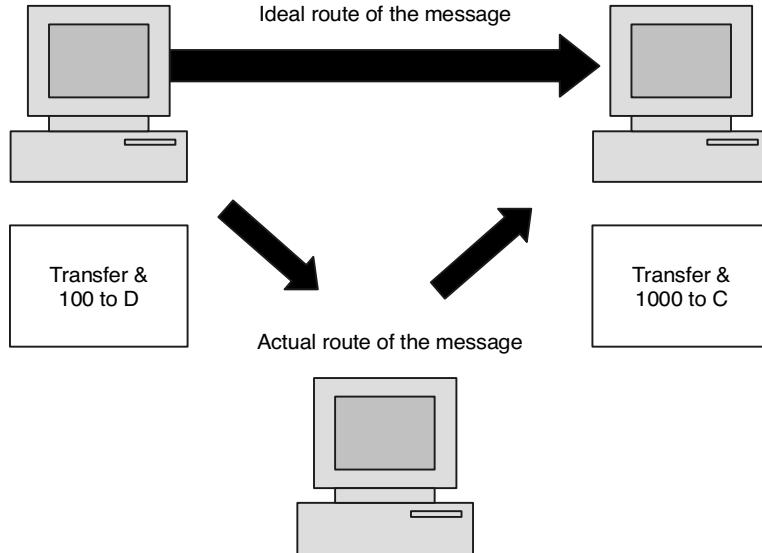


FIGURE 1.4 Loss of integrity

1.4.5 Access Control

The principle of *access control* determines *who* should be able to access what. For instance, we should be able to specify that user A can view the records in a database, but cannot update them. However, user B might be allowed to make updates, as well. An access control is broadly related to two areas: *role management* and *rule management*. Role management concentrates on the user side (which user can do what), whereas rule management focuses on the decisions taken, and so an access control matrix is prepared. A list of items is generated, including what they can access (*e.g.*, it can say that user A can write to file X, but can only update files Y and Z). An *Access Control List (ACL)* is a subset of an access control matrix.

1.4.6 Availability

The principle of *availability* states that resources should be available to authorized parties at all times. For example, due to the intentional actions of another unauthorized user C, the authorized user A may not be able to

contact a server computer B, as shown in Figure 1.5. This would defeat the principle of availability. Such an attack is called *interruption*.

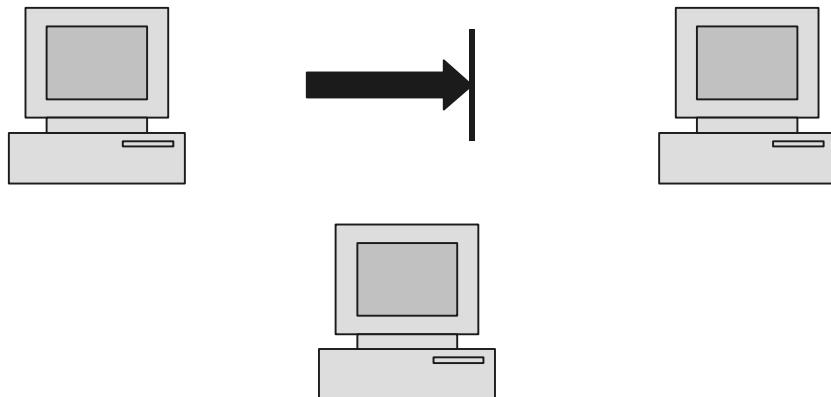


FIGURE 1.5 Attack on availability

Having discussed the various principles of security, let us now discuss the different types of attacks that are possible from a technical perspective.

1.5 TYPES OF ATTACKS

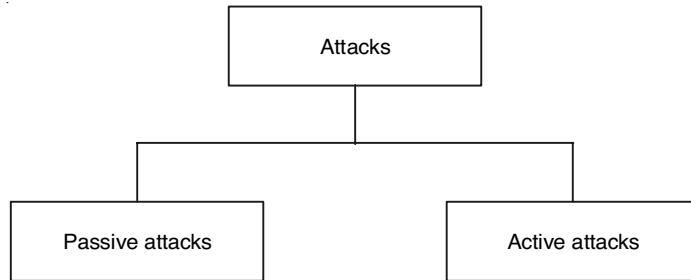
We can classify the types of attacks on computers and network systems into two categories for a better understanding: (a) the theoretical concepts behind these attacks and (b) practical approaches used by the attackers.

1.5.1 Theoretical Concepts

The principle of security faces threats from various attacks. These attacks are generally classified into four categories. They are

- **Interception:** Discussed in the context of *confidentiality* earlier.
- **Fabrication:** Discussed in the context of *authentication* earlier.
- **Modification:** Discussed in the context of *integrity* earlier.
- **Interruption:** Discussed in the context of *availability* earlier.

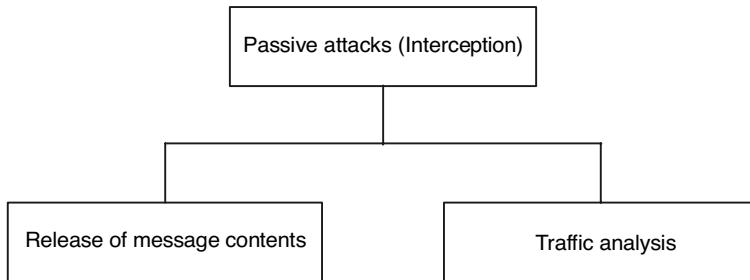
These attacks are further grouped into two types: passive attacks and active attacks, as shown in Figure 1.6.

**FIGURE 1.6** Types of attacks

1.5.1.1 Passive Attacks

Passive attacks are those wherein the attacker indulges in eavesdropping or the monitoring of data transmissions. The attacker attempts to obtain information that is in transit. The term passive indicates that the attacker does not attempt to perform any modifications to the data. In fact, this is also why passive attacks are harder to detect, thus, the general approach to deal with passive attacks is to think about prevention, rather than detection or corrective actions.

Figure 1.7 shows a further classification of passive attacks into two sub-categories. These categories are the release of the message contents and traffic analysis.

**FIGURE 1.7** Passive attacks

The *release of message contents* is quite simple to understand. When we send a confidential email message to our friend, we only want her to access it. Otherwise, the contents of the message are released against our wishes to someone else. Using certain security mechanisms, we can prevent the release of the message contents. For example, we can encode messages using a code language, so that only the desired parties understand the contents of a message because only they know the code language. However, if many such messages are passing through, a passive attacker could try to figure out the

similarities between them to come up with a pattern that provides her some clues regarding the communication that is taking place. Such attempts at analyzing (encoded) messages to come up with likely patterns are the work of *traffic analysis* attacks.

1.5.1.2 Active Attacks

Unlike passive attacks, *active attacks* are based on the modification of the original messages in some manner or on the creation of a false message. These attacks cannot be prevented easily. However, they can be detected with some effort, and attempts can be made to recover from them. These attacks can be in the form of interruption, modification, and fabrication.

- Interruption attacks are called masquerade attacks.
- Modification attacks can be classified further into the replay attacks and alteration of messages.
- Fabrication causes Denial of Service (DoS) attacks.

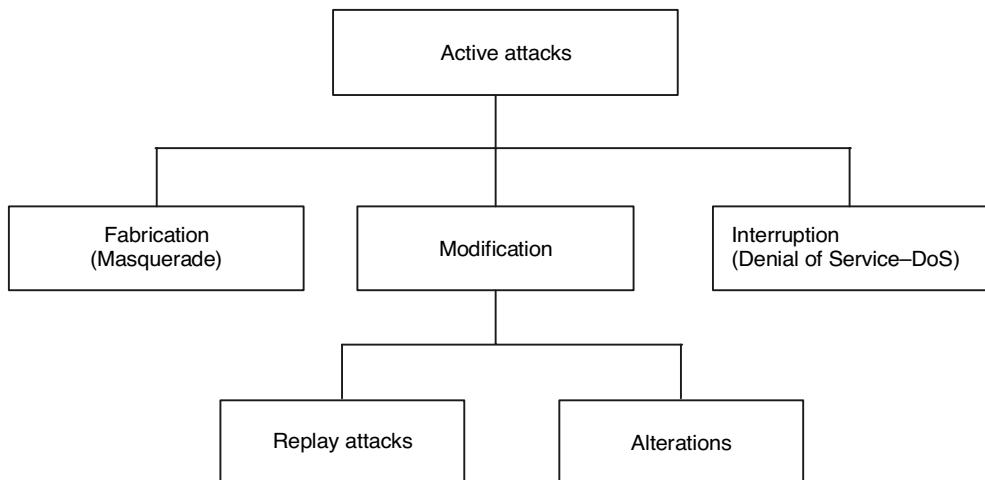


FIGURE 1.8 Active attacks

A *masquerade* is caused when an unauthorized entity pretends to be another entity. User C might pose as user A and send a message to user B. User B might be led to believe that the message indeed came from user A.

In a replay attack, a user captures a sequence of events, or some data units, and resends them. For instance, suppose user A wants to transfer some amount to user C's bank account. Both users A and C have accounts with bank B. User A might send an electronic message to bank B, requesting the

funds transfer. User C could capture this message and send a second copy of the same to bank B. Bank B would have no idea that it is an unauthorized message and would treat this as a second, and different, funds transfer request from user A. Therefore, user C would get the benefit of the funds transfer twice: once actually authorized and once through a replay attack.

The alteration of messages involves some change to the original message. For instance, suppose user A sends an electronic message to transfer \$10,000 to C's account. The beneficiary captures this and changes it to transfer \$100,000 to B's account. Note that both the beneficiary and the amount have been changed: only one of these could have caused the alteration of the message.

Denial-of-Service (DoS) attacks make an attempt to prevent legitimate users from accessing some services that they are eligible for. For instance, an unauthorized user might send too many emails so as to flood the network and deny other legitimate users access to the network.

1.5.2 The Practical Side of Attacks

The attacks discussed earlier can come in a number of forms in real life. They can be classified into two broad categories: application-level attacks and network-level attacks, as shown in Figure 1.9.

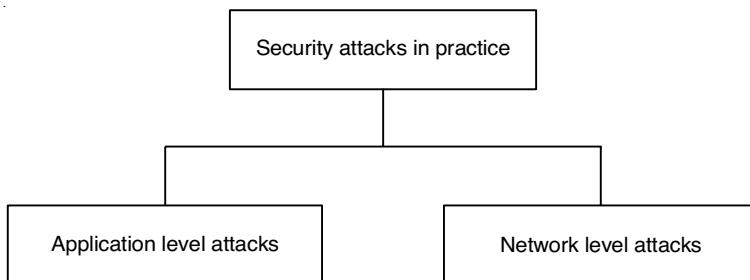


FIGURE 1.9 Practical side of attacks

- *Application level attacks:* These attacks happen at an application level in the sense that the attacker attempts to access, modify, or prevent access to information of a particular application or the application itself. Examples of this are trying to obtain someone's credit information on the Internet or changing the contents of message to change the amount in a transaction.
- *Network level attacks:* These attacks are generally aimed at reducing the capabilities of network. These attacks make an attempt to either slow

down, or completely bring to halt, a computer network. Note that this automatically can lead to application level attacks, because once someone is able to gain access to a network, she is able to access/modify at least some sensitive information, causing havoc.

These two types of attacks can be attempted by using varies mechanisms. These attacks are not encompassed in the above two categories, since they can span across application as well as network levels.

1.5.2.1 Virus

One can launch an application-level attack or a network level attack using a virus. A *virus* is a piece of program code that attaches itself to legitimate program code and runs when the legitimate program runs.

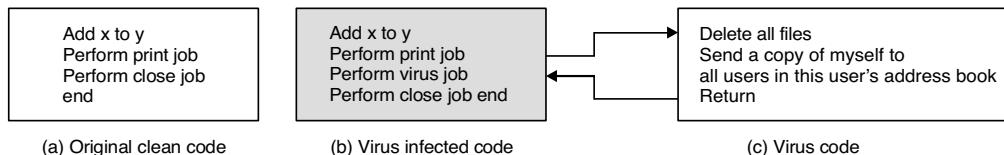


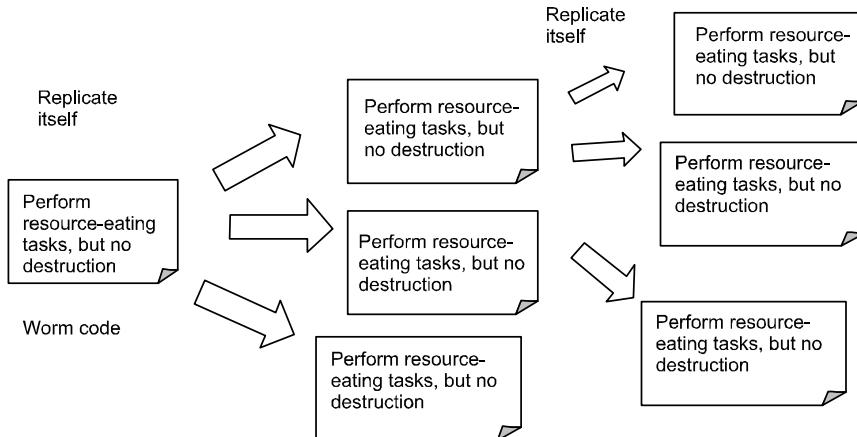
FIGURE 1.10 Virus

It can then infect other programs in that computer or programs that are on other computers but on the same network. This is shown in Figure 1.10. After deleting all the files from the current user's computer, the virus self-propagates by sending its code to all users whose e-mail addresses are stored in the current user's address book.

Viruses can also be triggered by specific events (*e.g.*, a virus could automatically execute at 12 p.m. every day). Viruses cause damage to computers and network systems, but this damage can be repaired, assuming that the organization deploys good backup and recovery producers.

1.5.2.2 Worms

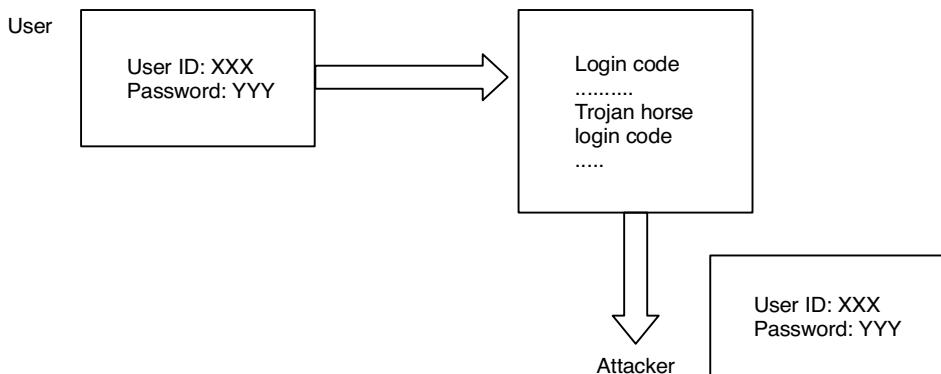
Similar in concept to a virus, a worm is actually different in implementation. A virus modifies a program (*i.e.*, it attaches itself to the program under attack). This is shown in Figure 1.11. The replication grows so much that ultimately the computer or the network, on which the worm resides, become very slow, finally coming to a halt. Thus, the basic purpose of a worm attack is different from that of a virus. A *worm* attempts to make the computer or the network under attack unusable by consuming all its resources.

**FIGURE 1.11** Worm

A worm does not perform any destructive actions, but instead only consumes system resources to bring it down.

1.5.2.3 Trojan Horse

A Trojan horse is a hidden piece of code, like a virus. However, the purpose of a Trojan horse is different. The main purpose of a virus is to make modifications to the target computer or network, whereas a *Trojan horse* attempts to reveal confidential information to an attacker. The name (*Trojan horse*) is taken from the secret attack executed by Greek soldiers, who hid inside a large hollow horse that was pulled into the city of Troy by its citizens, unaware of its contents. Once the Greek soldiers entered the city of Troy, they opened the gates for the rest of their army.

**FIGURE 1.12** Trojan horse

In a similar fashion, a Trojan horse could silently sit in the code for a login screen by attaching itself to it. When the user enters the user ID and password, the Trojan horse captures these details and sends this information to the attacker without the knowledge of the user who entered the ID and password. The attacker can then use the user ID and password to gain access to the system.

1.5.2.4 Applets and ActiveX Controls

Applets and ActiveX controls were born through the technological development of the World Wide Web (WWW) applications. In its simplest form, the Web consists of the communication between client and server computers using a communications protocol called the Hyper Text Transfer Protocol (HTTP). The client uses software called a Web browser. The server runs a program called a Web server. In its simplest form, a browser sends an HTTP request for a Web page to a Web server. The Web server locates this Web page (actually a computer file) and sends it back to the browser, again using HTTP. The Web browser interprets the contents of that file and shows the results on the screen to the user. This is shown in Figure 1.13. Here, the client sends a request for the web page *www.yahoo.com/info*, which the server sends back to the client.

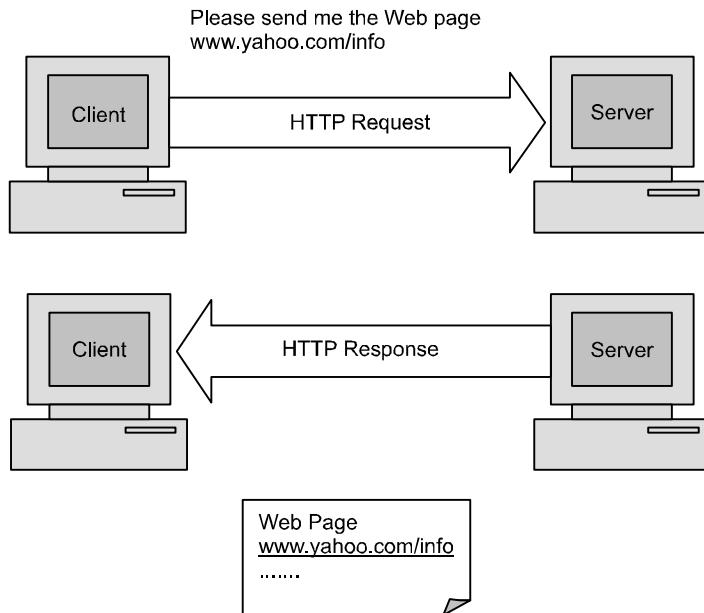


FIGURE 1.13 Example of an HTTP interaction between the client and server

Many Web pages contain small programs that get downloaded on to the client along with the Web page itself. These programs then execute inside the browser. Sun Microsystems created *Java applets* for this purpose, and Microsoft's technology makes use of *ActiveX controls* for the same purpose. Both are essentially small programs that get downloaded along with a Web page and then executed on the client. This is shown in Figure 1.14. The server sends an applet along with the Web page to the client.

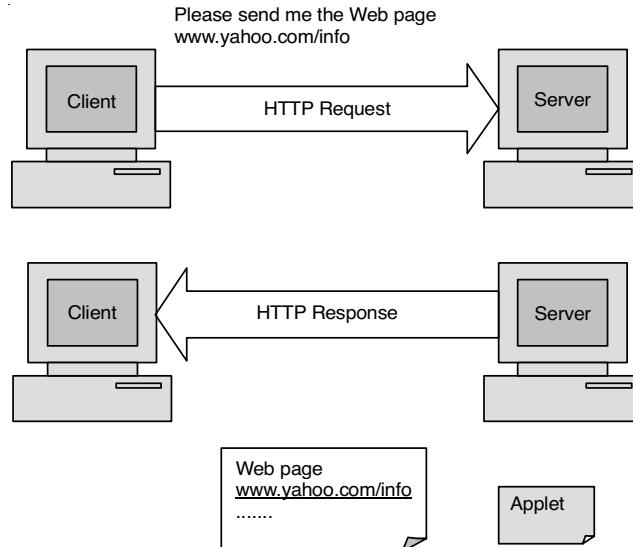


FIGURE 1.14 Applet sent back with a Web page

Usually, these programs (applets or ActiveX controls) are used to either perform some processing on the client side or to automatically and periodically request information from the Web server using a technology called *client pull*. For instance, a program can get downloaded on to the client along with the Web page showing the latest stock prices on a stock exchange, and then periodically issue HTTP requests for pulling the updated prices to the Web server. After obtaining this information, the program displays it on the user's screen.

These apparently innocuous programs can sometimes cause havoc. What if such a program performs a virus-like activity by deleting files on the user's hard disk, by stealing some personal information, or by sending junk e-mails to all the users whose addresses are contained in the user's address book?

To prevent these attacks, Java applets have strong security checks as to what they can and cannot do. ActiveX controls have no such restrictions. A new version of applets called *signed applets* allows accesses similar to those of ActiveX. Of course, a number of checks are in place to ensure that neither

applets nor ActiveX controls can do a lot of damage; even if they somehow manage to do it, the damage can be detected. However, at least in theory, they pose a security risk.

1.5.2.5 Cookies

Cookies are the result of a specific characteristic of the Internet. The Internet uses the HTTP protocol, which is *stateless*.

Suppose that the client sends an HTTP request for a Web page to the server. The Web server locates that page on its disk, sends it back to the client, and completely forgets about this interaction. If the client wants to continue this interaction, it must identify itself to the server in the next HTTP request. Otherwise, the server would not know that this same client had sent an HTTP request earlier. Since a typical application is likely to involve a number of interactions between the client and the server, there must be some mechanism for the client to identify itself to the server each time it sends a HTTP request to the server. For this, cookies are used. Cookies are perhaps the most popular mechanism of maintaining the state information. Actually, a Web server sends the Web browser a cookie and the browser stores it on the hard disk of the client computer. The browser then sends a copy of the cookie to the server during the next HTTP request. This is used for identification purposes, as shown in Figure 1.15(a) and 1.15(b).

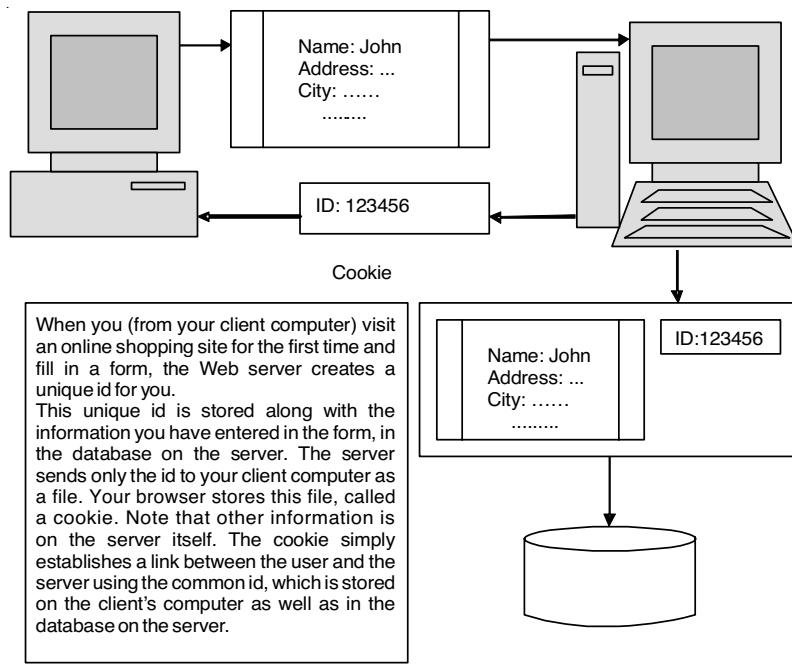
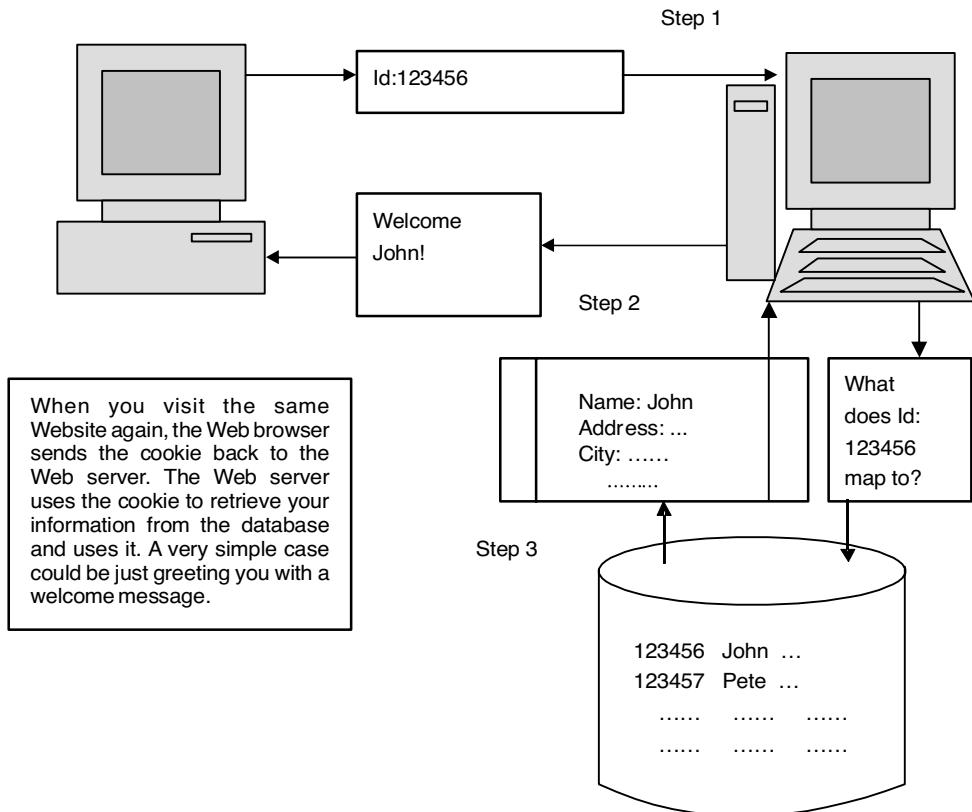


FIGURE 1.15(a) Creation of cookies (I)

**FIGURE 1.15(b)** Creation of cookies (II)

A *cookie* works as follows:

- When you interact with a Website for the first time, the site might want you to register yourself. Usually, this means that the Web server sends a page to you wherein you have a form to enter your name, address, and other details, such as date of birth and interests.
- When you complete this form and send it to the server with the help of your browser, the server stores this information in its database. Additionally, it also creates a unique ID for you. It stores this ID along with your information in the database and also sends the ID back to you in the form of a cookie.

- c. The next time you interact with the server, you do not have to enter any information, such as your name and address. Your browser automatically sends your ID along with the HTTP request for a particular page to the server.
- d. The server now takes this ID and tries to find a match in its database. When it finds it, it knows that you are a registered user. Accordingly, it sends you the next page, which could contain a simple welcome message. This can also be used for other purposes.

People perceive that cookies are dangerous. Actually, this is generally not true. Cookies can do little, if any, harm to you. First, the Web server that originally created a cookie can only access the cookie. Second, cookies can contain only text-based information. Third, the user can refuse to accept cookies.

1.5.2.6 Java Script, VBScript, and Jscript

A Web page is constructed using a special language called *Hyper Text Markup Language (HTML)*. It is a tag-based language. A tag begins with <> and it ends with </>. The boundaries of these tags contain the information for how things should be displayed on the user's computer. As an example, let us consider how the tag pair and can be used to change the font to boldface.

When a browser comes across this portion of the HTML document, it realizes that the portion of the text embedded within the and tags needs to be displayed in boldface. Therefore, it displays this text in boldface.

In addition to HTML tags, a Web page can contain *client-side scripts*. These are small programs written in scripting languages like Java Script, VBScript, or JScript, that are executed inside the Web browser on the client computer. For instance, let us assume that a user visits the Website of an online bookshop. Suppose that the Website mandates that the user must place an order for at least three books. Then the Web page uses a small JavaScript program that ensures that this condition is met before the user can place the order. Otherwise, the Java script program does not allow the user to proceed. Note that HTML cannot be used for this purpose, as its sole purpose is to display text on the client computer in a pre-specified format. To perform dynamic actions, such as the one discussed here, we need scripts.

Scripts can be dangerous. Since scripts are small programs, they can perform actions on the client's computer. Of course, there are restrictions as to what a scripting program can and cannot do. However, security breaches related to scripts have been reported.

1.5.3 Java Security

In this section we will discuss, Java security issues.

1.5.3.1 Introduction

For Java to become successful, it needed to avoid the security problems that had plagued other models of software distribution. Therefore, the early design of Java focused mainly on these concerns. Consequently, Java programs are considered safe as they cannot install, execute, or propagate viruses, and because the programs cannot perform any action that is harmful to the user's computer.

One of the key attributes of Java is the ability to download Java programs over a network and execute these programs on a different computer within the context of a Java-enabled browser. Developers were attracted to Java with different expectations. As a result, they had different ideas about Java security. Simply put, if we expect Java to be free from introducing viruses, any release of Java should satisfy our requirements. However, if we require special functionalities such as a digital signatures, authentication, and encryption in our program, we need to use at least release 1.1 of Java.

Interestingly, Java security discussions are centered on the idea of Java's applet-based security model. This security is contained inside Java-enabled browsers. This model was envisaged for use on the Internet.

1.5.3.2 The Java Sandbox

Java's security model is closely associated with the idea of *sandbox* model. A sandbox model allows a program to be hosted and executed, but there are some restrictions in place. The developers/end users may decide to give the program access to certain resources. However, in general, they want to make sure that the program is confined to the sandbox. The overall execution of a Java sandbox protects a number of resources, and it performs this task at a number of levels, as described below:

- A basic sandbox is one in which a program can access the CPU, screen, keyboard, mouse, and its own memory. It contains just enough resources for the sandbox.
- The default state of the sandbox is one in which a program can access the CPU and its memory, as well as access the Web server from which it was downloaded.

- A sandbox can exist in which a program can access the CPU, its memory, its Web server, and a set of resources (such as files computers) that are local.
- An open sandbox is one in which the program can access whatever resources the host machine can.

1.5.3.3 Java Application Security

There are some broad aspects of Java security.

- *The byte code verifier:* The byte code verifiers ensure that the Java class files obey the rules of the Java programs. However, not all files are required to go through byte code verification.
- *The class loader:* The class loader loads classes that are located in Java's default path (called CLASSPATH).
- *The access controller:* The security manager is the chief interface between the core Java API and operating system.
- *The security manager:* The security manager is the chief interface between the core Java API and the operating system. It has the ultimate responsibility for allowing or disallowing access to all operating system resources. The security manager uses the access controller for many of these decisions.
- *The security package:* The security package (that is, the classes in the Java security package)
- *The key database:* The key database is a set of keys used by the security manager and access controller to validate the digital signature that comes along with a signed class file. In the Java architecture, it is contained within the security package, although it may be an external file or database, as well.

1.5.3.4 Built-in Java Application Security

From version 1.2, the Java platform itself comes with a security model built for the applications it runs. The classes that are found in the CLASSPATH may have to go through a security check. This allows the running of the application code in a sandbox defined by a user or an administrator. The following points are salient:

- Access methods are strictly adhered to.
- A program cannot access an arbitrary memory location.

- Entities that are declared as *final* must not be changed.
- Variables may not be used before they are initialized.
- Array bounds must be checked during all array accesses.
- Objects cannot arbitrarily be casted into other object types.

The program simply declares a character pointer, and without allocating any memory, accepts user input in that pointer. This can cause havoc if an attacker finds intelligent ways to exploit such code. This is not possible in Java.

1.5.4 Specific Attacks

On the Internet, computers exchange messages with each other in the form of small groups of data, called *packets*. A packet is like an envelope that contains the actual data to be sent and the address information. Attackers target these packets as they travel from the source computer to the destination computer over the Internet. These attacks take two main forms: (a) *packet sniffing* (also called *snooping*) and (b) *packet spoofing*. The protocol used in this communication is called the *Internet Protocol (IP)*. Other names for these two attacks are (a) *IP sniffing* and (b) *IP spoofing*.

Understanding the Two Attacks

- a. *Packet sniffing*: Packet sniffing is a passive attack on a conversation. An attacker need not hijack a conversation, but instead, can simply observe (*i.e.*, sniff) the packets as they pass by. To prevent an attacker from sniffing packets, the information that is passed needs to be protected in some ways. This can be done at two levels: (i) The data that is traveling can be encoded in some way or (ii) the transmission link itself can be encoded. To read a packet, the attacker needs to access it in the first place. The simplest way to do this is to control a computer that the traffic goes through. Usually, this is a router. However, routers are highly protected resources. Therefore, an attacker might not be able to attack it and instead attack a less-protected computer on the same path.
- b. *Packet spoofing*: In this technique, an attacker sends packets with an incorrect source address. When this happens, the receiver (*i.e.*, the party who receives these packets containing a false source address) would inadvertently send replies back to this forged address (called the *spoofed address*), and not to the attacker. This can lead to three possible scenarios:

- (i.) The attacker can intercept the reply: If the attacker is between the destination and the forged source, the attacker can see the reply and use that information for the hijacking.
- (ii.) The attacker need not see the reply: If the attacker's intention was a Denial Of Service (DOS) attack, the attacker need not bother about the reply.
- (iii.) The attacker does not want the reply: The attacker could simply be angry with the host, so it may put that host's address as the forged source address and send the packet to the destination. The attacker does not want a reply from the destination, as it wants the host with the forged address to receive it and get confused.

Another attack, which is similar to these attacks, is the *DNS spoofing* attack. People usually can't identify Websites using the *Domain Name System (DNS)* because they are not really memorable (for example, 120.10.1.67). For this, a special server computer called as a DNS server maintains the mappings between domain names and the corresponding IP address. The DNS server could be located anywhere. Usually, it is with the *Internet Service Provider (ISP)* of the users. With this background, the DNS spoofing attack works as follows.

1. Suppose that there is a merchant (Bob), whose site's domain name is www.bob.com, and the IP address is 100.10.20. Therefore, the DNS entry for Bob in all the DNS is www.bob.com.
2. The attacker (Trudy) manages to hack and replace the IP address of Bob with her own (say 100.20.20.20) in the DNS server maintained by the ISP of another user, Alice. Therefore, the DNS server maintained by the ISP of Alice now has the following entry: www.bob.com, 100.20.20.20.
3. When Alice wants to communicate with Bob's site, her web browser queries the DNS server maintained by her ISP for Bob's IP address, providing it with the domain name (*i.e.*, www.bob.com). Alice gets the replaced (*i.e.*, Trudy's) IP address, which is 100.20.20.20.
4. Alice then starts communicating with Trudy, believing that she is communicating with Bob.

Such attacks of DNS spoofing are quite common and cause havoc. Even worse, the attacker (Trudy) does not have to listen to the conversation on the wire. She has to simply be able to hack the DNS server of the ISP and replace a single IP address with her own.

A protocol called the *DNSSec (secure DNS)* is being used to thwart such attacks. Unfortunately, it is not widely used.

EXERCISES

1. Find more examples of security attacks reported in the last few years.
2. What is the key principle of security?
3. Why is confidentiality an important principle of security? Think about ways of providing security. (*Hint:* Think about the ways in which children use a secret language.)
4. Discuss the reasons behind the significance of authentication. Find out the simple mechanism of authentication. (*Hint:* What information do you provide when you use a free e-mail service such as Yahoo or Hotmail?)
5. In real life, how is the message integrity ensured? (*Hint:* On what basis is a check honored?)
6. What is repudiation? How can it be prevented in real life? (*Hint:* Think what happens if you issue a cheque, and after that, tell the bank that you never issued that cheque).
7. What is access control? How different is it from availability?
8. Why are some attacks called passive? Why are others called active?
9. Discuss a passive attack.
10. What is a masquerade? Which principle of security is breached because of that?
11. What are replay attacks? Give an example of replay attacks.
12. What is a denial of service attack?
13. What is a worm? What is the significant difference between a worm and a virus?
14. Find out more about some recent worms.
15. Write a small virus-like program in plain English that accepts a file name and changes every character in the file to an asterisk.

16. Read more about computer viruses and their principles of working in detail.
17. What is a Trojan horse? What is the principle behind it?
18. What are Java applets?
19. Discuss ActiveX controls and compare them with applets.
20. Find out more about applets and ActiveX control technology.

CHAPTER 2

PUBLIC KEY CRYPTOGRAPHY AND SSL

Chapter Goals

- One-way functions
- Digital signatures
- Anatomy of a certificate
- Digital certificates
- Authentication method
- Challenge handshake authentication protocol
- Biometrics
- Mutual authentication

2.1 ONE-WAY FUNCTIONS INTRODUCTION

One-way functions, functions that are easy to compute and “hard” to invert, are an extremely important cryptographic primitive. Probably the best known and simplest use of one-way functions is for passwords. In a time-shared computer system, instead of storing a table of login passwords, we can store, for each password w , the value $f(w)$. Passwords can easily be checked for correctness at login, but even the system administrator cannot deduce any user’s password by examining the stored table. Let’s examine a theoretical treatment of one-way and trapdoor functions and discuss the candidate one-way functions proposed in the literature.

We begin by explaining why one-way functions are important to cryptography.

2.1.1 Motivation

In this section, we provide the motivation for one-way functions. We argue that the existence of one-way functions is a necessary condition to the existence of most known cryptographic primitives (including secure encryption and digital signatures). As the current state of knowledge in complexity theory does not allow us to prove the existence of a one-way function (even using more traditional assumptions, such as $P \neq NP$), we will have to assume the existence of one-way functions. We will later try to provide evidence for the plausibility of this assumption. As stated in the introductory chapter, modern cryptography is based on a gap between efficient algorithms guaranteed for the legitimate user versus the unfeasibility of retrieving protected information for an adversary. To make the following discussion clearer, let us concentrate on the cryptographic task of secure data communication, namely encryption schemes. In secure encryption schemes, the legitimate user is able to decipher the messages (using some private information available to him), yet for an adversary (not having this private information) the task of decrypting the cipher text should be infeasible. Clearly, the breaking task can be performed by a non-deterministic polynomial time machine. Yet, the security requirement states that the breaking task should not be feasible, *i.e.*, it cannot be performed by a probabilistic polynomial time machine. Hence, the existence of a secure encryption scheme implies that there are tasks performed by non-deterministic polynomial time machines that cannot be performed by deterministic (or even randomized) polynomial time machines. In other words, a necessary condition for the existence of secure encryption schemes is that NP (nondeterministic polynomial time) is not contained in BPP (bounded-error probabilistic polynomial time; hence, $P \neq NP$). However, the necessary condition (*e.g.*, $P \neq NP$) is not a sufficient one. $P \neq NP$ only implies that the encryption scheme is hard to break in the worst case. It does not eliminate the possibility that the encryption scheme is easy to break in almost all cases. In fact, one can easily construct “encryption schemes” for which the breaking problem is NP -complete and yet there exists an efficient breaking algorithm that succeeds in 99% of the cases. Hence, worst-case hardness is a poor measure of security. Security requires hardness for most cases or at least average-case hardness. Hence, a necessary condition for the existence of secure encryption schemes is the existence of languages in NP that are hard on average. Furthermore, PNP is not known to imply the existence of languages in NP that are hard on

average. The mere existence of problems (in NP) that are hard on the average does not suffice. In order to be able to use such problems, we must be able to generate such hard instances together with auxiliary information that enables us to solve these instances fast. Otherwise, the hard instances are also hard for the legitimate users and they gain no computational advantage over the adversary. Hence, the existence of secure encryption schemes implies the existence of an efficient way (*i.e.*, a probabilistic polynomial time algorithm) of generating instances with the corresponding auxiliary input so that

1. it is easy to solve these instances given the auxiliary input and
2. it is hard on the average to solve these instances (when not given the auxiliary input)

We avoid formulating the above definition. We only remark that the coin tosses used to generate the instance provide sufficient information to allow us to efficiently solve the instance. Hence, without the loss of generality, we can replace a condition by requiring that these coin tosses are hard to retrieve from the instance. The last simplification of the above conditions essentially leads to the definition of a one-way function.

2.2 ONE-WAY FUNCTIONS: DEFINITIONS

In this section, we present several definitions of one-way functions. The first version, hereafter referred to as a strong one-way function (or just one-way function), is the most convenient one. We also present weak one-way functions that may be easier to find and yet can be used to construct strong one-way functions and non-uniform one-way functions.

2.2.1 (Strong) One-Way Functions

The most basic primitive for cryptographic applications is a one-way function. Formally, this is a function that is “easy” to compute but “hard” to invert. Any Probabilistic Polynomial Time (PPT) algorithm attempting to invert the one-way function on an element in its range will succeed with no more than a “negligible” probability, where the probability is taken over the elements in the domain of the function and the coin tosses of the PPT attempting the inversion. This informal definition introduces a couple of erasures that are prevalent in complexity theoretic cryptography. An easy computation is one that can be carried out by a PPT algorithm. A function $v: N \rightarrow R$ is negligible

if it vanishes faster than the inverse of any polynomial. More formally, this means v is negligible if, for every constant $c \leq 0$, there exists an integer K_c such that $v(k) < k^{-c}$ for all $k \geq K_c$.

Another way to think of it is $v(k) = k^{-w}(1)$.

The above definition and discussion consider the success probability of an algorithm to be negligible if, as a function of the input length, the success probability is bounded by any polynomial function. It follows that repeating the algorithm polynomial (in the input length) many times yields a new algorithm that also has a negligible success probability. In other words, events that occur with negligible (in n) probability remain negligible, even if the experiment is repeated for the polynomial (in k) many times. Hence, defining negligible success as occurring with a probability smaller than any polynomial fraction is naturally coupled with defining feasible as computed within polynomial time. A “strong negation” of the notion of a negligible fraction/probability is the notion of a non-negligible fraction/probability. We say that a function v is non-negligible if there exists a polynomial p such that for all sufficiently large k s it holds that $v(k) > 1/p(k)$. Note that functions may be neither negligible nor non-negligible.

2.3 DIGITAL SIGNATURES

A digital signature is a mathematical scheme for demonstrating the authenticity of a digital message or documents. A valid digital signature gives a recipient reason to believe the message was created by a known sender. Digital signatures are commonly used for software distribution and financial transactions. A major benefit of public key cryptography is that it provides a method for employing digital signatures. Digital signatures enable the recipient of information to verify the authenticity of the information’s origin and also verify that the information is intact. Thus, public key digital signatures provide authentication and data integrity. A digital signature also provides evidence of non-repudiation, which means that it prevents the sender from claiming that he or she did not actually send the information. These features are every bit as fundamental to cryptography as privacy, if not more. A digital signature serves the same purpose as a handwritten signature. However, a handwritten signature is easy to counterfeit. A digital signature is superior to a handwritten signature in that it is nearly impossible to counterfeit, plus it attests to the contents of the information as well as to the identity of the signer. Some people tend to use signatures more than they use encryption. For example, you may not care if anyone knows that you just deposited \$1,000 in your account, but you do

want to be sure it was a bank teller you were dealing with. The basic manner in which digital signatures are created is illustrated in Figure 2.1. Instead of encrypting information using someone else's public key, you encrypt it with your private key. If the information can be decrypted with your public key, then it must have originated with you.

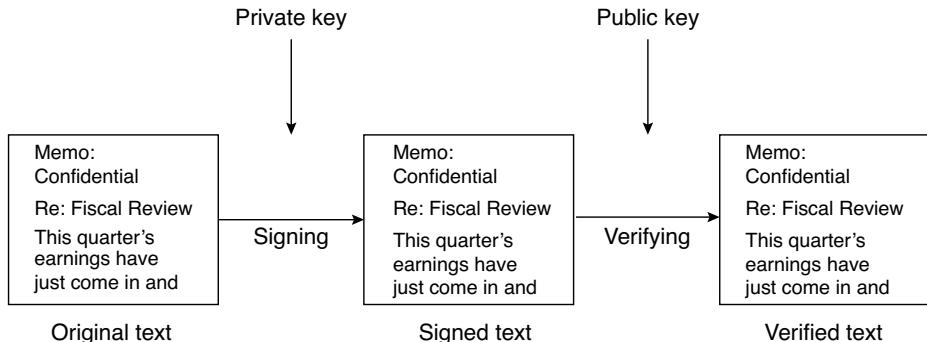


FIGURE 2.1 Simple digital signatures

2.4 HASH FUNCTIONS

A hash function is any well-defined procedure or mathematical function that converts a large variable amount of data into a small amount of data, usually using a single integer that may serve as an index to an array. The value returned by a hash function is called a *hash value*. The system described above has some problems. It is slow, and it produces an enormous volume of data at least double the size of the original information. An improvement on the above scheme is the addition of a one-way hash function. A one-way hash function takes variable-length input, in this case, a message of any length (it can even be thousands or millions of bits), and produces a fixed-length output (for example, 160 bits). The hash function ensures that, if the information is changed in any way, even by just one bit, an entirely different output value is produced. PGP uses a cryptographically strong hash function on the plaintext the user is signing. This generates a fixed-length data item known as a *message digest*. (Again, any change to the information results in a totally different digest.) Then PGP uses the digest and the private key to create the “signature.” PGP transmits the signature and the plaintext together. Upon receipt of the message, the recipient uses PGP to recompute the digest, thus verifying the signature. PGP can encrypt the plaintext or not; signing plaintext

is useful if some of the recipients are not interested in or capable of verifying the signature. As long as a secure hash function is used, there is no way to take someone's signature from one document and attach it to another or to alter a signed message in any way. The slightest change in a signed document will cause the digital signature verification process to fail.

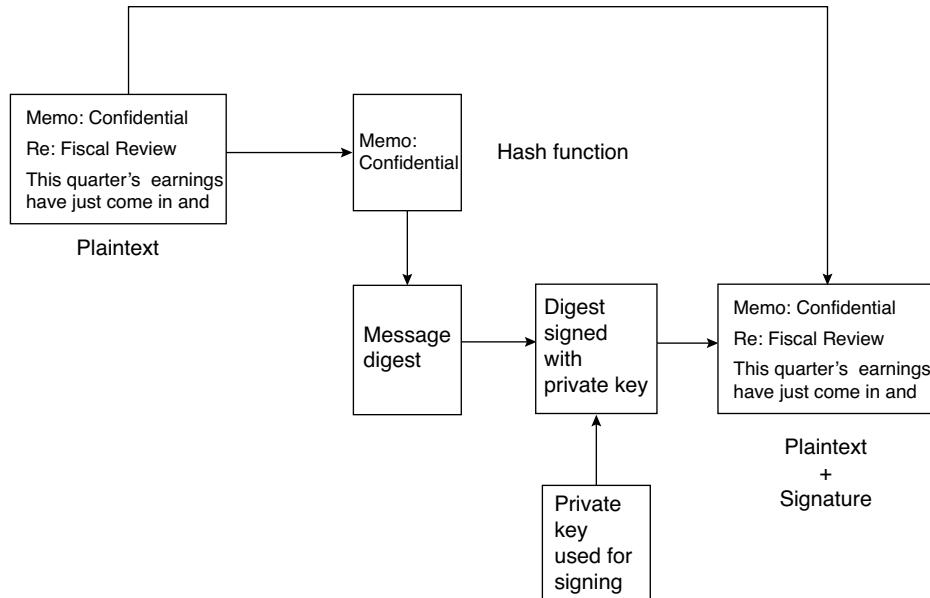


FIGURE 2.2 Secure digital signatures

Digital signatures play a major role in authenticating and validating other PGP users' keys.

2.5 CENTRALIZED CERTIFICATES

One issue with public key cryptosystems is that users must be constantly vigilant to ensure that they are encrypting to the correct person's key. In an environment where it is safe to freely exchange keys via public servers, *man-in-the-middle attacks* are a potential threat. In this type of attack, someone posts a phony key with the name and user ID of the user's intended recipient.

Data encrypted to, and intercepted by, the true owner of this bogus key is now in the wrong hands. In a public key environment, it is vital that you are assured that the public key to which you are encrypting data is in fact the public key of the intended recipient and not a forgery. You could simply encrypt

only to those keys which have been physically handed to you. But suppose you need to exchange information with people you have never met; how can you tell that you have the correct key?

Digital certificates, or *certs*, simplify the task of establishing whether a key truly belongs to the purported owner. Webster's dictionary defines a certificate as "a document containing a certified statement, especially as to the truth of something." A certificate is a form of credential. Examples might be your passport, your social security card, or your birth certificate. Each of these has some information on it identifying you and some authorization stating that someone else has confirmed your identity. Some certificates, such as your driver's license, are important enough confirmation of your identity that you would not want to lose them, lest someone use them to impersonate you. A digital certificate is data that functions much like a physical certificate. A digital certificate is information included with a person's public key that helps others verify that a key is genuine or valid. Digital certificates are used to thwart attempts to substitute one person's key for another. A digital certificate consists of three things:

- a public key
- certificate information ("identity" information about the user, such as the name and user ID)
- one or more digital signatures

The purpose of the digital signature on a certificate is to state that the certificate information has been attested to by some other person or entity. The digital signature does not attest to the authenticity of the certificate as a whole; it vouches only that the signed identity information goes along with, or is bound to, the public key. While some security experts believe it is not a good practice to mix professional and personal identity information on one key, but rather have separate keys for each, you will come across certificates containing a public key with several associated identities (for example, the user's name and corporate email account, the user's nickname and home email account, or the user's maiden name and college email account—all in one certificate). The list of signatures of each of those identities may differ; signatures usually attest to the authenticity of one of the identities, not that all three are authentic. For example, suppose your coworker, Alice, asks you to sign her certificate. You look it up on the server and see that Alice has two pieces of identity information associated with the certificate. The first one reads "Alice Petucci, alice@securecompany.com." Depending on how well you know Alice, you might want to choose to sign only the one that relates to the Alice you know at work.

When a user encrypts plaintext with PGP, PGP first compresses the plaintext.

Data compression saves modem transmission time and disk space and, more importantly, strengthens cryptographic security. Most cryptanalysis techniques exploit patterns found in the plaintext to crack the cipher. Compression reduces these patterns in the plaintext, thereby greatly enhancing resistance to cryptanalysis. (Files that are too short to compress or that don't compress well aren't compressed.)

PGP then creates a session key, which is a one-time-only secret key. This key is a random number generated from the random movements of your mouse and the keystrokes you type. This session key works with a very secure, fast conventional encryption algorithm to encrypt the plaintext; the result is ciphertext. Once the data is encrypted, the session key is then encrypted to the recipient's public key. This public key-encrypted session key is transmitted along with the ciphertext to the recipient.

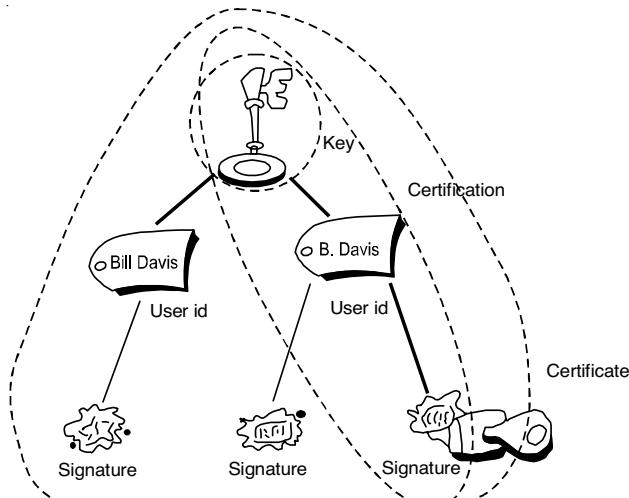


FIGURE 2.3 Anatomy of a certificate

2.6 RANDOM KEY GENERATION

A key is a value that works with a cryptographic algorithm to produce a specific cipher text. Keys are basically really big numbers. Key size is measured in bits; the number representing a 1024-bit key is enormous. In public key

cryptography, the bigger the key, the more secure the cipher text. However, public key size and conventional cryptography's secret key size are totally unrelated. A conventional 80-bit key has the equivalent strength of a 1024-bit public key. A conventional 128-bit key is equivalent to a 3000-bit public key. Again, the bigger the key, the more secure, but the algorithms used for each type of cryptography are very different and thus comparison is like that of apples to oranges. While the public and private keys are related, it's very difficult to derive the private key given only the public key; however, deriving the private key is always possible given enough time and computing power. This makes it very important to pick keys of the right size: large enough to be secure, but small enough to be applied fairly quickly. Additionally, you need to consider who might be trying to read your files, how determined they are, how much time they have, and what their resources might be.

Larger keys will be cryptographically secure for a longer period of time. If what you want to encrypt needs to be hidden for many years, you might want to use a very large key. Of course, who knows how long it will take to determine your key using tomorrow's faster, more efficient computers? There was a time when a 56-bit symmetric key was considered extremely safe. Keys are stored in encrypted form. PGP stores the keys in two files on your hard disk; one for public keys and one for private keys. These files are called *keying*. As you use PGP, you will typically add the public keys of your recipients to your public keying. Your private keys are stored on your private keying. If you lose your private keying, you will be unable to decrypt any information encrypted to keys on that ring.

2.7 AUTHENTICATION METHODS

All methods of authentication require you to specify who or what you are and to relay the appropriate credentials to prove that you are who you say you are. These credentials generally take the form of something you know, something you have, or something you are. What you know may be a password. What you have could be a smart card. What you are pertains to the field of biometrics, in which sophisticated equipment is used to scan a person in some sense to provide authentication. The important element is to recognize that different mechanisms provide authentication services with varying degrees of certainty. Choosing the proper authentication technology largely depends on the location of the entities being authenticated and the degree of trust placed in the particular facets of the network.

2.8 EMAIL SECURITY

In the last few years, several important research papers have examined how to protect information or resources from unauthorized users. Password authentication is the most commonly used method because it is easy to use and relatively inexpensive. The user name and password are often used in the real world. It's a method that is based on "what you know" for authentication. It is the simplest way of authentication and it provides each user with a unique user name and a secret password. Unfortunately, it is also easy to get attacked by password guessing.

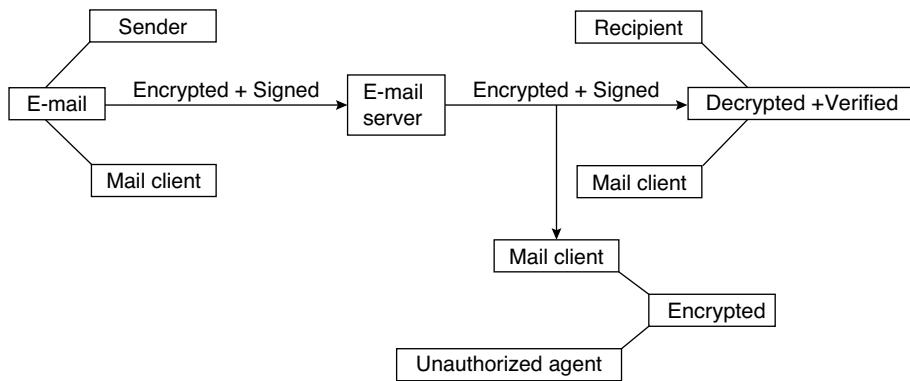


FIGURE 2.4 Protecting information or resources from unauthorized users

The biggest problem of a good password is that users forget them easily. Although most users can remember one or two IDs and passwords for sites they often visit, it is becoming impossible to remember several different IDs and passwords for sites that are visited less frequently. As long as a user can enter the correct user name and password, the computer and system will assume that the operator who using the computer is the legal user or original user. Actually, many of the users often use a character string, which is easy to guess, such as the date of a birthday, telephone number, or a pet's name to prevent forgetting their password. Many users also write down the password on paper and hide it some place where they think it might be safe. Thus, it may cause the information to get lost or to be stolen. Even if one can ensure that the user name and password won't be stolen, it's still not very safe. This is because the user's name and password are stored in a static state in a database. Thus, the name and password need to be transferred on the network platform and over computers' memories during the authentication process. The authentication information will be the same every time, so it is easy to get

attacked by Trojan horses or spy monitors watching traffic on the network. Thus, from a security point of view, using a user name and password is an unsecure method for authentication.

2.9 CHALLENGE HANDSHAKE AUTHENTICATION PROTOCOL

The *Challenge Handshake Authentication Protocol (CHAP)* is used to periodically verify the identity of the peer using a 3-way handshake. CHAP is a remote access authentication protocol that is used in conjunction with a Point-to-Point Protocol (PPP) to provide security and authentication to users of remote resources. The process of authentication works in the following manner:

1. After the link establishment phase is complete, the authenticator sends a challenge message to the peer.
2. The peer responds with a value calculated using a one-way hash function.
3. The authenticator checks the response against its own calculation and is acknowledged; otherwise the connection is terminated.
4. At random intervals, the authenticator sends a new challenge to the peer, and repeats steps 1 through 3.

CHAP provides protection against the playback attack by the peer through the use of an incrementally changing identifier and a variable challenge value. The use of repeated challenges is intended to limit the time of exposure to any single attack. The authenticator is in control of the frequency and timing of the challenges. This authentication method depends upon a “secret” known only to the authenticator and the peer. The secret is not sent over the link. Although the authentication is only one-way, by negotiating CHAP in both directions, the same secret set may easily be used for mutual authentication. Since CHAP may be used to authenticate many different systems, the name fields may be used as an index to locate the proper secret in a large table of secrets. This also makes it possible to support more than one name/secret pair per system, and to change the secret in use at any time during the session. However, CHAP requires that the secret be available in plaintext form. The irreversibly encrypted password databases commonly available cannot be used. It is not useful for larger installations, since every possible secret is maintained at both ends of the link.

2.10 AUTOMATIC REKEYING

Most companies and organizations have an open distributed environment, with employees free to access services on servers from workstations located anywhere in the world. The server needs to be able to control access to authenticated users and also authenticate users' requests for the service. However, in this environment, any of the workstations might not be trusted, and it is hard to authenticate their users through the network. One of the potential threats is that an unauthorized user might pretend to be another user to access the workstation. Kerberos was developed to address this problem. It is a centralized authentication method used to authenticate the user to the server and the server to user.

Kerberos is an authentication protocol based on conventional encryption that has received widespread support and is used in a variety of systems. Kerberos includes the Authentication Server (AS) and Ticket-Granting Server (TGS). When a user sends a request to the server for login, then the AS will verify the user's identity in the system's database and generate the TGS ticket and a session key to the workstation. The workstation decrypts the message and sends the ticket and authenticator to the TGS and server. The server verifies the ticket and authenticator; if they are a match, then the server grants access to the services.

A *token* is a security device that authenticates the user by having the appropriate permission (such as a password) embedded into the token itself. The essence of token-based authentication is that you must have the token in your possession to authenticate yourself to the computer. The computer will not recognize you without the token, regardless of whether the token was lost, lent, or stolen. Here is a summary of the fundamental properties of tokens from a security standpoint:

- A person must physically possess the token to use it.
- A good token is hard to duplicate.
- A person can lose a token and unintentionally lose access to a critical resource.
- People can detect stolen tokens by taking an inventory of the tokens they should have in their possession.

Tokens generally fall into two categories: passive and active. In both cases, the token incorporates a base secret, and one must copy the base secret to make a working copy of a particular token. A passive token is simply a storage device for the base secret. Some examples are ATM cards, mechanical keys,

and specialized devices like “data keys.” An active token can generate different outputs under different circumstances. For example, an active token can take part in a challenge response authentication protocol or provide other crypto functions that use the token’s base secret. Traditionally, active tokens have been either commercial one-time password tokens or smart cards, though other models have evolved that plug into existing ports on desktop and laptop computers.

Tokens are a passive authentication method in the form of a plastic card with a magnetic strip. Today, they appear everywhere: ATM cards, credit cards, driver’s licenses, supermarket VIP cards, and employee badges to operate electronic door locks.

2.11 BIOMETRICS

Biometrics was first introduced in the 1970s and early 1980s. This technology gathers the unique physiological or behavioral attributes of a person for storage in a database or comparison with data already found in the database. *Biometrics* look for specific characteristics, often unique and measurable, that allow for the recognition and identification of human. These days, biometric technologies are typically used to analyze human characteristics for security purposes. Five of the most common physical biometric patterns analyzed for security purposes are taken from fingerprints, hands, eyes, face, and voice. A definition used in the biometrics industry includes the following:

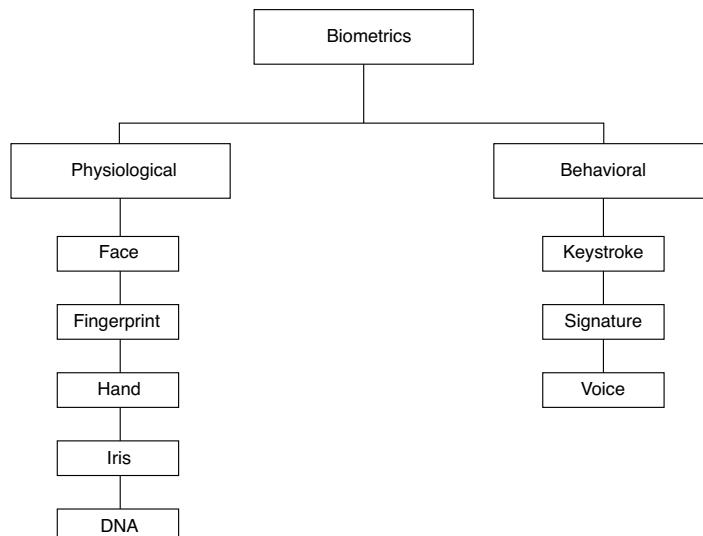


FIGURE 2.5 Biometric technologies are typically used to analyze human characteristics for security purposes

- The lowest level of security is defined as something you have in your possession, such as an ID badge with a photograph on it.
- The second level of security is something that you know, such as a password used with a computer login or PIN code to use your bank account card on an ATM.

The highest level of security is information about what you are and something that you do. This is essentially what biometric technology is all about.

Physiological biometrics use algorithms and other methods to define identities in terms of data gathered from the direct measurement of the human body. Biometric information includes that from fingerprints, hand geometry, iris and retina scanning, facial geometry, iris patterns, facial features, and retinal blood vessels.

Behavioral biometrics involves analyzing a specific action of a person. How a person talks, signs their name, or types on a keyboard is a method of determining identity when analyzed correctly. Biometrics can be defined as either passive or active. Passive biometrics does not require a user's active participation and can be successful without a person even knowing that they have been analyzed. Active biometrics does require a person's cooperation and will not work if they deny their participation in the process. Mouse dynamics is one type of biometric that can be used in both active as well as passive monitoring. Although biometric technologies differ, they all work in a similar fashion: the user submits a sample that is an identifiable, unprocessed image or recording of the physiological or behavioral biometric via an acquisition device (for example, a scanner or camera). This biometric is then processed to extract information about distinctive features to create a trial template or verification template. Figure 2.6 shows some often-used biometrics methods.

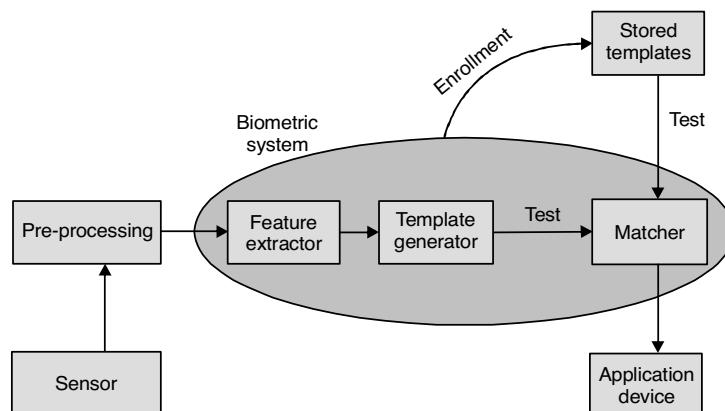


FIGURE 2.6 Face recognition consists of several steps

It can be scanned optically, but the cameras needed are bulky. A capacitive technique uses the differences in the electrical charges of the whorls on the finger to detect those parts of the finger touching a chip and not touching it. The data is converted into a graph where the ridges are represented by vertices, and the vertices corresponding to the adjacent ridges are connected. Each vertex has a number approximating the length of the corresponding ridge. At this point, determining matches becomes a problem of graph matching. This problem is similar to the classical graph isomorphism problem, but because of the imprecision in measurements, the graph generated from the fingerprint may have different numbers of edges and vertices. Thus, the matching algorithm is an approximation.

Authentication by voice, also called *speaker verification* or *speaker recognition*, involves the recognition of a speaker's voice characteristics or verbal information verification. The former uses statistical techniques to test the hypothesis that the speaker's identity is as claimed. The system is first trained on fixed pass-phrases or phonemes that can be combined. To authenticate, either the speaker says the pass-phrase or repeats a word (or set of words) composed of the learned phonemes. Verbal information verification deals with the contents of utterance. The system asks a set of question such as "What is your mother's middle name?" and "In which city were you born?" It then checks that the answers spoken are the same as the answers recorded in its database. The key difference is that speaker verification techniques are speaker-dependent, but verbal information verification techniques are speaker-independent, and only rely on the content.

Authentication by eye characteristics uses the iris and the retina. Patterns within the iris are unique for each person. Hence, one verification approach is to compare the patterns statistically and ask whether the differences are random. A second approach is to correlate the images using statistical tests to see if they match. Retinal scans rely on the uniqueness of the patterns made by blood vessels at the back of the eye. This requires a laser beam to scan the retina, which is highly intrusive. This method is typically used only in more secure facilities.

Facial recognition consists of several steps. First, the face is located. If the user places her face in predetermined position (for example, by resting her chin on a support), the problem becomes somewhat easier. However, facial features such as hair and glasses may make the recognition harder. Techniques for doing this include the use of a neural network and templates. The resulting image is then compared with the relevant image in the database. The correlation is affected by the differences in the lighting between the current image and the reference image, by distortion, by noise, and by the view of the

face. The correlation mechanism must be trained. Several different methods of correlation have been used, with varying degrees of success. An alternative approach is to focus on the facial features, such as the distance between the nose and the chin, and the angle of the line drawn from one to the other.

Keystroke dynamics methods are now used to identify individuals. Many security analysts believe the measurement of keystroke pressure, intervals of keystrokes, and how hard the key is pushed, can help to identify users.

Table 2.1 Overview of Biometrics.

Biometric	Acquisition device	Sample	Feature extracted
Iris	Infrared-enabled video camera, PC camera	Black and white iris image	Furrows and striations of iris
Fingerprint	Desktop peripheral, PC card, mouse chip, or reader	Fingerprint image (optical, silicon, ultrasound, or touchless)	Location and direction of ridge endings and bifurcations on fingerprint, minutiae
Voice	Microphone, telephone	Voice recording	Frequency, cadence, and duration of the vocal pattern
Signature	Signature tablet, motion-sensitive stylus	Image of signature and record of related dynamics	Speed, stroke order, pressure, and appearance of signature
Face	Video camera, PC camera, single-image camera	Facial image (optical or thermal)	Relative position and shape of nose, position of cheekbones
Hand	Proprietary Wall-mounted unit	3-D image of top and sides of hand	Height and width of bones and joints in hands and fingers
Retina	Proprietary desktop or wall mountable unit	Retina Image	Blood vessel patterns and retina

2.12 PUBLIC KEY CRYPTOGRAPHY

Certificates are messages that are authorized by trusted third parties. All parties using certificates must ensure the publicly-trusted third parties are absolutely secure and can be trusted. *Public Key Certificate* “formats” are used to guarantee this high level of security. The X.509 V3 public key certificate is a popular format used to secure documents appearing on Web sites, organizations, etc.

The certificate contains various types of information: version, serial number, signature, issuer, validity, subject, subject public key information, issuer’s unique ID, subject’s unique ID, extensions, signature algorithm, and signature value.

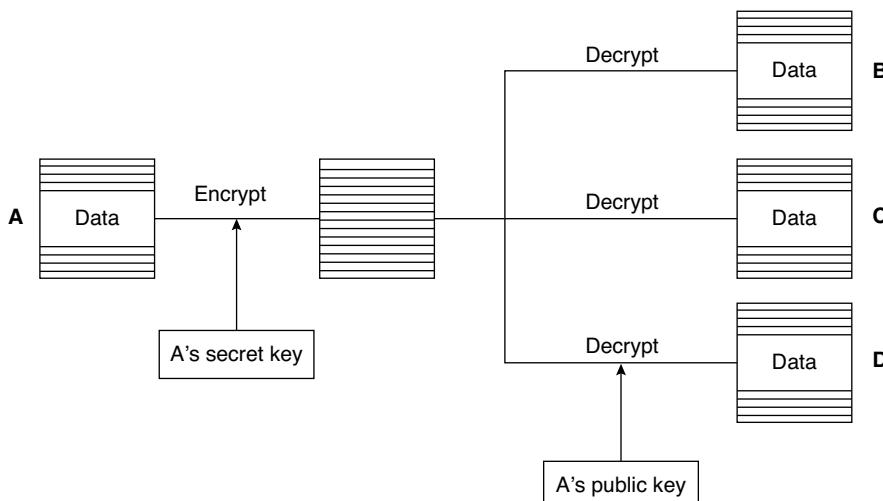


FIGURE 2.7 Public key authentication method

Certificate authentication (CA) is performed via the public key authentication method. Instead of the client sending just a public key, it sends a certificate containing a public key. In brief, the certificate authentication works in the following way. The client sends the user certificate (which includes the user’s public key) to the server. The server uses the CA certificate to check that the user’s certificate is valid. The server uses the user certificate to check from its mapping file(s) whether the login is allowed or not. Finally, if the connection is allowed, the server makes sure that the user has a valid private key by using a challenge. Compared to traditional public key authentication, this

method is more secure because the system checks that the user certificate was issued by a trusted CA. In addition, certificate authentication is convenient because no local database of the users' public keys is required on the server.

2.13 MUTUAL AUTHENTICATION

Mutual authentication is also called two-way authentication consisting of a client and a server. The client needs to prove its identity to the server, and the server proves its identity to the client before any application will start. It doesn't involve user interaction in either the client process or the server process, so the client can trust the actual organization or entity's certification chain and build the connection between them. The importance of using mutual authentication is obvious.

The mutual authentication is often used on e-business and online banking. By using mutual authentication, users trust the trade entity's certificate or an entity in the certificate chain, which means users trust a third-party authority. For example, when we shop on eBay, we do not need to pay money directly to the shop owner, since we do not trust the offline private payment and the store owner. eBay accepts payments from the online payment service PayPal, which means we can send money through the PayPal platform. Then, PayPal will transfer money to the shop owner. PayPal is the third party in the online transaction.

2.14 MULTIFACTOR AUTHENTICATION

Multifactor authentication is used to determine the right of an individual to access a physical facility or to access data within an information system.

There are many commonly used authenticate methods available, but they all have limitations or might be easy affected by the environment. In order to develop a stronger authentication, we must combine those methods. For example, when you use the ATM machine, you need to insert your bank card and then enter your PIN code. The bank card is something you have, and the PIN code, only known to you, is something you know. You can then access your account profile and take out money. The multifactor authentication method makes transactions more reliable. Individually, any one of these approaches has its limitations. "Something you have" can be stolen, whereas "something you know" can be guessed, shared, or forgotten. "Something you

are” is generally the strongest approach, but it can be costly to implement. To make authentication stronger, we can combine methods, often referred to as multifactor or strong authentication. The most common type is two-factor authentication, such as using a PIN code as well as a Secure ID token to log on to the network. An example of two-factor authentication that we are probably most familiar with is our ATM card—we insert the card (something you have) into the ATM machine and enter PIN (something you know) to access our account number and perform transactions. We can also use three-factor authentication. For example, if we use biometrics to authenticate users on a network, we can store the fingerprint information on the system’s database, which is only accessible with the user’s PIN. Without the PIN and a fingerprint scan, the user cannot access the system.

2.15 ELEMENTS OF AN AUTHENTICATION SYSTEM

There are several elements that need to be considered in an authentication system. The first set of elements includes people, principals, and entities. This group represents certain people or groups of people who need to authenticate the authenticating target. The second set of elements encompasses the need to distinguish different characteristics between certain people and groups. The third type of element includes each system, server, or organization that has its owner or administrator who can authorize users and distinguish certain people or group characteristics from other people or groups. The authentication mechanism validates information from users. It can respond to words or numbers entered by users and then it decides if they match or not. In the fifth set of elements, the administrator grants certain people or groups access to the system or server when the authentication has succeeded by using access control mechanisms. However, if the authentication process fails, then the mechanisms do not allow the user access to the system or server.

Table 2.2 Examples of the Five Elements in an Authentication System.

Authentication element	Cave of the 40 thieves	Password login	Teller machine	Web server to client
Person, principal, entity	Anyone who knew the password	Authorized user	Owner of a bank account	Web site to client

(continued)

(continued)

Distinguishing characteristic	The password “open, sesame”	Secret password	ATM card and PIN	Public key within a certificate
Proprietor, system owner	The forty thieves	Enterprise owning the system	Bank	Certificate authority
Authentication mechanism	Magical device that responds to the words	Password validation software	Card validation software	Certificate validation software
Authentication element	Cave of the 40 thieves	Password login	Teller machine	Web server to client
Access control mechanism	Mechanism to roll the stone from in front	Login process, access controls	Allows banking transactions	Brower marks the page “secure”

Table 2.2 shows an example based on the story *Alibaba and the 40 Thieves*. Alibaba found a cave with treasure that was guarded by 40 thieves. His brother also knew about this cave and tried to find the secret code that would open the cave’s door. The distinguishing characteristic was the knowledge of the password: Alibaba heard the thieves say “open sesame” to access the cave. The owner and proprietors were the 40 thieves. The stone would move away by the access control mechanism that granted access to the cave. Another example of authentication is that of student user names and passwords. By entering a user name and password, the computer system then identifies the person as a student. The distinguishing characteristic for each student is the password, which allows the user to access the system. During the login process, the computer system compares the entered password and user name against the ones in the database. If the two match, then the access control mechanism allows the user to proceed with using the system. By using unique user names and passwords, the system can make the access control decision for protected resources.

2.16 ATTACKS

In *The Art of War*, Sun Tzu said that the best way to defend yourself is to know your enemy well. In the world of computer security, users are always trying to protect themselves against different attacks, threats, and potential enemies. Thus, we need to know our enemy better.

An *attack* is different than a threat. Threats appear every day and target numerous computer systems, Web sites, etc. often randomly, without specific targets. Attacks, however, represent specific actions against a controlled system, especially targeting the weaknesses of that system in order to result in some type of failure or loss.

Attacks are classified into five different levels of prevalence. The different levels indicate the relative knowledge and resources the attacker will need, as well as the degree to which such attacks seem to occur. Attackers faced with very attractive targets will expend a significant amount of effort on intrusions.

- *Trivial attacks:* Anyone who knows the “trick” behind one of these attacks can perform them using conventional software already present on a typical workstation. The attack doesn’t rely on special hacker tools or non-standard software. Anyone can do it if they know how.
- *Common attacks:* The attacker must acquire specific software tools or take similar steps that may indicate premeditation. Common attacks often have password sniffing and/or viruses.
- *Physical attacks:* These attacks require the attacker’s physical presence at the point of attack. They may also require the manipulation of computer hardware and/or special hardware tools. The attacks may also depend on special knowledge and training.
- *Sophisticated attacks:* A sophisticated attack is an attack that requires a sophisticated knowledge of security vulnerabilities. While common attacks may be performed by people with good tools and limited knowledge or skills, a sophisticated attack may require the attacker to construct a tool to implement the attack.
- *Innovative attacks:* These are attacks that exploit theoretical vulnerabilities that have not been publicly demonstrated to be practical. They often utilize a significant amount of resources to breach a strong security mechanism.

Table 2.3 Attack Summary.

Attack	Security problem	Prevalence	Attack description
Keystroke confusion	Masquerade as someone else	Obsolete	A bug found in the timesharing software allowed a peculiar sequence of characters to skip password checking
Password file theft	Recover all other users' passwords	Obsolete	Weak protection of password file allowed its contents to be stolen
Trojan horse	Recover hidden information, like a password file	Common, sophisticated, or innovative	Attacker writes a program that gets used by the victim. Unknown to the victim, the program copies or modifies the victim's data. A virus is a well-known example
On-line password guessing	Recover a user's password	Trivial	Interactive trial-and-error attack to try to guess a user's password
Password audit make review	Recover a user's password	Common	Review audit records of a user's mistakes while logging on to make guesses of the user's password
Helpful disclosure	Recover a user's password	Trivial	Attacker convinces a victim to reveal a password in support of an apparently important task
Bogus password change	Recover a user's password	Trivial	Attacker convinces victim to change their passwords to a word selected by the attacker
Rubber hose disclosure	Recover hidden information, like a user's password	Physical	Attacker uses threats or physical coercion to recover secret information from the victim

(continued)

Attack	Security problem	Prevalence	Attack description
Shoulder surfing	Recover a user's password	Trivial	Attacker watches a user type his password, then uses it himself
Keystroke sniffing	Recover a user's password	Common	Software watches keystrokes transmitted from the user to the system for typed-in user names and passwords, saved for later user
Trojan login	Recover a user's password	Common	Run a program that mimics the standard login program, but collects user names and passwords when people try to log on

1. *Password crack*: Attempting to reverse-calculate a password is often called *cracking*. A cracking attack is a component of many dictionary attacks. It is used when a copy of the security account manager data file can be obtained.
2. *Brute force*: The application of computing and network resources to try every possible combination of options of a password is called a brute force attack. Since this is often an attempt to repeatedly users' passwords to commonly used accounts, it is sometimes called a *password attack*.
3. *Dictionary*: This is another form of the brute force attack noted above for guessing passwords. The dictionary attack narrows the field by selecting specific accounts to attack and uses a list of commonly used passwords (the dictionary) instead of random combinations.
4. *Spoofing*: Spoofing is a technique used to gain unauthorized access to computers, wherein the intruder sends messages to a computer that has an IP address that indicates that the messages are coming from a trusted host. To engage in IP spoofing, a hacker must first use a variety of techniques to find an IP address of a trusted host and then modify the packet headers so that it appears that the packets are coming from that host.

2.17 IP SECURITY ENCRYPTION ROUTER

Routers are a critical element of both the Internet and corporate network infrastructures. They control the flow of data packets on a network and determine the best way to reach the appropriate destination. On corporate networks, they are often used to separate network segments. In addition, border routers are often the first line of defense in firewall configurations and are a key component of most VPNs. Routers are network devices that operate at the network layer (layer 3) of the OSI model that are employed to connect two or more networks. They serve three primary purposes. First, they route network traffic based on predetermined rules or routing tables. Second, they segment frames for transmission between LANs. For example, they can frame 10-Mbps Ethernet LAN frames for transmission on a 16-Mbps token ring LAN or a frame relay connection WAN. Third, routers provide the ability to deny or block unauthorized traffic. This can be accomplished through filtering commands that limit certain protocols (*i.e.*, HTTP, FTP, or SNMP) or by employing access lists.

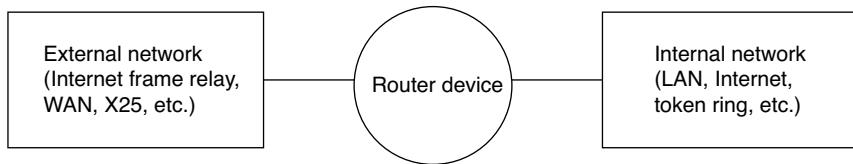


FIGURE 2.8 Basic router configurations

Even though routers are ubiquitous, they tend to be overlooked when security measures are developed. No security measures can be considered to be comprehensive unless they include the control and management of routers.

Risks

It is important to understand that routers are subject to many of the same risks associated with computers. In fact, the first routers were actually modified computers. A router has an operating system that needs to be configured and, like any OS, that can be subject to bugs. Just as with computers, proper password controls are critical to router security. Routers should not run unnecessary services or protocols. Routers can be affected by denial-of-service attacks. They need to be monitored, just like computers. How well the router is configured and maintained is critical to the availability of the network. In many ways, an incorrectly configured router is an even greater risk

than an incorrectly configured computer. An incorrectly configured computer usually only affects local users of the system. An incorrectly configured router can affect everyone on the network. There are severe consequences that can result from incorrectly modifying routing tables. In 1997, a major portion of the Internet was practically shut down by the incorrect routing tables of a small backbone service provider. The service provider sent the incorrect routing tables to other backbone providers that essentially sent all network traffic to the small provider. The problem took three hours to resolve, during which time it is estimated that 30–40% of all Internet traffic was lost. Such changes can have a crippling effect on a network if a hacker is able to gain privileged access to its routers. A simple denial-of-service attack launched against a router can cripple a network.

Cisco IOS

The dominant player in networking today is Cisco Systems. They have approximately 80% to 90% of the market for routers, switches, and hubs. The vast majority of routers on corporate networks and on the Internet are Cisco products. To illustrate how similar routers and servers are when it comes to security, we can use Cisco's IOS. IOS is the operating system that Cisco routers run. An example of one of the concerns that IOS shares with computer operating systems is the concept of the banner or message of the day. IOS can be configured with a banner. Just as with a server, you run the risk of providing information in a banner that could be useful to a hacker. Of course, this should be avoided. Cisco's IOS supports multiple password levels and encrypted passwords. However, the default at installation is not to encrypt the password. This is important because if the password is not encrypted, it is readable in the configuration file if you do a “show startup” or “show run.” In addition, it is a common practice to store router configuration files on network TFTP (Trivial File Transfer Protocol) servers. This is done so that the network administrator can update the non-volatile RAM (NVRAM) on the router from the copy of the configuration file on TFTP server. For example, an administrator may use the configuration file on the TFTP server to reload a “clean” configuration file onto a router if he or she garbled the existing file in NVRAM. A TFTP server is designed to facilitate access and as such, is notoriously easy to hack (see below). As a result, if the password was stored in the configuration file in an unencrypted format, it would not be too difficult for someone to view the file and obtain the password. In addition to the risk of password disclosure associated with using a TFTP server, you also run the

risk of unauthorized modifications being made to the configuration file stored on the TFTP server. TFTP is considered not secure because it doesn't require password authentication. If a host runs the TFTP service without restricting the access in some manner, an attacker can read and write files anywhere on the system. For example, it is possible to obtain the password file from a system running the TFTP service. The steps are as follows:

```
$ tftp anyhost (IP address or alias)
tftp> get/etc/passwd/tmp/passwd.
tftp> quit
```

Generally, it is a very bad idea for any server to run the TFTP daemon. This protocol is an example of an unnecessary service that a computer should not run. Even if the password is securely encrypted, there are programs available to decrypt the Cisco login and enable passwords from a Cisco configuration file or to sniff the password on the network. These programs are easy to find.

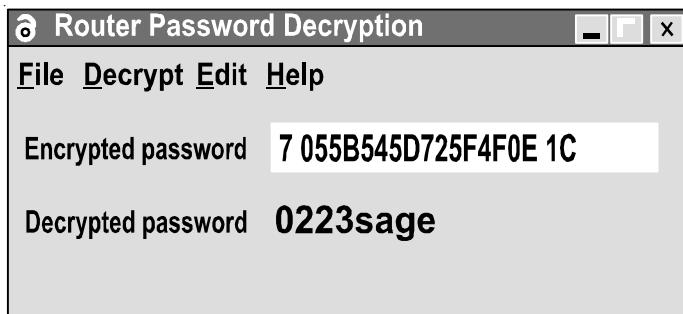


FIGURE 2.9 Solar winds

Router Password Decrypt is a tool from Solar Winds that allows you to reset the enable password for a Cisco router and change any Cisco configuration parameter via SNMP. While this tool has legitimate uses, it can also be used as a tool for hacking.

Cisco Discovery Protocol (CDP) is an example of a protocol that should be disabled on most routers. CDP protocol makes it very easy for hackers to gather information about routers on the network. The CDP protocol broadcasts platform and protocol information to other devices on the network. This information can be useful to any potential hacker. By default, CDP is enabled on a router and its interfaces when IOS is installed. It should be disabled

unless there is a specific purpose for running it. This is not meant to be a lesson on the configuration and commands for Cisco IOS, but is simply offered as an illustration of the similarities between the servers and routers. Servers are normally protected behind firewalls on the internal network, while routers, due to their unique function, are often exposed to the outside world.

2.18 CRYPTOGRAPHY

Cryptography is the science of securing data. It addresses four major concerns: confidentiality, authentication, integrity, and non-repudiation. Encryption is the transformation of data into an unreadable form using an encryption/decryption key. Encryption ensures privacy and confidentiality, keeping information hidden from anyone for whom it is not intended, including those who can see the encrypted data.

2.19 CRYPTOSYSTEMS

A cryptosystem obeys a methodology (procedure). It includes one or more encryption algorithms (mathematical formulae), keys used with the encryption algorithms, a key management system, plaintext (the original text), and ciphertext (the original text that has been obscured).

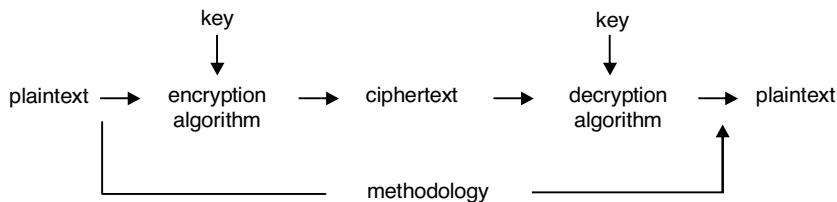


FIGURE 2.10 Cryptosystems

The methodology first applies the encryption algorithm and key to the plaintext to produce ciphertext. The ciphertext is transmitted to a destination where the same algorithm is used to decrypt it to produce the plaintext. The procedure (included in the methodology) to support key creation and distribution is not shown in the diagram.

2.20 KEY-BASED METHODOLOGY

In this methodology, the encryption algorithm combines with a key and plain-text to create ciphertext. The security of a strong key-based system resides with the secrecy of the key used with the encryption algorithm rather than the supposed secrecy of the algorithm. Many encryption algorithms are publicly available and have been well tested (*e.g.*, the Data Encryption Standard). However, the main problem with any key-based methodology is how to create and move the keys securely among the communicating parties. How does one establish a secure channel between the parties prior to transmitting keys? Another problem is authentication. There are two potential areas of concern here: The message is encrypted by whomever holds the key at a given moment. This should be the owner of the key, but if the system has been compromised, it could be a spoof. When the communicating parties receive the keys, how do those parties know that the keys were actually created and sent by the proper authority? There are two types of key-based methodologies: symmetric (private-key) and asymmetric (public-key). Each methodology uses its own procedures, key distribution, types of keys, and encryption/decryption algorithms. The terminology employed in discussing these methodologies can be very confusing.

2.21 SYMMETRIC (PRIVATE) METHODOLOGY

In this methodology, both encryption and decryption operations use the same key with the sender and receiver agreeing on the key before they can communicate. Provided the keys have not been compromised, authentication is implicitly resolved because only the sender has a key capable of encrypting and only the receiver has the same key capable of decrypting. Because the sender and the receiver are the only people who know this symmetric key, if the key is compromised, only these two users' communication is compromised. The problem, which is the same for all types of cryptosystems, is how to distribute the symmetric (private) key securely. Symmetric key encryption algorithms use small-length keys and can quickly encrypt large quantities of data.

The process involved with symmetric key systems is as follows:

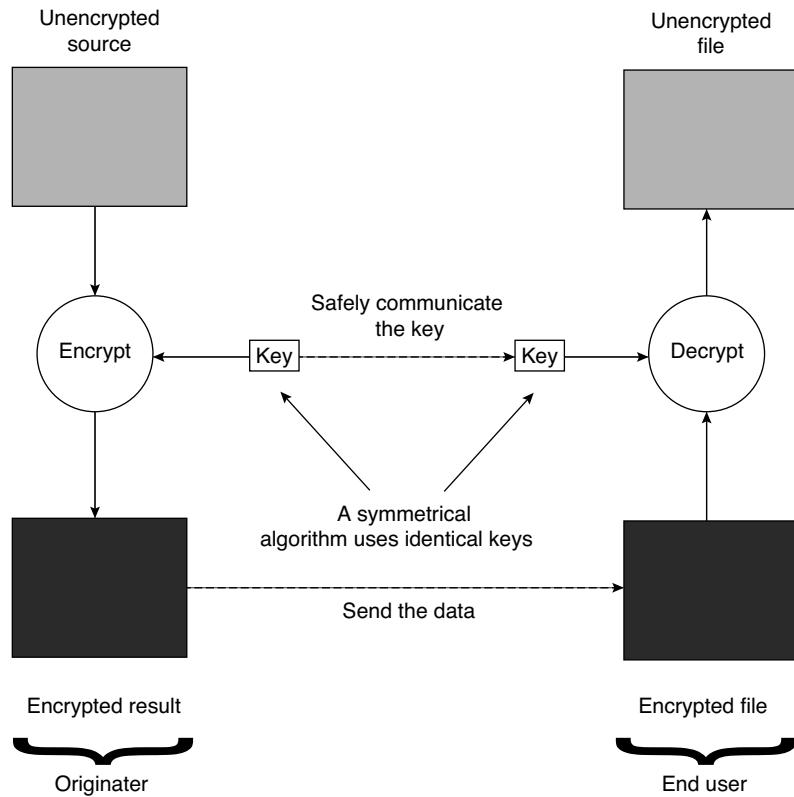


FIGURE 2.11 Symmetric private key to the package to produce the ciphertext

1. Create, distribute, and store the symmetric private key securely.
2. The sender creates a digital signature by hashing the plaintext and attaching the resulting string to the plaintext.
3. The sender applies the fast symmetric encryption/decryption algorithm with the symmetric private key to the package (plaintext and attached digital signature) to produce the ciphertext. Authentication happens inherently because only the sender has the symmetric private key and can encrypt the package. Only the receiver holding the symmetric private key can decrypt this package. The sender transfers the ciphertext. The private symmetric key is never transmitted over the unsecured communication lines.
4. The receiver applies the same symmetric encryption/decryption algorithm with the same symmetric key (which the receiver already has) to the ciphertext to produce the original plaintext and digital signature. This authenticates whoever holds the private key.

5. The receiver detaches the digital signature from the plaintext.
6. The receiver creates a digital signature by hashing the plaintext.
7. The receiver compares the two digital signatures to prove the message integrity (unaltered data). The services available today use symmetric methodologies, which include Kerberos and ATM banking networks. Kerberos was designed to authenticate access to network resources rather than to verify data. It uses a central database that generates and keeps copies of the secret keys of all users. ATM Banking Network (automated teller machines) systems are proprietary and are not for resale, although they use symmetric methodologies.

2.22 ASYMMETRIC (PUBLIC) METHODOLOGY

The encryption and decryption keys are different from each other, although they are produced together. One key is made public; the other key is kept private. While both keys can encrypt and decrypt, data encrypted by one can only be decrypted by the other. All asymmetric cryptosystems are subject to shortcut attacks as well as brute force, and therefore, must use much larger keys than symmetric cryptosystems to provide equivalent levels of security. This immediately impacts the computing cost, although using elliptic curve algorithms may reduce this problem. Bruce Schneier, in his book *Applied Cryptography: Protocols*, compared equivalent key lengths.

Symmetric key length	Public key lengths
56 bits	384 bits
64 bits	512 bits
80 bits	768 bits
112 bits	1792 bits
128 bits	2304 bits

In order to circumvent the slowness of the asymmetric encryption algorithms, a temporary, random, small, symmetric session key is generated for each message, and it is the only partly encrypted by the asymmetric algorithm. The message itself is encrypted using this session key and an

encryption/decryption algorithm. The small session key is then encrypted using the sender's asymmetric private key and encryption/decryption algorithm. This encrypted session key, along with the encrypted message, is then transmitted to the receiver. The receiver uses the same asymmetric algorithm and the sender's asymmetric public key to decrypt the session key, and the recovered plaintext session key is used to finally decrypt the message. It is important in asymmetric cryptosystems that the session and asymmetric keys are comparable in terms of the security they produce. If a short session key is used (*e.g.*, 40-bit DES), it does not matter how large the asymmetric keys are: hackers will attack the session key instead. The asymmetric public keys are susceptible to brute-force attacks partly because it is difficult to change them. Once broken, all current and future communication is compromised, often without anyone knowing. The process involved with asymmetric-key systems is as follows:

1. Create and distribute the asymmetric public and private keys securely. The asymmetric private key is delivered to the owner. The asymmetric public key is stored in an X.500 database and managed by the Certification Authority (CA). Users must implicitly trust the secure creation, distribution, and management of the keys. Further, if the creator is the only person for this transmission (the session key), apply it and the symmetric encryption/decryption algorithm to the plaintext and attached encrypted digital signature. The systems managing the keys are different, so the end user must implicitly trust that the creator of the keys has actually deleted his copies.
2. Create a digital signature by hashing the plaintext. Encrypt the resulting digital signature using the sender's asymmetric private key and attach the resulting string to the plaintext (only the sender has created the digital signature).
3. Create a private symmetric key using ciphertext.
4. The problem of sending the session key to the receiver must now be addressed.
5. Make certain the sender has the Certification Authority's (CA) asymmetric public key. Interception of unencrypted requests for the public key is a common form of attack. There may be a whole hierarchy of certificates attesting to the validity of the CA's public key. X.509 describes different methods for establishing user access to the CA public keys, all of which provide an entry point to spoofers, and show that there is no system that guarantees the identity of the CA.

6. Ask the CA for the receiver's asymmetric public key. This process is vulnerable to the man-in-the-middle attack. The receiver's asymmetric public key has been digitally signed by the CA. This means that the CA has used the CA's asymmetric private key to encrypt the receiver's asymmetric public key. Since only the CA holds the CA's asymmetric private key, then the receiver's asymmetric public key came from the CA.
7. Once received, decrypt the receiver's asymmetric public key using the CA's asymmetric public key and an asymmetric encryption/decryption algorithm. Implicit trust in the CA, that the CA has not been compromised, is required. If the CA is compromised, the entire infrastructure is unusable. Those holding the public key can encrypt, but there is no way of knowing if the key has been compromised. (When you requested the CA's public key, did you actually receive the CA's public key or something else?)
8. Using the receiver's asymmetric public key (now received from the CA and decrypted) and an asymmetric encryption/decryption algorithm, encrypt the session key. Only those holding the receiver's public key can encrypt, but there is no way of knowing if the key has been compromised. Attach the encrypted session key to the ciphertext (which includes the previously encrypted digital signature).
9. Transfer the package (ciphertext that includes the digital signature and the attached encrypted session key). The encrypted session key is transmitted across the unsecured network and is an obvious target for various types of attacks.
10. The receiver detaches the encrypted session key from the ciphertext.
11. The problem of decrypting the session key by the receiver must now be addressed.
12. Make certain the receiver has the CA's asymmetric public key.
13. Using the receiver's asymmetric private key and the same asymmetric encryption/decryption algorithm, the receiver decrypts the session key.
14. The receiver applies the same symmetric encryption/decryption algorithm with the now unencrypted symmetric key (session key) to the ciphertext to produce the plaintext and attached hash or digital signature.
15. The receiver detaches the hash from the plaintext.
16. The receiver asks the CA for the sender's asymmetric public key.

17. Once received, the receiver decrypts the sender's asymmetric public key using the CA's public key and the correct asymmetric encryption/decryption algorithm.
18. Using the sender's asymmetric public key and an asymmetric encryption/decryption algorithm, the receiver decrypts the hash string.
19. Create a digital signature by hashing the plaintext.
20. Compare the two hashes to prove that the data has not been altered.

2.23 KEY DISTRIBUTION

Both types of cryptosystems have a problem distributing the keys. Symmetric methodologies squarely face up to this fact and define how keys are to be moved between the parties before communication can take place. How this is done depends upon the security required. For lower security requirements, sending keys by a delivery mechanism of some kind (such as postal mail or a parcel delivery service) may be adequate. Banks use the postal service to deliver PINs, which are, in essence, easily crackable symmetric keys that may or may not unlock other keys or your money. Very high security requirements may require the hand delivery of keys, possibly in parts by several people. Asymmetric methodologies try to get around the problem by encrypting the symmetric key and attaching it to the encrypted data. They then try to make it possible to distribute the asymmetric keys used to encrypt the symmetric key by employing a CA to store the public asymmetric key. The CA in turn digitally signs the keys with the CA's private asymmetric key. Users of the system must also have a copy of the CA's public key. In theory, this means that the communicating parties do not need to know about each other ahead of secure communication. Proponents of asymmetric cryptosystems maintain that this mechanism proves authenticity and is sufficient. The problem still remains, however. The asymmetric key pair must be created together. Both keys, whether they can be made publicly available or not, must be sent securely to the owner of the key, as well as to the Certification Authority. The only way to do this is by some kind of delivery mechanism for low security requirements, and hand-delivery for high security requirements. The problems of the asymmetric mechanism include the following:

- X.509 assumes that keys are securely distributed and does not address the issue other than identifying it. There are no standards covering this

area. To be safe, keys (whether symmetric or asymmetric) must be hand-delivered. Even then, people could be intimidated or bribed.

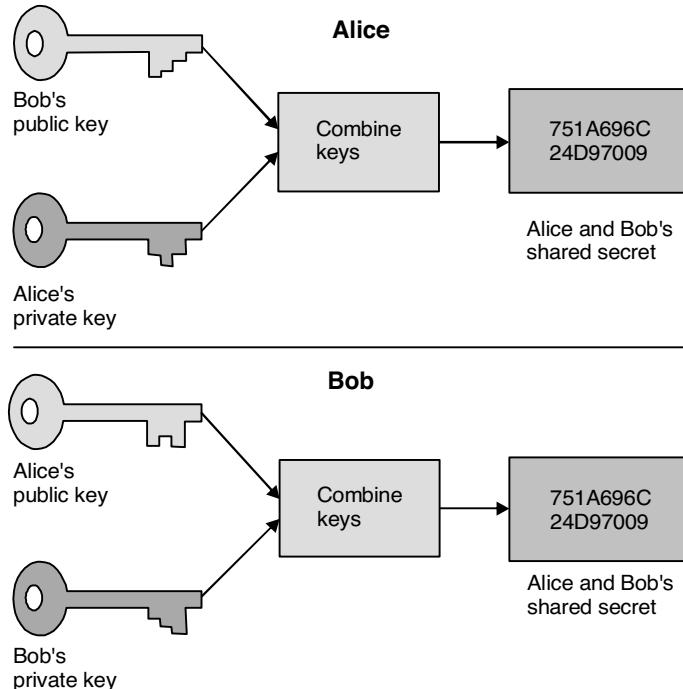


FIGURE 2.12 Key distribution

- There is no mechanism in place to reliably validate what system is actually talking to what system. The man-in-the-middle attack is an attack by a spooper masquerading as the CA and getting the data before it is realized the spooper was actually in the picture. All the spooper has to do is to capture the request to the CA and substitute his own keys in its place. This type of spooper has come and gone long before users become aware that something might be wrong.
- The digital signing by the CA of a key still does not prove the authenticity of the key because the CA's own key could have been compromised. X.509 describes the digital signing of CA keys by higher level CAs, and describes this as a certification path (a hierarchy of CA public keys). X.509 discusses the problems associated with verifying the correctness of a public key, suggesting that it can only operate if there is an unbroken chain of trusted points in the directory between the users required to authenticate. The standards do not offer any mechanism to get around this.

- X.509 assumes the user had prior access to the CA's public key. How this is achieved is not defined in the standards document. Compromise of the Certification Authority is a very real threat. Compromise of the CA means that ALL users of the system are compromised, and no one might ever know. X.509 assumes that storage of all keys, including the CA keys, is secure. The deployment of X.500 directory systems (where X.509 keys are stored) is difficult and prone to misconfiguration. There are very few people with the technical knowledge required to manage these systems properly. Further, it is a well-known fact that people in trusted positions can be subverted, kidnapped, or bribed. The CA may become a bottleneck. To provide for fault tolerance, X.509 suggests that the CA database be replicated or shadowed by using the X.500 standard directory services; this considerably raises the cost of the cryptosystem. If spoofing occurs, it is difficult to identify which system was attacked. Furthermore, all the data must be sent across communication lines somehow when the data is being distributed.
- An X.500 directory system is costly to install, configure, and maintain. Access to this directory is either by using an outside subscription service or by an organization providing its own. The X.509 certificate is based on each individual possessing a unique name. The allocation of names is the responsibility of yet another trusted authority, the naming authority.
- Full keys, even though encrypted, are transmitted across the unsecured communications medium. In spite of these major drawbacks, users must blindly trust the asymmetric cryptosystem. Key management refers to the distribution, authentication, and handling of keys. No matter what kind of cryptosystem is used, keys must be managed. Secure methods of management are very important as many attacks on key-based cryptosystems are aimed at key management procedures.

2.24 ASYMMETRIC ALGORITHMS

Asymmetric algorithms are used by asymmetric cryptosystem methodologies in order to encrypt a symmetric session key (which is actually used to encrypt the data). Two distinct keys are used: one that is publicly available, and the other that is kept private and secret. Usually, both keys perform encryption and decryption functions. However, data encrypted by one can only be decrypted by the companion key.

Type	Description
RSA	Popular asymmetric encryption algorithm, whose security depends on the difficulty in factoring large integers.
ECC (Elliptic Curve Cryptosystem)	Uses the algebraic system defined on the points of an elliptic curve to provide asymmetric cryptographic algorithms. Emerging as competition to other asymmetric algorithms because it offers equivalent security using shorter key lengths and faster performance. Current implementations indicate that these systems are far more efficient than other public key systems. Performance figures show an order of magnitude improvements in efficiency over RSA, Diffie-Hellman and DSA.
ElGamal	Variant of the Diffie-Hellman which can be used for both digital signatures and encryption.

2.25 HASH FUNCTIONS VS. KEY-BASED CRYPTOSYSTEMS

Hash functions are central to key-based cryptosystems. They are relatively easy to compute, but almost impossible to decrypt. A hash function takes a variable size input and returns a fixed size string (sometimes called a Message Digest), usually 128 bits. Hash functions are used to detect modification of a message.

Type	Description
MD2	Slowest, optimized for 8-bit machines
MD4	Fastest, optimized for 32-bit machines. Now broken.
MD5	Most commonly used of the MD functions similar to MD4, but with added security features, making it 33% slower than MD4. Provides data integrity. Considered secure.

(continued)

Type	Description
SHA (Secure Hash Algorithm)	Produces 160-bit hash values from variable-sized input proposed by NIST and adopted by the US Government as a standard Designed for use with the proposed DSS (Digital Signature Standard) and part of the US Government's Capstone project.

EXERCISES

1. What is the public key certificate?
2. Explain the centralized certificate.
3. What is a hash function?
4. Discuss random key generation.
5. Explain biometrics.
6. What is a digital signature? How does it work?
7. Discuss e-mail security.
8. Explain automatic rekeying.

CHAPTER 3

WORLD WIDE WEB TRANSACTION SECURITY

Chapter Goals

- Internet Service Provider (ISP)
- Network Access Point (NAP)
- Routers
- Addressing
- ATM
- Ethernet
- Fiber Distributed Data Interface (FDDI)
- Multi-Protocol Label Switching (MPLS)
- Point-to-Point Protocol (PPP)
- High-level Data Link Control (HDLC)

3.1 INTERNET INFRASTRUCTURE

With the growth of the Internet for personal use (for example, Facebook, Google, and Gmail) and business purposes (such as file storage, web applications, and communication), it is useful to talk about what actually powers its various facets. There are five areas that encompass the Internet's infrastructure.

3.1.1 Internet

In 1969, the Internet started with four interconnected computers in the US and was known as ARPA net, a project funded by the Advanced Research Projects Agency of the US Department of Defense. Today, it is made up of hundreds of millions of hosts and hundreds of thousands of networks all over the world, carrying various kinds of information and services, such as electronic mail, the World Wide Web, and file transfer protocols.

3.1.2 Internet Service Providers (ISPs)

If you want to access the Internet or obtain Internet services, your computer must be part of an *Internet Service Provider (ISP)* network. ISPs are the companies that provide access to the Internet. Residential users may use a modem and a dial-up line to connect to an ISP. Commercial companies or educational institutes also require ISPs to provide connections from their LANs to the Internet.

3.1.3 Point of Presences (POPs)

Most large communications companies have many *Point of Presences (POPs)* in various regions, and these POPs are interconnected via high-speed links. A POP is a service provider's location for connecting users. For example, suppose ISP-A is a large ISP that has a POP in each state of the US and owns its dedicated fiber-optic backbones connecting the POPs. Its customers in the same state should connect to the same POP in that state, and all of ISP-A's customers in the US can talk to each other, even though they are located in different states. However, at this stage, they cannot talk to the customers of another ISP.

3.1.4 Network Access Point (NAP)

To achieve the intercommunication between two ISPs' customers, both of the ISPs have to agree to connect to a common *Network Access Point (NAP)* simultaneously, which is also known as an *Internet Exchange Point (IXP)*. A NAP is a location where ISPs can connect with one another and exchange traffic. NAPs are usually operated by Internet backbone providers. Currently, there are dozens of large ISPs connected through NAPs all over the world, allowing computers on the Internet to talk to each other. NAPs are critical components of the global Internet infrastructure, as the connectivity they provide determines how data traffic is actually routed. For example, there are

a number of ISPs in Hong Kong. Since the Internet is still US-centric, each of the ISPs has its own links to the US. However, in the early days of connectivity in Hong Kong, there was no NAP. Therefore, the traffic between two computers connecting two different ISPs in Hong Kong had to flow through the NAPs in the US. This process consumed a significant amount of Hong Kong's precious international link bandwidth. In 1995, Hong Kong set up its own NAP, called HKIX, which connected the local ISPs together. It allows the exchange of intra-Hong-Kong traffic locally and provides faster and less expensive paths to local sites. There are plenty of significant NAPs in the world. Some of them are as follows:

- US—MAE-West California, MAE-East Washington DC, Chicago NAP, New York NAP, and the NAP of the Americas
- UK—MaNAP, LINX, LoNAP, and ScotIX
- Japan—JPIX, Media Exchange (TTNet), and NSPIXP
- China—TerreNAP and SHIX (ShangHai IX)
- Singapore—SingTel IX
- Hong Kong—HKIX, ReachIX, and Pilhana

3.1.5 Local Area Network (LAN)

A *Local Area Network (LAN)* is a computer network for communication between end computers in a local area such as a company or an institute. By using wide area network (WAN) links, LANs can be connected (*e.g.*, LANs in the POP of an ISP) to form a larger network (*e.g.*, an ISP network) on the Internet.

3.2 NETWORK INFRASTRUCTURE

The Internet infrastructure is essentially a global collection of networks. End computers are connected to a LAN, and LANs are connected to an ISP (a kind of network). Access-level ISPs are usually connected through national and international ISPs that are connected at the NAPs (another kind of network) operated by Internet backbone providers (see Figure 3.1).

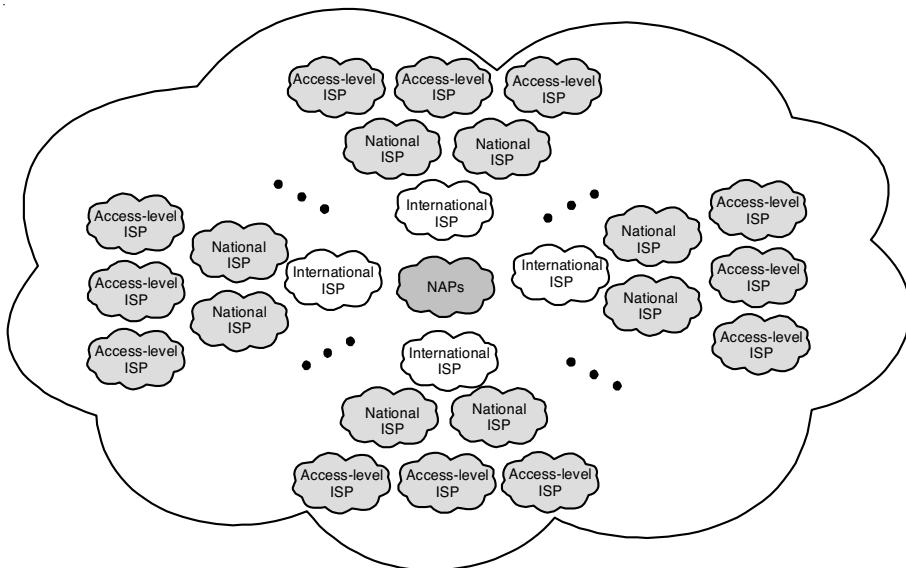


FIGURE 3.1 The Internet is essentially a global collection of networks

3.3 BASIC ISSUES IN SECRET KEY MANAGEMENT

Each ISP and NAP is essentially a network of routers and communications links. Since the Internet infrastructure is made up of ISPs and NAPs, it can be said that the Internet Infrastructure is made up of links and routers. However, to implement the host-to-host communication on the Internet, in addition to these physical components, it also requires an addressing scheme and a naming system. That is, the hosts on the Internet conform to certain naming and address conventions.

3.3.1 Links

The links on the Internet are made up of different types of physical media, ranging from copper wire and coaxial cable to optical fiber and the radio spectrum. Different types of media transmit data at different rates, and the rates are typically measured in bits per second (bps). The final leg of delivering connectivity from an ISP to a customer (which can be a residential user or a company's LAN) occurs over what is called the *last mile*. It is actually about 2–3 miles and includes

- Integrated Service Digital Network (ISDN)
- Digital Subscriber Line (DSL), *e.g.*, ADSL, HDSL, and VDSL
- Cable and the cable modems
- Leased lines, *e.g.*, T1 and T3
- Wireless connections, *e.g.*, 802.11, 802.20, and Wi MAX

As the Internet backbones are the points of most Internet congestion, they are typically made up of fiber optic trunk lines that transmit data at extremely high rates. The trunk line uses multiple fiber optics in parallel to increase the link speed. Optical Carrier (OC) levels are used to specify the speed of fiber optic networks (or example, OC-1 = 51.85 Mbps and OC-3 = 155.5 Mbps).

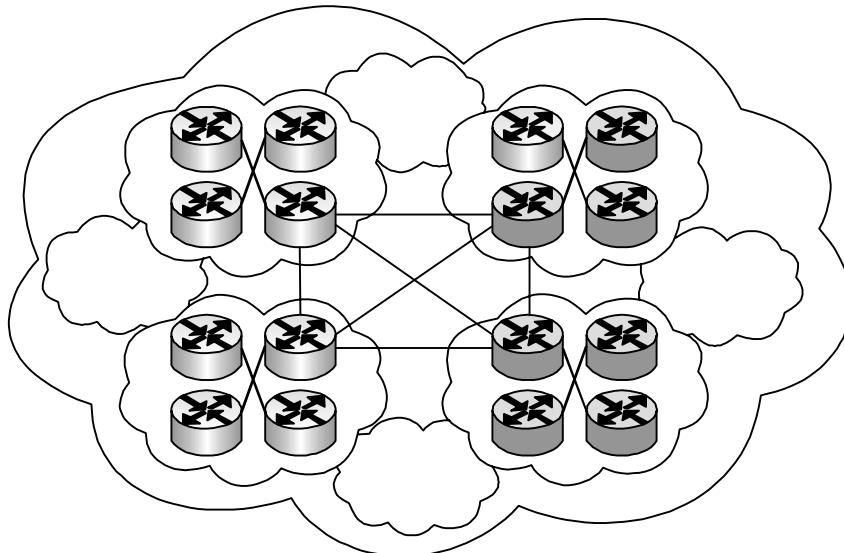


FIGURE 3.2 The Internet is essentially a network of routers and communication links

3.3.2 Routers

Networks on the Internet are not usually directly connected. Instead, they are indirectly connected through many intermediate network devices known as routers. A *router* is a special-purpose dedicated computer that attaches to two or more links (networks). When it receives a packet from one of its incoming links, it makes a routing decision, and then forwards that packet to one of its outgoing links. The decision is usually made based on the current state of the

networks the router is connected to. No matter how many networks a router is connected to, its basic operation remains the same. To make the selection of the next “hop” efficient, each router uses a routing table to keep track of routes to a particular network destination. A simple routing table looks like this.

Table 3.1 Routing Table.

Destination	Outgoing link
Network 1	Serial line 1
Network 2	Serial line 2
:	:
Network N	Serial line 1

For example, if the router with the above routing table receives a packet destined for Network 2, it will forward that packet to its attached serial line number 2. Routing tables are built according to the routing algorithm that the routers in the network use. The router forwards it to one of its nearby routers, which in turn forwards that packet to one of its nearby routers. After a series of links and routers, the destination can be reached. The computers at the end are not usually directly connected to routers. To form a local area network, switches are commonly used to interconnect end computers. Switches operate at the data link layer (of the Open Systems Interconnection (OSI) reference model), and split up networks into smaller individual collision domains. When a switch receives a frame, it first reads the destination data-link address from the header information in the frame, then establishes a temporary circuit between the source and destination switch ports, and finally sends that frame on its way.

3.4 ADDRESSING

On the Internet, every participating machine is identified by an *Internet Protocol (IP) address*, which is a unique 32-bit binary number. To make it easier to remember, IP addresses are normally expressed as a string of four decimal octets separated by periods, ranging from 0.0.0.0 to 255.255.255.255, with some reserved values for specific purposes. Therefore, the IP address of 00001011 00010110 00100001 00101100 can be written as 11.22.33.44.

Internet addresses are not only used to identify a host, but also to specify routing information on the Internet. Data packets traverse the Internet by following a path from their source through a number of routers to the final destination. The data packets are called *IP packets* or *datagrams*, which is the basic unit of transmission across the Internet and contains both the source and destination IP addresses. Upon receiving a datagram, based on the destination address, a router determines the next hop to which the datagram should be sent. Since IP addresses exhibit a hierarchical structure, they can be used to make routing decisions. Each 32-bit IP address is divided into two parts: the *network ID* and *host ID*. The addresses of the hosts in the same network should have the same network ID but different host IDs. The IP defines three classes of networks: classes A, B, and C. Their network IDs are 8, 16, and 24 bits long, respectively. In classful IP addressing, the network portion can take only these three predefined number of bits. A better practice is to use classless addressing. In classless addressing, any number of bits can be assigned to the network ID. To determine the length of the network ID, the use of a *subnet mask* is needed. The subnet mask is a kind of bit mask containing a number of ones starting from the left side, which can be expressed by the slash form or the decimal-octets-periods form. For example, if the network ID is 24 bits long, the subnet mask can be expressed by “/24” or “255.255.255.0.” By performing a bit-wise AND on the IP address and the subnet mask, the corresponding network ID can be obtained. Since each network on the Internet has a unique network ID, routers can use the network ID to make routing decisions. Upon receiving a packet, a router first identifies the network ID that the packet targeted on. After that, it checks its routing table to see where to forward that packet. Version 6, or IPv6, expands the IP address to 128 bits long, providing a theoretical address space of 340,282,366,920,938,463,463, 374,607,431,768,211,456, which surely solves the IP address shortage problem. However, due to the difficulty of deploying IPv6, the majority of today's routers are routing IPv4 packets. The specification of IPv6 can be seen in RFC 2460.

3.5 SYSTEM SECURITY

As IP addresses are in numeric form, they are difficult for humans to remember. Therefore, in addition to an IP address, we can also assign a symbolic name to a machine on the Internet. The symbolic name consists of a series of alpha-numeric text separated by periods. For example, the machine with the

IP address 11.22.33.44 can be assigned the name www.mysite.com. However, although users prefer to the more mnemonic symbolic names, the underlying network protocols and routers operate based on IP addresses that are of a fixed-length and hierarchically structured. Thus, the application software (e.g., Web browsers and email clients) in the sending machine, which allows users to enter the symbolic name, is responsible for translating the name into an equivalent IP address of the destination, and assigning the IP address in binary form in IP packets. The translation process requires a directory service that maps the symbolic names to the IP addresses. It is the main task of the Internet's Domain Name System (DNS). The DNS is a distributed database implemented with many servers located all over the world. The servers are called *name servers* or *DNS servers*. Each of them only maintains part of the database and none of them has a complete copy.

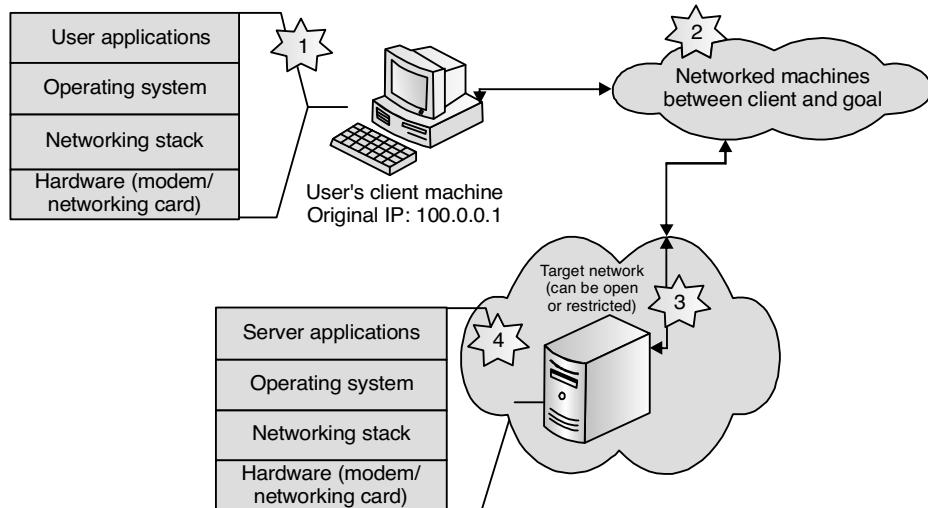


FIGURE 3.3 System security

More specifically, a name server only holds the name-to-address mappings of the machines under its management. When a name server receives a query for a symbolic name that is out of its database, the name server is responsible for asking the corresponding name server that maintains that symbolic name. In this way, the DNS allows for a decentralized administration. The decentralization of the administration makes the task of keeping mappings up-to-date much easier, and more importantly, improves the DNS scale. The DNS provides the infrastructure for translating domain names into their equivalent IP addresses for application software on the Internet. The naming system is

a critical operational part of the Internet infrastructure. Without it, the Internet would shut down very quickly. In spite of its importance, the DNS has no security mechanisms. The DNSSEC (short for *DNS Security Extensions*) adds a set of security extensions to the DNS to provide authenticity and integrity. The extensions are mostly based on the use of the cryptographic digital signature.

3.6 BASIC ISSUES IN INTERNET TRANSACTION SECURITY

Internet infrastructure security focuses on the protection of the key infrastructure components, such as links, routers, DNS servers, and naming systems. Since the Internet was assumed to work in a trustworthy environment in the beginning, it was designed without security risks in mind. As a result, the infrastructure is vulnerable to a variety of security threats and attacks, which can lead to various kinds of network problems. Let's consider an example to illustrate the problem caused by a packet mistreatment attack. Assume all links have a unit cost (e.g., in terms of the bandwidth), except the link connecting routers R1 and R3 together. When the packets sourced from R1 are targeted on R4, the shortest path in terms of cost should be R1-R2-R3-R4. This is the expected path in normal packet forwarding. However, if R3 is compromised to mishandle the packets by forwarding them back to R1 maliciously (not to R4), a routing loop occurs. In this case, the packets will circulate among R1, R2, and R3 until their time-to-live period expires. This loop prevents the packets from reaching their destination, R4. This also causes extra router load and network traffic. The problem becomes even more intractable if the malicious router only misroutes packets selectively (e.g., only for selected networks or hosts at random time intervals), or if there is a larger number of routers involved in the routing loop.

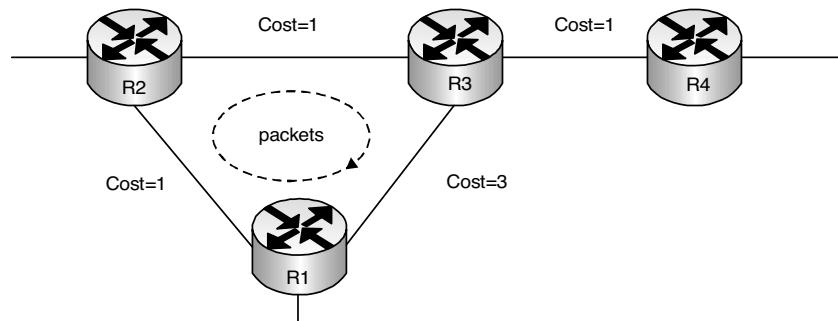


FIGURE 3.4 An illustration of triangle routing

3.7 NETWORK INFORMATION AND NETWORK INFRASTRUCTURE SECURITIES

In the past decades, the research on Internet security has mainly focused on information assurance, which is based on the principle of confidentiality and integrity. *Confidentiality* is the property whereby data is accessible to only legitimate receivers and is not disclosed to unauthorized persons, whereas *integrity* is the property that the data received is sent from legitimate senders and has not been altered in an unauthorized manner. The techniques commonly used to achieve these goals are based on encryption/decryption, digital signatures, and a message authentication code. However, although information assurance is important, it becomes meaningless if the data, no matter how secure its content is, cannot be delivered through the Internet infrastructure to the targeted destination. If the infrastructure is attacked, a large area (or even all) of the Internet will be affected, causing economic damages. Therefore, securing the Internet infrastructure is of significant importance. Unlike network information security that focuses on the information protection, infrastructure security focuses on the protection of the network infrastructure itself, that is, it focuses on how to detect network attacks and how to prevent routers or other network devices from being compromised. Security includes the use of secure Internet protocols, traffic monitoring, and firewalls.

3.8 IMPORTANCE OF NETWORK INFRASTRUCTURE SECURITY

One of the reasons of why network infrastructure security is important and has drawn much research interest is that attacks to the infrastructure affect a large portion of the Internet and create service disruptions. Since many daily operations highly depend on the availability and reliability of the Internet, the security of its infrastructure is a high priority issue. Let us examine a scenario in which enormous destruction is caused by attacks on the Internet infrastructure. In this scenario, the links have the unity cost; all are fairly heavily loaded, but still under capacity. The attacker compromises router A and sets the cost of link B to a very higher value. Since Internet traffic is routed along the path with the least cost, packets will be routed around link B (because of its high cost). As a result, packets will be routed through router C. This makes router C receive more packets than it can handle. Since router C is the border router of domain Z, its overloaded condition causes congestion at domain Z. Hence, the services to clients of domain Z (also W, X, and Y) will noticeably

slow down. Another reason to secure the network infrastructure is the growing fear of cyber terrorism. As can be seen the example shown in Figure 3.4, simply increasing the routing cost of a link (accidentally or maliciously) can affect a large portion of the network. Therefore, if terrorists manage to attack the core network of a country, it would imply that they had already attacked the country's economy and caused ruinous financial damage because today's business operations highly rely on the availability of the cyber network.

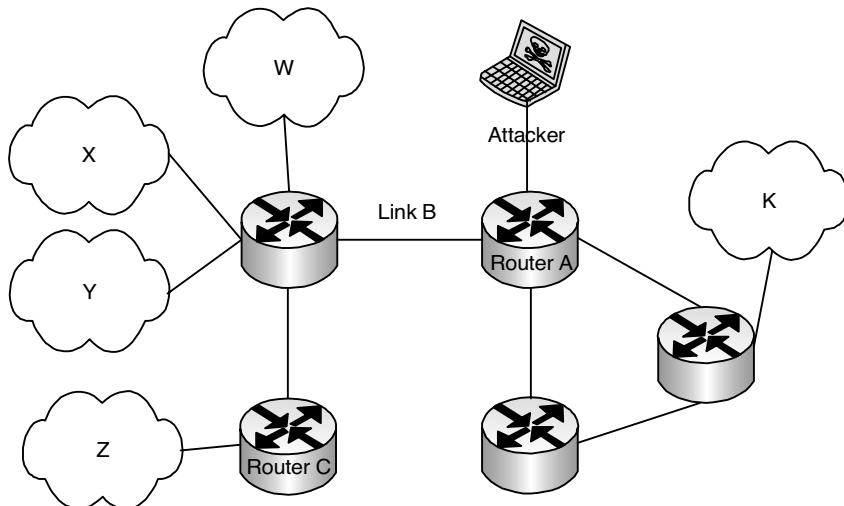


FIGURE 3.5 The attacker increases the cost of link B so that traffic from domains W, X, and Y to domain K takes the suboptimal path, causing a denial of service

3.9 INTERNET INFRASTRUCTURE VULNERABILITY

Many network devices and protocols were designed without security concerns in mind at the beginning. Unfortunately, those protocols now form part of the Internet infrastructure, making it vulnerable to various kinds of attacks. The original protocols were designed with the assumption that the Internet was a completely trustworthy environment.

3.9.1 Solutions Usually Require Large Scale Modifications

As mentioned above, as the design of network infrastructure is fundamentally insecure, new security solutions usually require a certain level of modification of the existing network devices, such as firmware updates or even device replacement. Thus, the costs and effort expended on large-scale deployments

can be very high, which makes some service providers take a conservative view of new infrastructure security solutions.

3.9.2 Security and Performance Tradeoffs

Security usually requires extra processes to run (or more CPU cycles to execute) the security process. It reduces the performance of current devices. For example, routers can activate the Access Control List (ACL) to perform packet access control. However, the activation of the ACL may reduce their routing performance, resulting in a lower packet throughput. Therefore, the ACL is not recommended for use in the routers of core networks.

3.9.3 Security is Only as Strong as the Weakest Link

Since the Internet is heterogeneous and made up of various kinds of networks, the overall security level highly depends on the weakest link or computer on the networks. Though there are advanced security technologies, if any one of the nodes in the network does not (or fails to) employ them, the security of the network is not guaranteed. Besides, though some nodes employ self-defense technologies, which makes them less vulnerable, they may still believe some malicious messages (as they look legitimate) from the compromised node. There is no central authority or organization to ensure the security level of each network on the Internet. Therefore, it is difficult to produce a reasonable quality of security on the Internet.

3.9.4 Attacks Can Be Easily Launched and Are Difficult to Trace

Because of the inherent openness of the Internet, anyone with a computer and an Internet connection can reach any point on the Internet, making it possible to launch attacks from anywhere in the world. Compounding the problem, a computer can easily pretend it is a switch or router (by running a special software package) and broadcast malicious information to mislead the real network devices to perform abnormally.

3.10 NETWORK INFRASTRUCTURE SECURITY—SWITCHING

Layer 2 of the seven-layer OSI model is the data link layer, which is on top of the physical layer (layer 1). Since the physical layer is only concerned with the transmission of a raw bit stream over the physical medium, the data link layer attempts to provide reliable data transfer across the physical link. The services

provided by the link layer include the flow control, acknowledgment, error recovery, and maintenance (activation and deactivation) of the link. Data link technologies include



FIGURE 3.6 Network infrastructure security—switching

- ATM
- Ethernet
- Fiber Distributed Data Interface (FDDI)
- Multi-Protocol Label Switching (MPLS)
- Point-to-Point Protocol (PPP)
- High-Level Data Link Control (HDLC)

The data link layer is generally independent of the network layer (layer 3) and deals with layer 2 addresses only. Therefore, a layer 2 network can transfer traffic using different kinds of layer 3 protocols, such as IPX and IP. There are two types of link-layer channels: broadcast and point-to-point. When two network nodes are connected with a point-to-point link (*e.g.*, a serial link between two routers, such as a telephone line between a residential dial-up modem and an ISP router), a data link layer protocol is also needed to coordinate the traffic within the link. Examples of data link protocols for point-to-point connections are PPP and HDLC. This coordination is simpler than that with a broadcast channel. Hosts are connected to the same communication channel with a data link layer using a broadcast channel. Thus, the Media Access Control (MAC) protocol is needed to coordinate the transmissions

to determine who is allowed to access the media at any one time. Ethernet is a typical example. An Ethernet switch is the most commonly used layer 2 device in today's local area networks.

3.11 SWITCH SECURITY IS IMPORTANT

The OSI and other models commonly use the layered approach. In the layered model, the functions of a communication protocol are divided into a series of layers. Each layer performs a subset of the functions. Each layer provides services to its next upper layer and requires services from its next lower layer, whereas the operation of one layer is independent of the other layers. The advantage of layer independence is that it enables interoperability and interconnectivity. However, such kind of independence also causes security challenges because if any layer is compromised or attacked, other layers will not be aware of this. Since the data link layer is at the bottom of a layered protocol model, its upper layers (namely, the network, transport, and applications layers) rely on it to provide the reliable data transfer across the physical link. If it is compromised, the entire communication session is compromised.

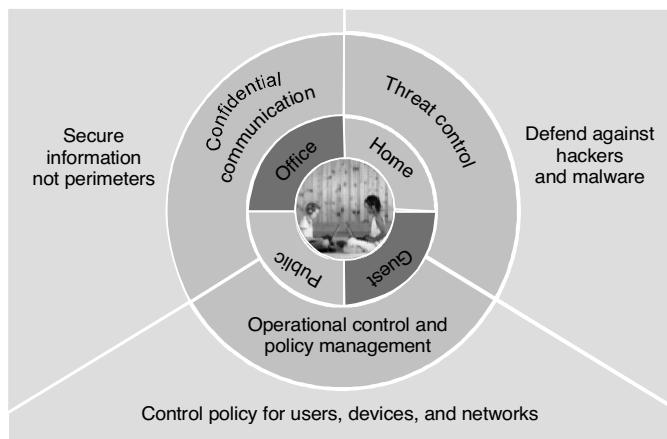


FIGURE 3.7 Switch security is important

In WAN environments, the links are more secure, though they are still subject to attacks, because the links are usually private or operated by trusted telecommunication links that are maintained in a more secure way by guards and a security room. LAN environments, however, are less secure. The major reason is that the layer 2 devices, such as Ethernet switches, are directly accessible to end computers. With the many hacking tools available, a hacker

can generate malicious advertisements or control messages to compromise the layer 2 devices in the network. For example, a Linux machine with the BRIDGE-UTILS package installed can generate Bridge Protocol Data Units (BPDU) frames, which are used to communicate among switches. Since the machine “speaks” the “switch language,” its directly connected switch will believe that it is a switch. As a result, by fitting the right BPDU content, the Linux machine can easily disrupt the structure of the existing switched network. Since an Ethernet switch is the most commonly used layer 2 device in today’s LANs, its security is crucial. Unfortunately, the security of switches is commonly overlooked. Since network administrators usually believe the layer 2 networks are trusted, it is uncommon for them to monitor or examine the operation of layer 2 infrastructures unless there is a connectivity problem. Security measures usually focus on layer 3 and above. For example, intrusion detection systems typically examine the layer 3 and 4 packets, in particular, TCP/IP packets, to detect any suspicious activities. Switch attacks are difficult to discover. One of the reasons is that detecting such attacks requires a good knowledge of the data link protocol operations. For example, to exploit the spanning tree protocol (STP), one has to first understand the logic of a switch port (it can be a *root*, *designated*, or *blocked* port) and their possible states (such as *forwarding*, *listening*, and *learning*) in a STP network. The operation of STP will be detailed later.

3.12 HOW SWITCHES CAN BE ATTACKED

There are various kinds of switch attacks. Many attacks can be initiated from outside of a switch, making it easy to launch for those with the LAN access. Although taking full control of a switch is difficult, if an attacker manages to do it, the network structure can be instantly changed. For example, the attacker could change some configuration values in the compromised switch to disrupt the current switching structure. The attacker could change the manufacturer’s switch operating system with its own modified version. In this case, the switch no longer functions as if it is produced by the manufacturer, but functions in the way the attacker wants, such as constantly returning important network information to the attacker’s machine.

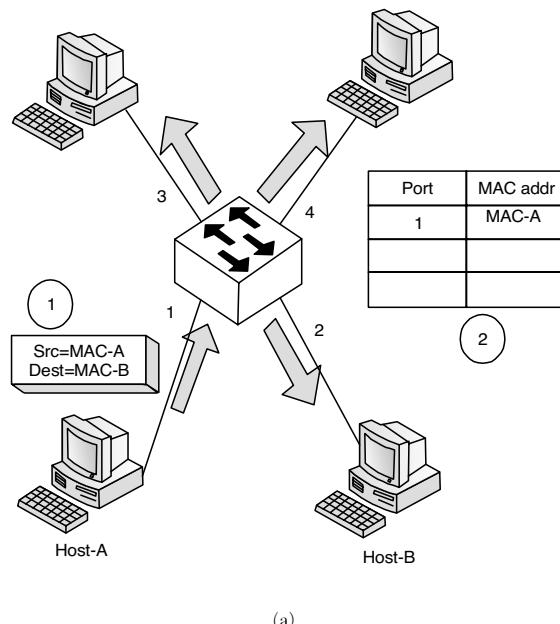
3.12.1 MAC Flooding

Unlike hubs that operate under a broadcast model, switches operate under a virtual circuit model. That is, a switch is capable of determining the destination MAC address in a frame and selectively forwarding the frame to the

correct outgoing switch port. In this case, one can capture, using a sniffer program, the network traffic between two other hosts. MAC flooding attacks make a switch revert to broadcast mode, acting like a hub, so that sniffing can be performed easily.

3.12.2 Content Addressable Memory Table

To determine which frames go to which ports, switches use a *Content Addressable Memory (CAM) table* to store the switching information. The table is built by extracting the source MAC address from the frame transmitted on each port. For example, when the switch in Figure 3.8 receives a frame targeted on Host-B with the MAC address MAC-B from Host-A with the MAC address MAC-A, it will record in CAM that the source MAC address is from port 1. At this point, because the CAM table does not contain an entry for MAC-B, the switch will broadcast this frame to all outgoing ports. When Host-B receives the frame and replies to Host-A, the switch learns that the MAC-B is from port 2. Once the mapping is learned and recorded in the CAM, all future frames destined for MAC-B will be forwarded to the outgoing port 2 only, not other ports. As switches only forward traffic to the destination port, the hosts other than Host-A and Host-B cannot see the traffic between them.



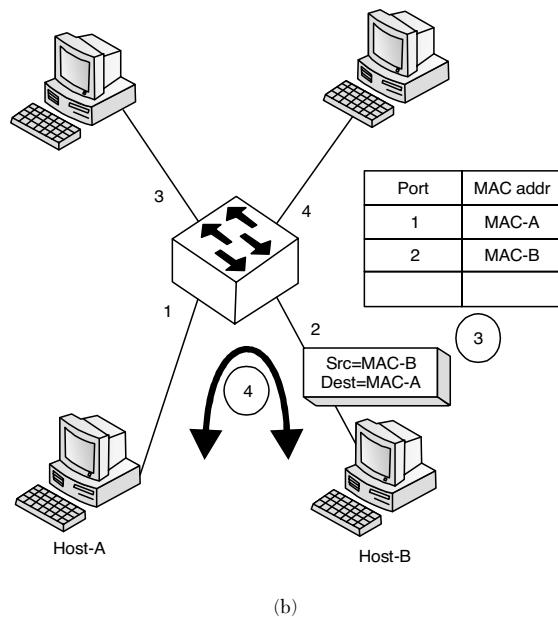


FIGURE 3.8 The CAM table is built by extracting the source MAC address on frames

3.12.3 MAC Flooding Attacks

When there is a frame with a source MAC address that the switch has never seen, a new entry for that new address is added to the CAM table. However, the size of the table is limited. When the table overflows, the switch does not rely on it and broadcasts all the frames to all outgoing ports, like a hub does. The MAC flooding attack forces the switch to act like a hub by making the CAM table overflow. To do that, the attacker just needs to generate (flood) frames with different bogus source MAC addresses. When a sufficient number of frames have been generated, the CAM table overflows, and the attacker can then see traffic he wouldn't ordinarily see. Unfortunately, there are tools, such as `mascot` and `d sniff`, that can generate 155,000 MAC entries on a switch per minute, making this attacking process even more trivial, even with large CAM tables and high-end switches.

3.12.4 Mitigation

The common way to mitigate the MAC flooding attack is to limit the number of MAC addresses that can be connected to a switch port. This feature is called *port security*, which is found on high-end switches. It allows the administrator

to specify the number of MAC address that can be connected to a switch port, or to specify which MAC addresses can access a particular port. In the event of a security violation, the corresponding port can be shut down, disabled for a certain period, or perform other actions configured by the administrator.

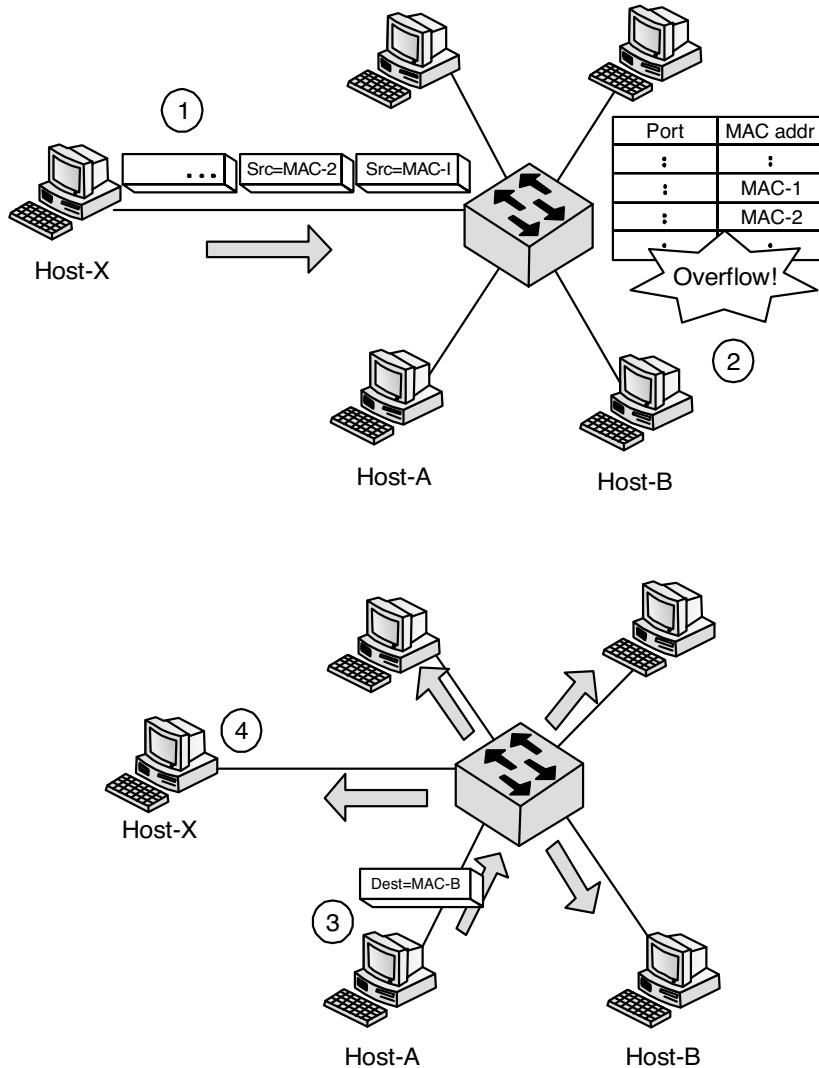


FIGURE 3.9 Illustration of a MAC flooding attack

3.12.5 ARP Spoofing

By exploiting the interaction of the Address Resolution Protocol (ARP), an attacker is able to poison the ARP cache with a forged IP-MAC mapping, and hence cause traffic to be redirected from the correct target to a target of the attacker's choice. This process is called *ARP spoofing*. After performing ARP spoofing, one can launch various attacks, such as man-in-the-middle, DoS, and broadcasting.

3.13 ARP

On an Ethernet/IP network, when a host wants to send an IP packet to another host, only knowing the destination IP is not enough. It also needs to know the MAC address of the destination host so that Ethernet frames can be built and transferred in the network. Therefore, a process is required to find the MAC address of a host given its IP. The *Address Resolution Protocol (ARP)* is used for this purpose. Figure 3.6 shows an example of an ARP operation in which the ARP packet content is also shown. In the example, Host-A wants to communicate with Host-B. The corresponding ARP operations are as follows:

1. Host-A sends out an ARP request packet by broadcasting, asking, "If you have IP-B, please send your MAC to me."
2. Only the machine with the specified IP address, IP-B, will send back an ARP reply packet in unicast directly to Host-A, saying that "I am Host-B, and my MAC is MAC-B." At this point, Host-A is able to build an Ethernet frame with Host-B by filling the destination MAC address field with MAC-B. To reduce the number of ARP requests, operating systems normally cache resolved addresses for a short period of time. Before making an ARP request, a host will check the ARP cache to see if the required IP-MAC mapping exists. If the mapping exists, it will get the corresponding MAC address from the cache without the need to make a new ARP request. Windows provides the command to view, add, or delete entries in the ARP cache. Figure 3.9 shows an example of the output where we can see that the ARP cache contains four IP-MAC mappings. Note that whenever a host receives an ARP reply, it will update its ARP cache with the newly received IP-MAC mapping. It can be exploited by attackers to poison the ARP cache.

Ethernet	
Destination:	FF : FF : FF : FF : FF : FF
Source:	MAC-A
Type:	ARP
ARP	
Opcode	Request
Sender MAC	MAC-A
Sender IP	IP-A
Target MAC	0 : 0 : 0 : 0 : 0 : 0
Target IP	IP-B

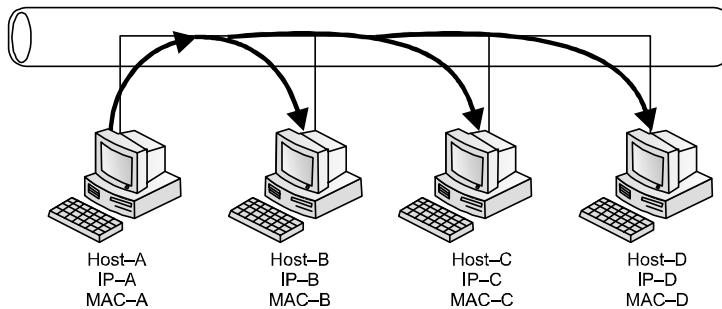


FIGURE 3.10 The ARP request packet broadcasted by Host-A

Ethernet	
Destination:	MAC-A
Source:	MAC-B
Type:	ARP
ARP	
Opcode	Reply
Sender MAC	MAC-B
Sender IP	IP-B
Target MAC	MAC-A
Target IP	IP-A

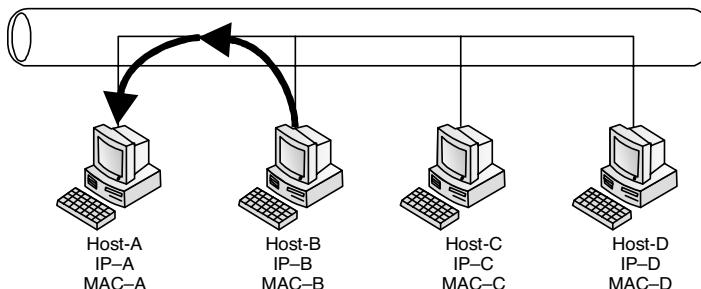


FIGURE 3.11 The ARP reply packet sent from Host-B

```
C:\>arp -a

Interface: 202.175.25.68 on Interface 0x10000003
  Internet Address      Physical Address      Type
  202.175.25.2          00-d0-95-86-65-b5    dynamic
  202.175.25.12         00-40-a5-b1-d8-f2    dynamic
  202.175.25.13         00-02-a5-29-07-34    dynamic

Interface: 192.168.0.1 on Interface 0x10000004
  Internet Address      Physical Address      Type
  192.168.0.40          00-0d-93-59-24-e8    dynamic

C:\>
```

FIGURE 3.12 The ARP command in windows

3.14 THE ARP POISONING PROCESS

Since ARP does not require authentication, whenever a computer receives an ARP reply, it updates its cache regardless of whether it has sent out an ARP request. Therefore, one can easily poison other hosts by sending out an ARP reply packet with the wrong information. Figure 3.9 shows an example where Host-X is attacking Host-A and Host-B.

1. Before the attack, the ARP caches of Host-A and Host-B containing the correct mappings.
2. To poison the ARP cache of Host-A, Host-X sends Host-A an ARP reply packet with the wrong mapping: IP-B-to-MAC-X. Similarly, Host-X sends to Host-B an ARP reply with the wrong mapping: IP-A-to-MAC-X. Since ARP does not require authentication, Host-A and Host-B believe that the mappings are valid and update their ARP tables as usual. Now, their ARP caches are poisoned.
3. As a result, the traffic between Host-A and Host-B will be redirected to flow through Host-X first, instead of directly to each other. In this case, Host-X can intercept the traffic between them, even though they are connected using a switch. Note that the distance is not important for ARP

poisoning. Even though the two hosts are in different buildings, as long as they are in the same broadcast domain, one in the same domain can poison their ARP caches using the method described above. The attacker can perform the following attack with its cache poisoning.

3.15 MAN-IN-THE-MIDDLE ATTACK

If Host-X re-routes packets to the correct destination in both directions, Host-A and Host-B will not be aware that the traffic between them is being monitored (or modified) by Host-X.

3.15.1 DoS Attack

If Host-X chooses not to re-route packets, it can launch a Denial of Service (DoS) attack because Host-A and Host-B cannot communicate with each other anymore. However, since no traffic happens to Host-A and Host-B, the corresponding entries in the ARP caches will timeout. To keep denying service, Host-X has to continue poisoning their ARP caches (*i.e.*, repeat the above process) regularly.

3.15.2 Hijacking

In another case, Host-X could hijack the connection between Host-A and Host-B. Suppose Host-B is a server. When Host-X receives packets from Host-A, it then injects its own packets into Host-A, pretending to be Host-B, the server. Now, Host-X takes control of the connection, and both Host-A and Host-B are not aware of this.

3.15.3 Spoofing WAN Traffic

In a typical LAN, there is a default gateway (router) connecting the local hosts to the Internet. To reach the hosts on the Internet, the hosts in the LAN send packets to its default gateway (destination MAC = the gateway's MAC). That gateway then routes the packets to the next hop, which then routes to its next hop. This process goes on until the final destination is reached. When packets come back from the Internet, the default gateway sends them to the correct hosts in the LAN (by setting the correct destination MAC). Imagine that one of the two hosts in Figure 3.9 is the default gateway. By performing

the same attack, the attacker is able to sniff all of the host's outgoing traffic to the Internet, which would include the Website login name and password.

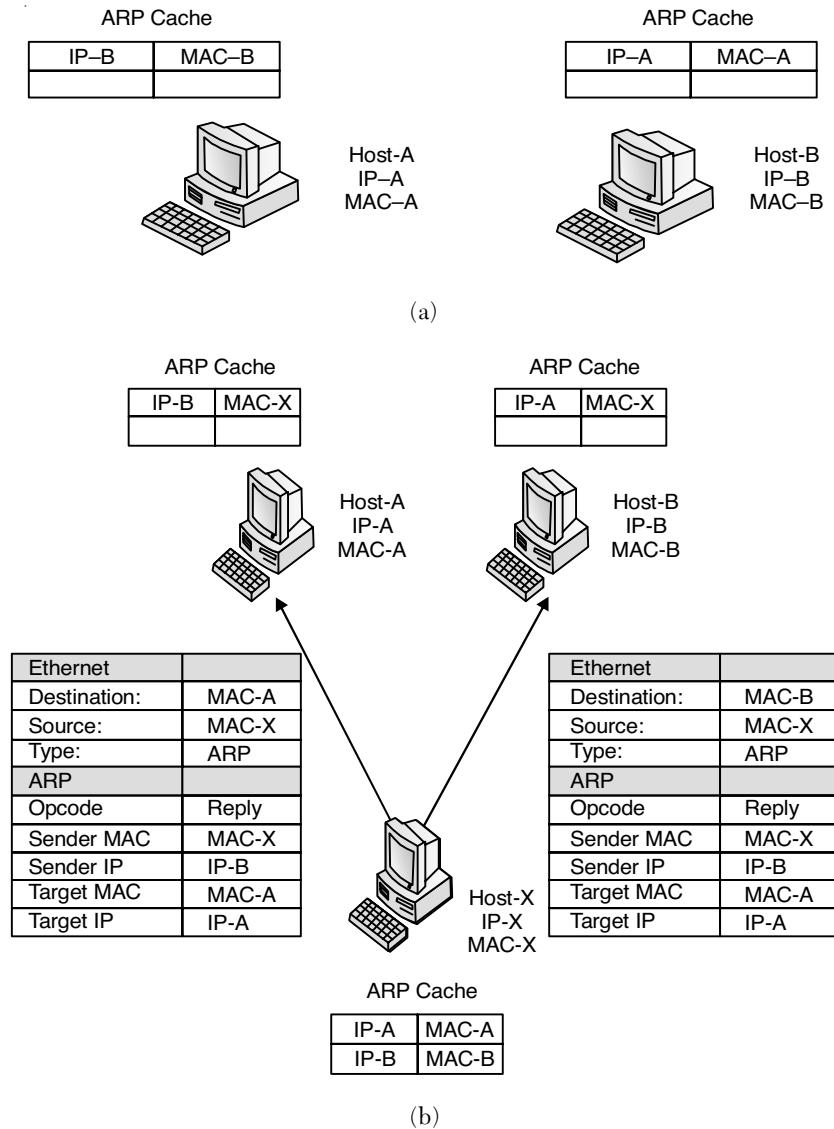


FIGURE 3.13 A normal ARP cache

Limitations

There are a number of limitations to an ARP attack.

1. ARP poisoning only works in the same broadcast domain, and therefore, it cannot redirect traffic between hosts on different subnets or VLANs that separate the broadcast domain.
2. To sniff the traffic between two hosts, the attacker must be able to reroute traffic to the correct destinations in both directions. To reroute the traffic, the attacker also needs to know the IP-MAC mappings of the hosts before the poisoning process.
3. ARP poisoning only updates an existing record in the ARP cache. That is, the record must be already present before the poisoning process.

3.16 STATIC ARP ENTRIES

To prevent spoofing, the ARP cache can store static IP-MAC mappings. As the entries are unchanging, any spoofed ARP replies will be ignored. This concept works; however, requiring the ARP cache to always keep up-to-date IP-MAC mappings of every computer on the network makes this solution impractical for most networks.

3.16.1 Detection

To detect possible ARP spoofing-based attack, the IP-MAC mappings for the machines on the network can be monitored. There are programs, such as ARP Watch (<http://www.securityfocus.com/tools/142>), to do that. When any suspicious changes of the mappings occur, the program will notify the administrator to investigate.

3.16.2 No Cache Update

A simple and practical solution is to only accept ARP replies and update the entries in the cache when they are expired (after the timeout period). In this case, in order to poison the ARP cache, the attacker has to send an ARP reply packet faster than the legitimate host can, making the spoofing process difficult.

3.17 STP ATTACKS

Switches have been widely used to connect computers in LANs. Multiple switches and links are used for redundancy. In this case, a physical loop may exist in the switched network. The *Spanning Tree Protocol (STP)*, IEEE 802.1d, is used to create and maintain a loop-free topology in the switched network. The loop-free topology, or spanning-tree topology, is automatically reconfigured when there are switch failures, link failures, or physical topology changes. Note that since switches are considered multi-port bridges, and STP uses the term “bridge” in the protocol specification, we mainly use the term “bridge” instead of the term “switch” in this section. The following section discusses the basics of STP, which helps explain the attacks discussed later this section. (The Rapid Spanning Tree Protocol, 802.1w, the enhanced version of STP, suffers similar security attacks as those that afflict the STP, but this is not discussed in this section). Bridge Protocol Data Unit (BPDU) Type bridges regularly exchange BPDUs, which contain configuration information, between neighbors. There are two types of BPDUs: Configuration BPDUs (Type value = 0) and Topology Change Notification (TCN) BPDUs (Type value = 0x80). In the Configuration BPDUs, there are two flags.

3.18 TOPOLOGY CHANGE (BIT 1)

TOPOLOGY CHANGE ACKNOWLEDGEMENT (BIT 8)

The Topology Change BPDU is similar to the Configuration BPDU except that no data is transmitted after the Type field. The functions of the above BPDU types are discussed later in this section.

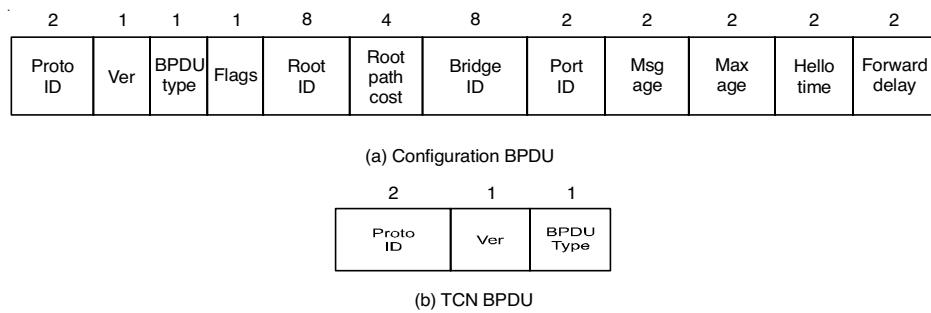


FIGURE 3.14 Formats of BPDUs

3.18.1 Bridge ID

In a spanning tree network, each bridge has a unique *Bridge ID*, and each port of a bridge is given a cost (based on bandwidth). A bridge ID consists of a 2-byte bridge priority and a 6-byte MAC address. The default priority is 32768.

3.18.2 Port State

Each port has a port state under the STP. The possible states are listed below along with the default timers that control the transition times.

1. Disabled—Administratively down and all frames discarded
2. Blocking—Only BPDUs are received (20 sec)
3. Listening—BPDUs are received and forwarded (15 sec)
4. Learning—Bridging table is built (15 sec)
5. Forwarding—User traffic is forwarded; BPDUs are forwarded

At power up, all ports are disabled. They then go through from the blocking state and the transitory states of listening and learning states. The ports finally stabilize to the forwarding or blocking state. The convergence time for the default timer is 30 to 50 seconds.

3.18.3 STP Timer

A number of *timers* are used in STP:

- Hello—2 sec (the time between the periodic Configuration BPDUs)
- Max Age—20 sec (controls the duration of the blocking state)
- Forward Delay—15 sec (the time spent by a port in the Listening and Learning states)

3.19 HOW STP WORKS

To construct a spanning tree topology, the bridges have to go through the following steps:

Step 1. Electing a Root Bridge

Step 2. Electing Root Ports

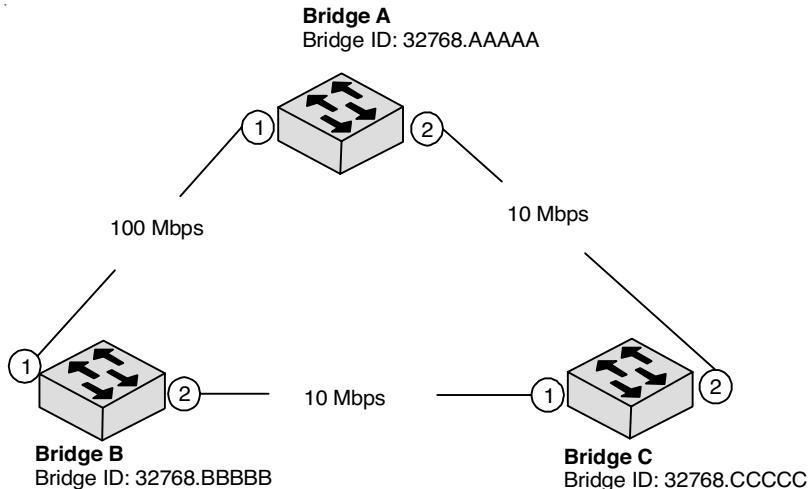
Step 3. Electing Designated Ports**Step 4.** Changing Port States (Forwarding and Blocking)**Step 5.** Maintaining the Spanning Tree

FIGURE 3.15 A network consists of three bridges. Each of them has a connection with the others

Step 1. Electing a Root Bridge

A spanning tree network has one and only one *root bridge*. The bridge with the smallest bridge ID is the root bridge. Therefore, the bridge ID is critical in the root election. The bridge ID consists of a 2-byte bridge priority and a 6-byte MAC address. The default priority is 32768. How do the bridges know which bridge has the smallest bridge ID?

When the bridges in the network are powered up, all of the ports on the bridges are in a blocking state. They are all listening for BPDUs.

1. At this time, they all think of themselves as the root bridge, and mark the root IDs with their bridge IDs. After that, each of them will send out the Configuration BPDU to announce that it is the root with a “Hello Time” periodicity.
2. Suppose that Bridge B is the first one to send out the BPDU with the root ID of 32768.BBBBBB announcing that “I am the root.”
3. When Bridge C receives that BPDU, its root ID will change from 32768.CCCCCC to 32768.BBBBBB (agreeing that Bridge B is the root), because Bridge B’s root ID is smaller. However, when Bridge A receives the

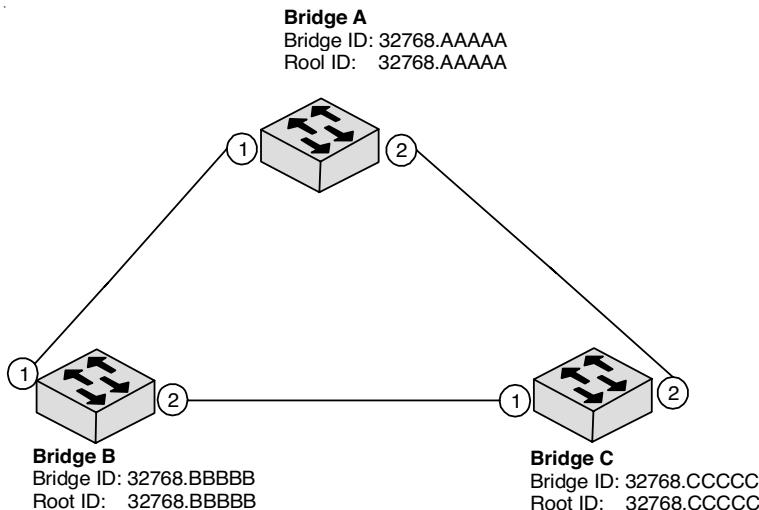
BPDU, it will not change its root ID because its root ID (32768.AAAAA) is smaller than Bridge B's root ID. At this point, both Bridges A and B consider themselves as the root. Assume that the next step is for Bridge A to send out BPDUs announcing that it is the root.

4. When Bridges B and C receive the BPDU from Bridge A, they all accept Bridge A is the root and change their root IDs to 32768.AAAAA.
5. At this point, Bridge A has been elected to be the root bridge by all bridges.
6. Here is more technical information about the root identification. When a port receives a better BPDU (with a smaller bridge ID), it will stop sending out the Configuration BPDU on that port. Therefore, at the end, only one switch will continue speaking, and the root election phase can then conclude that that one is the root.

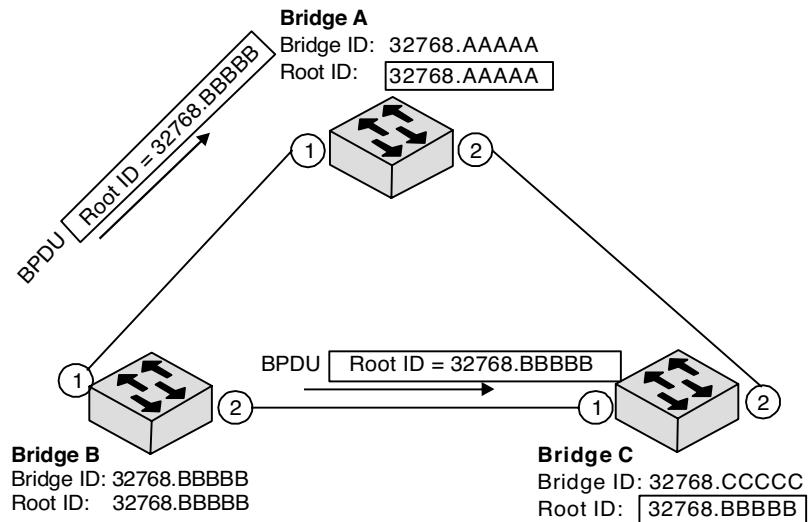
Step 2. Electing Root Ports

Each non-root bridge must elect a *root port*, which is the port providing the least path cost (based on bandwidth) to the root.

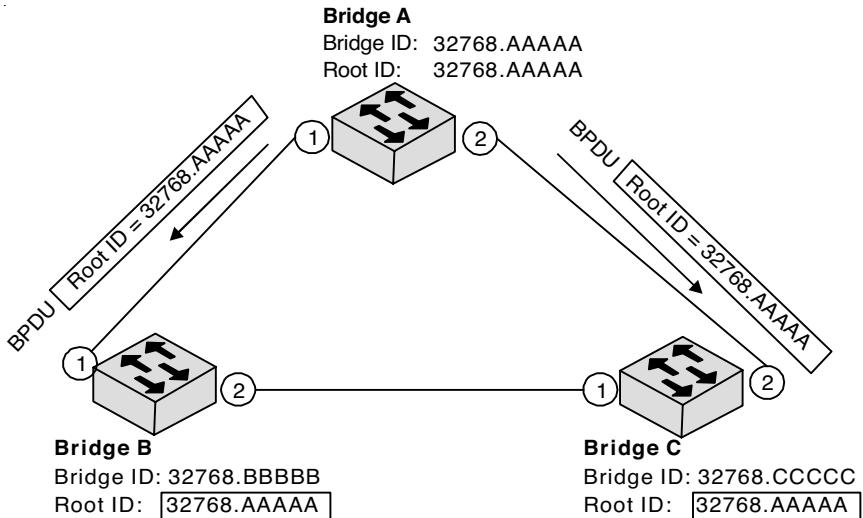
1. By examining the Root Path Cost field of the BPDUs, Bridge B knows its port 1 is the nearest to the root (port 1's cost is 19 whereas port 2's is 200. It then assigns port 1 as the root port).



(a) All consider themselves the root.



(b) Both Bridges A and B consider themselves the root after Bridge B sends out its BPDU.



(c) All bridges agree that Bridge A is the root bridge.

FIGURE 3.16 Electing a root bridge

Step 3. Electing Designated Ports

Each LAN segment must elect a *designated port*, which is the port providing the least path cost from the segment to the root. All traffic from a segment will be forwarded to the root via its corresponding designated port.

1. Segment 1 and Segment 2 elect Bridge A's port 1 and port 2 as their designated ports, respectively, because the two ports are closest to the root in the two segments (they are directly connected).
2. Segment 3 chooses Bridge B's port 2 as its designated port because it provides smaller path cost (cost=19) than Bridge C's port 1 provides (cost=100).

Step 4. Changing Port States (Forwarding and Blocking)

Bridges set root ports and designated ports to the *forwarding state* and set other ports to the *blocking state*. Bridges forward frames only between the root port and designated ports for the corresponding segments. The blocked ports will not be included in the spanning tree topology. Although the blocked ports discard the user frames, they still accept STP BPDUs.

1. Only Bridge C's port 1, which is neither root nor designate port, is blocked.
2. At this point, the spanning tree has fully converged.

Step 5. Maintaining the Spanning Tree

In normal STP operations

1. The root bridge periodically sends Configuration BPDUs, called Hello BPDUs (cost of 0), through all its ports, at an interval called Hello Time. In other words, all non-root bridges keep receiving Hello BPDUs from the root bridge on its root port.
2. All non-root bridges then forward the Hello BPDUs out of their designated ports.

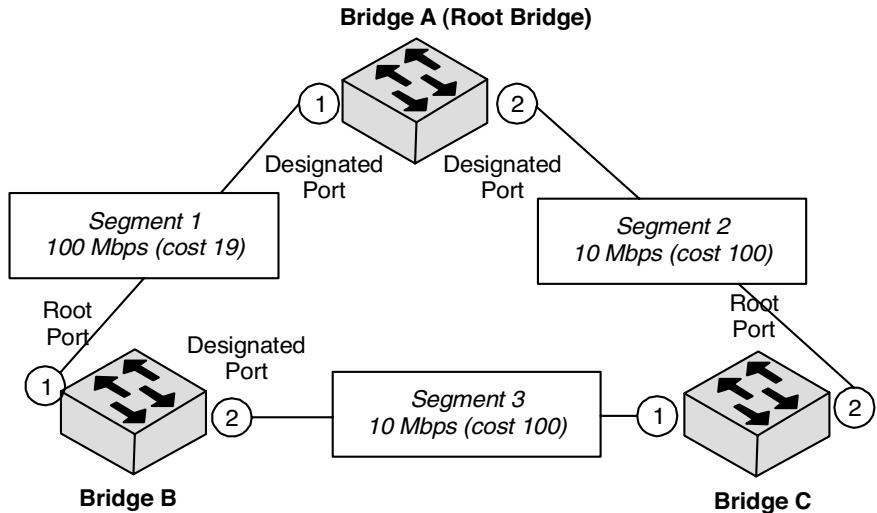


FIGURE 3.17 Root port and the designated port assignments

All non-root bridges keep receiving Hello BPDUs from the root bridge on its root port. All non-root bridges then forward the Hello BPDUs out their designated ports.

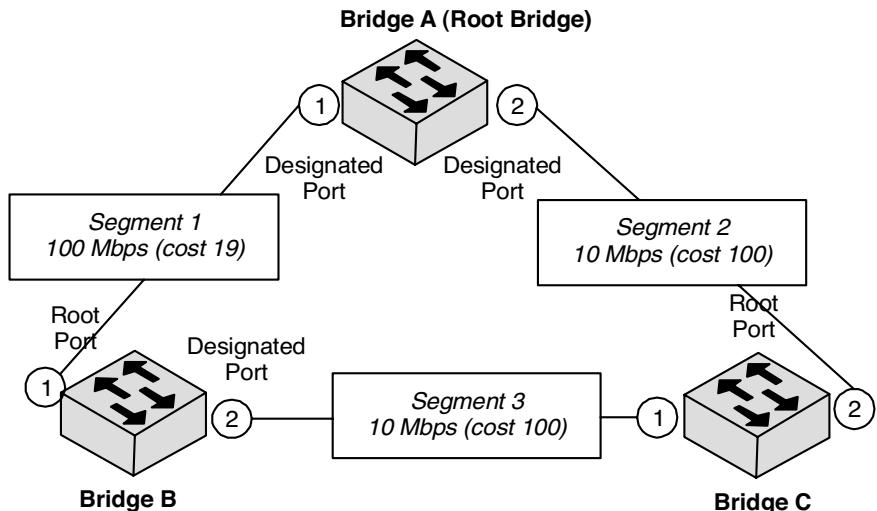


FIGURE 3.18 Root port and designated port assignments

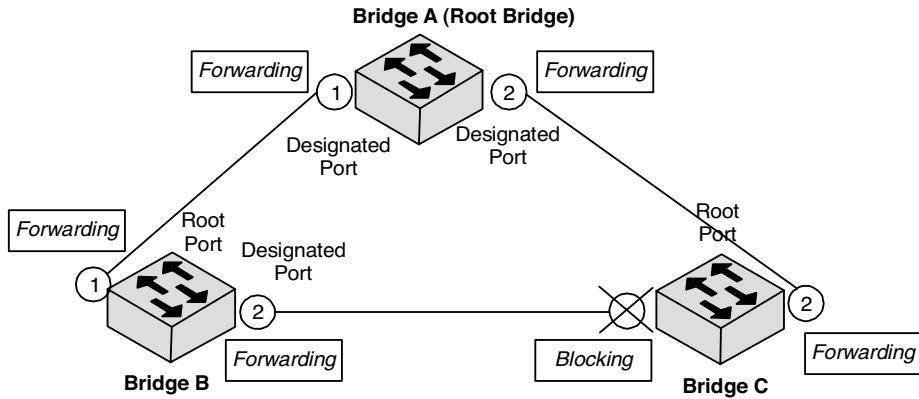


FIGURE 3.19 When the port states have been set, the spanning tree converges

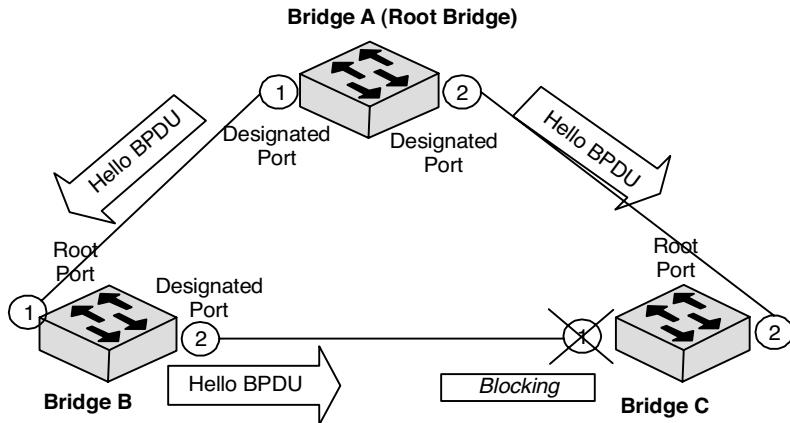


FIGURE 3.20 Maintaining the spanning tree

3.20 TOPOLOGY CHANGE

3.20.1 Failure to Receive the Hello Bpdus

As mentioned above, all bridges keep receiving periodical Hello BPDUs from the root bridge. If a bridge does not receive a Hello BPDU for a period of time, as defined by a parameter called the *Max Age Time* (20 seconds, by default), it will assume that either the root bridge is down or that the link to it is broken. In this case, it initiates network topology reconfiguration by

sending Topology Change Notification (TCN) BPDUs. When a bridge detects a topology change (such as when a link goes down and a port transitions to its forwarding state), it advertises the event to the whole bridged network:

1. It first advertises Topology Change Notification (TCN) BPDUs on its root port to its designated bridge. (This kind of BPDU is sent each Hello interval until it receives the acknowledgment from its upstream bridge).
2. After receiving the TCN BPDU, its designated bridge acknowledges it by replying with a Configuration BPDU with the TCA flag on. This bridge then generates another TCN BPDU for its own root port to inform the upstream bridge.
3. This process continues until the TCN BPDU hits the root bridge.
4. When the root receives the TCN BPDU, it then sets the Topology Change (TC) flag in all Configuration PDUs sent out downstream (for a period of the Forward Delay + Max Age).
5. These Configuration PDUs with a TC flag on them propagate the entire spanning tree. As a result, all bridges become aware of this topology change.
6. When the bridges detect the TC flag, they age out the CAM table entries faster (by using the Forward Delay instead of the regular 300 seconds). The use of the shorter timeout is to avoid inconsistencies in their CAM tables. Note that, during this period, it can disconnect or temporary loops can occur. Only after the root clears the TC flag, the (non-root) bridges resume their normal aging time (300 seconds), and begin the learning and forwarding operations for the new topology. It takes about 30–50 seconds to converge again. Later, we discuss how attackers can exploit this process.

3.21 STP ATTACK SCENARIOS

With the lack of authentication in the BPDU messages in STP, any host running bridging software can participate in the spanning tree. For example, any Linux computer can be configured as a software-based bridge by using the BRIDGE-UTILS package. It is not difficult for an insider to send out bogus BPDUs to attack the spanning tree. The following section shows some of the attacks.

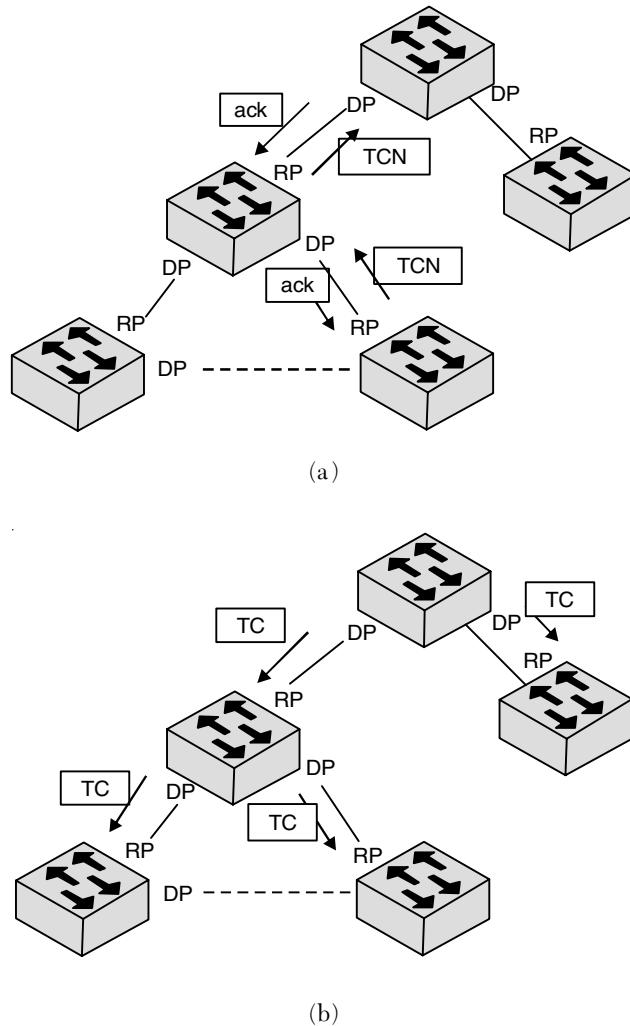


FIGURE 3.21 Operations involved in a topology change

3.22 ROOT CLAIM AND MITM

By sending a bogus BPDU with the lowest Bridge ID, the attacker station is elected as the new root bridge. This malicious root can accomplish different kinds of attacks. The new root can sniff all frames passing through it, as it is in the “middle” of the spanning tree network. As mentioned before, this new

root can be any host running bridge software, e.g., a Linux computer, and therefore, it is not difficult for that host to further process the sniffed frames to snoop on any sensitive information, such as passwords.

On the other hand, if it deliberately ignores TCNs, the bridges in the network will not adjust to any new network change. In this case, the spanning tree network no longer guarantees a loop-free topology.

Eternal root election: Another way to cause network instability is to force the network to keep selecting a root. To do that requires the following steps:

1. The attacker first learns the identifier of the root bridge, `root_id`. This can be done by using a sniffer to monitor the periodic configuration BPDUs from the root bridge, which contains its identifier.
2. Then, it injects a BPDU with the identifier `id=root_id-1` (just lower than the existing root bridge's ID) into the network. This will cause a root to elect itself.
3. After the election process, the attacker sends into the network another BPDU with the identifier `id=id-1`. This will cause another root election.
4. The attacker keeps doing this.
5. When the ID reaches its lowest value, the attacker can use the value calculated at the beginning of attack, and repeat the above process again. As a result, the network will always be in the root election process, and the ports of the bridges will never become the forwarding state, making the network unstable, and this process can disable the network.

Persistent TCN messages: A TCN BPDU from a bridge can cause the root to broadcast BPDUs with the TC flag on, and upon receiving that, the non-root bridges will age out their CAM tables more quickly. Therefore, by sending a steady stream of TCNs, an attacker can cause every bridge of the network to age out continuously. This could cause path loops and drop the network into an unusable state. The attacker's bridge can be just a computer running a software bridge or a computer with a packet injection tool.

3.23 AFFECTING NETWORK PERFORMANCE

By gaining the root role in the tree, an attacker could change the traffic flow in the switched network. The switches in the core layer have a much higher bandwidth than that in the access layer. If the computer under the access

switch runs a bridge application and claims that it is the root (by advertising a BPDU with the lowest bridge ID), the spanning tree will be reconstructed, and finally, the Gigabit link will be blocked. In this case, all the data will flow through the 100 Mbps link, and the network performance will be substantially deteriorated.

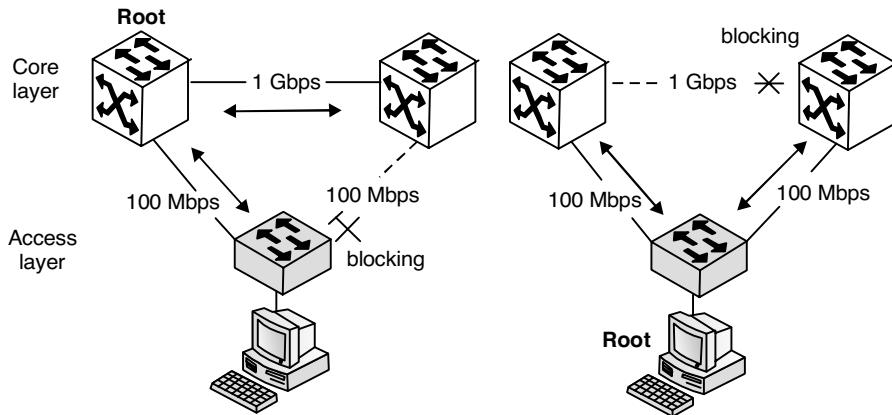


FIGURE 3.22 Changing the traffic flow

3.24 COUNTERMEASURES

3.24.1 BPDU Guard

As can be seen, without authentication, any host running a bridge application can easily participate in the spanning tree and take over the root role. The PortFast BPDU Guard feature, one of the STP enhancements created by Cisco, was designed as a defense against STP attacks. BPDU Guard is configured on a per-port basis. Administrators can enable one or more ports of a switch to be BPDU guarded. The ports with BPDU Guard enabled do not allow the hosts behind them to send BPDUs. As soon as those ports receive BPDUs, they are blocked. Without being able to send BPDUs, the hosts behind the BPDU Guard ports cannot affect the active STP topology. Those hosts can only receive and send normal data frames.

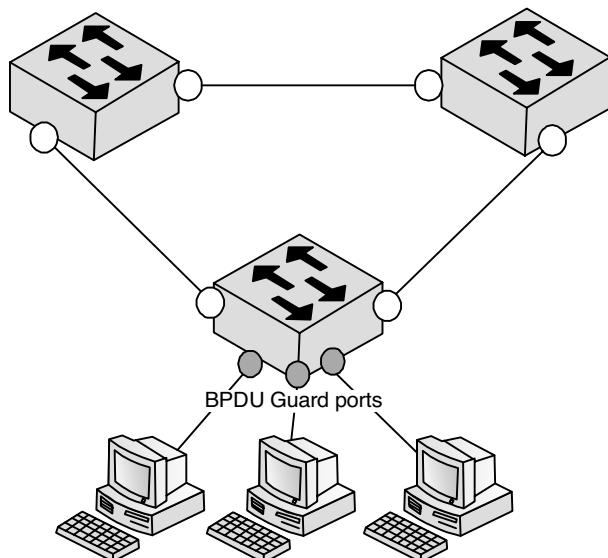
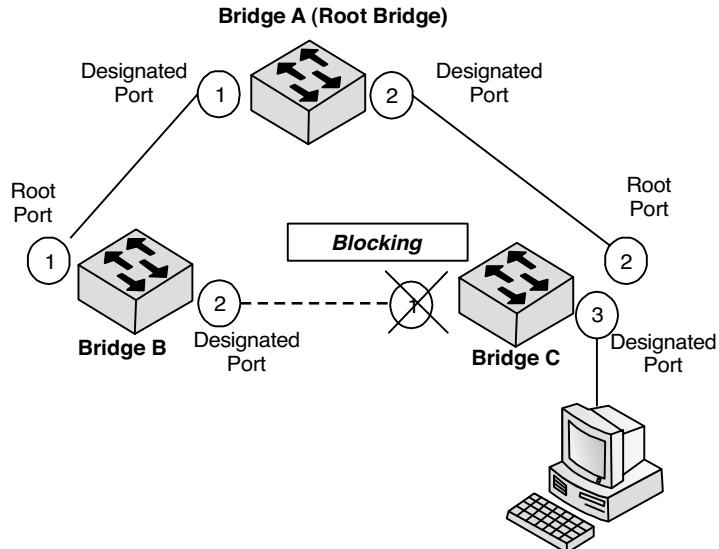


FIGURE 3.23 The ports with BPDU Guard block BPDUs from the hosts behind them

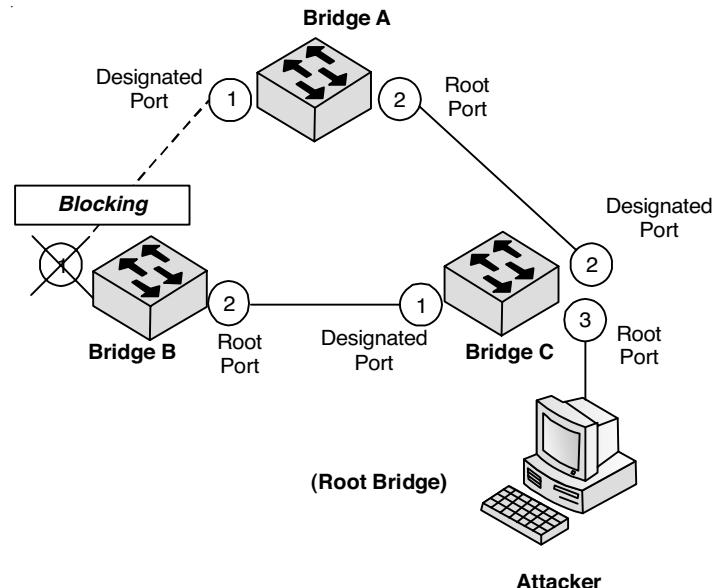
3.25 ROOT GUARD

Root Guard is another STP enhancement created by Cisco. It is used to enforce the root bridge placement in the network. When an attack takes over the root role in the network, the tree topology changes and the ports of the bridges in the network are reconfigured appropriately. The new root makes port 3 of Bridge C change from designated port to the root port. Root Guard can prevent the above attack. It is also configured on a per-port basis. The ports with Root Guard enabled cannot become root ports and must be designated ports. If these ports receive superior BPDUs, they will be moved to a root-inconsistent STP state (instead of becoming root ports) and no traffic will be forwarded across them. Therefore, if port 3 of Bridge C is enabled with Root Guard, it cannot become a root port, thus preventing the attacker from becoming the new root. As a result, the position of the original root bridge (Bridge A) is enforced.

Note that, to successfully enforce the root bridge placement in the network, Root Guard has to be enabled on all ports where the root bridge should not appear.



(a) Normal situation.



(b) Port 3 of Bridge C becomes a root port after an attack.

FIGURE 3.24 The attacker takes over the root bridge

3.26 VLAN ATTACKS

A Virtual LAN (VLAN) is a logical group of network stations and devices. A switch can be partitioned into multiple VLANs. Although the VLANs are in the same switch, they are isolated. That is, they have their own broadcast domains, and the computers in a VLAN are restricted to communicate with the computers in the same VLAN. Different VLANs cannot communicate without the use of a router and network layer addresses. VLANs are especially useful in creating workgroups where users share the same resources, such as databases and disk storage. For example, users in the Marketing Department are placed in the Marketing VLAN, whereas users in the Engineering Department are placed in the Engineering VLAN. VLANs offer many advantages.

3.26.1 Easier Network Administration

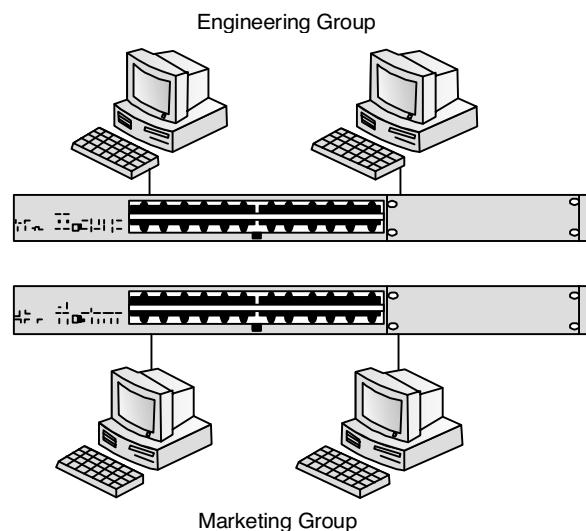
VLANs facilitate the easy assignment of logical groups of computers and easy modifications of the groups. These can be done by simply configuring the switch ports without the physical movement of the computers.

3.26.2 Improved Bandwidth Usage

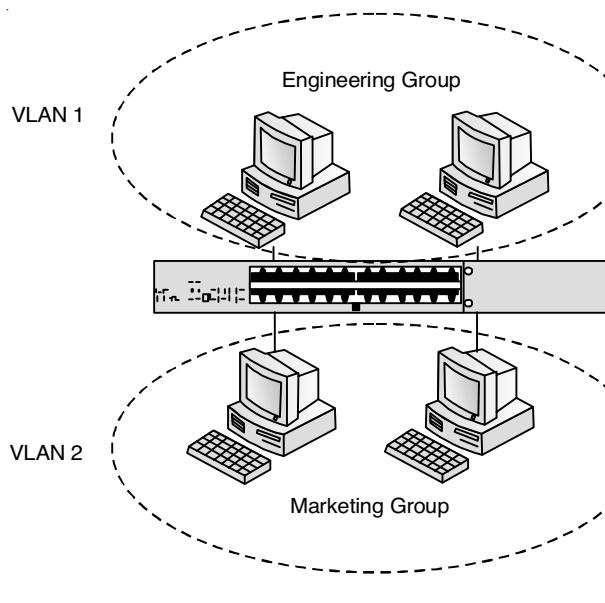
Users in the same workgroup share the similar resources, such as databases and disk storage. If workgroups can be isolated, the traffic within each workgroup does not affect the traffic of others, making better usage of the bandwidth.

3.26.3 Blocking Broadcast Traffic

All the computers connected to a switch share the same broadcast domain. However, some broadcasts are only useful to some stations. For example, Novell broadcast frames target only the hosts running Novell software only. With VLANs, different groups have their own broadcast domain, which can prevent broadcasts from reaching unrelated computers.



(a) Without a VLAN



(b) With a VLAN

FIGURE 3.25 The VLANs in a switch are logically isolated

EXERCISES

1. What is an Internet Service Provider (ISP)? Explain how it works.
2. Discuss the Network Access Point (NAP).
3. What is a router? Explain how it works.
4. Discuss addressing.
5. Give a brief description of ATM.
6. Explain Ethernet.
7. What is a Fiber Distributed Data Interface (FDDI)?
8. What is Multi-Protocol Label Switching (MPLS)?
9. Explain the Point-To-Point Protocol (PPP).
10. What is the High-Level Data Link Control (HDLC)?
11. What is a BPDU guard?
12. Explain how STP works.

CHAPTER 4

IP SECURITY AND FIREWALLS

Chapter Goals

- Protective devices
- Denial of service
- Spies (Industrial and otherwise)
- Network taps
- Host security
- A firewall can log Internet activity efficiently
- Buying versus building
- A firewall cannot fully protect against viruses

4.1 INTERNET FIREWALLS

A firewall is a system or group that enforces an access control policy between two or more networks. The firewall can be thought of as a pair of mechanisms that exist to block traffic. A firewall's purpose is to keep unauthorized users out of a network while still allowing people to get their jobs done. It is scarcely possible to go anywhere, read a magazine or a newspaper, or listen to a news broadcast without seeing or hearing about the Internet. It is so popular that no advertisement is complete without a reference to a Webpage. While non-technical publications are obsessed with the Internet, technical publications are obsessed with security. It's a logical progression: Once the first excitement of having a superhighway in your neighborhood wears off, people notice that

not only does it allow for rapid travel, but it also lets in a very large number of strangers to the neighborhood. (Not all of them are people you would have invited.)

The Internet is a marvelous technological advance that provides access to information and the ability to publish information in revolutionary ways. This book is about one way to balance the advantages and the risks to take part in the Internet while still protecting yourself.

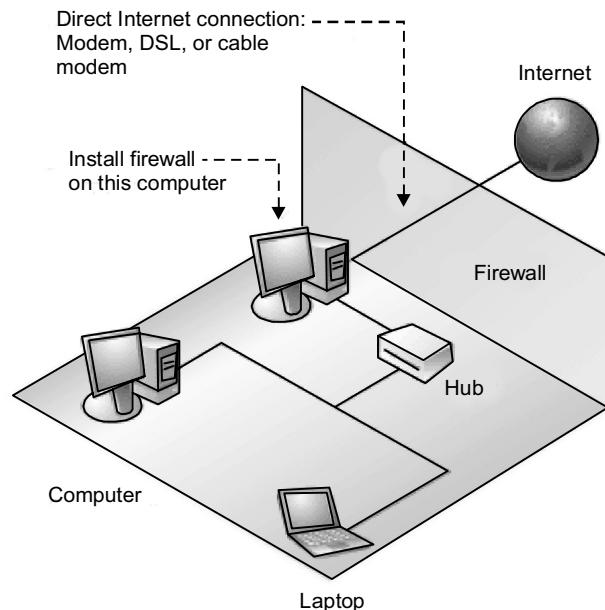


FIGURE 4.1 An example of a firewall

Later in this chapter, we describe different models of security that people have used to protect their data and resources on the Internet. Our emphasis in this book is on the network security model and in particular, the use of Internet firewalls. A firewall is a form of protection that allows a network to connect to the Internet while maintaining a degree of security. The section later in this chapter called “What is an Internet firewall?” describes the basics of firewalls and summarizes what they can and cannot do to help make a site secure. There are some important questions addressed here: What are you protecting on your systems? What types of attacks and attackers are common? What types of security can you use to protect your site?

4.2 PROTECTIVE DEVICES

A firewall is basically a protective device. If you are building a firewall, the first thing you need to worry about is what you're trying to protect. When you connect to the Internet, you're putting three things at risk:

- Your data: The information you keep on the computers
- Your resources: the computers themselves
- Your reputation

4.2.1 Your Data

Your data has three separate characteristics that need to be protected.

Secrecy

You might not want other people to know it.

Integrity

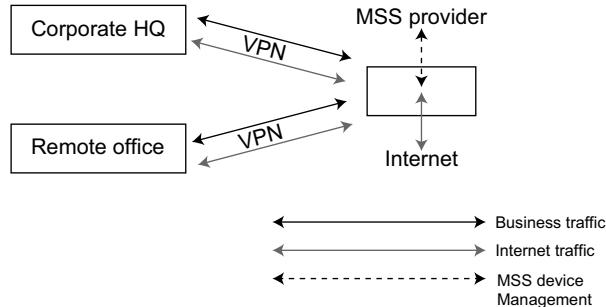
You probably don't want other people to change it.

Availability

You almost certainly want to be able to use it yourself.

People tend to focus on the risks associated with secrecy, and it's true that those are usually large risks. Many organizations have some of their most important secrets—the designs for their products, financial records, or student records—on their computers. However, you may find that for your site, it is relatively easy to separate the machines containing highly secret data from the machines that connect to the Internet. (Or you may not. You cannot carry out e-commerce without having information about orders and money pass through Internet-accessible machines.)

Suppose that you can separate your data in this way, and that none of the information that is Internet accessible is secret. In that case, why should you worry about security? Because secrecy isn't the only thing that must be protected. There are also important concerns about integrity and availability. (After all, if your data isn't secret, and if you don't mind it being changed, and if you don't care whether anybody can get to it, why are you wasting disk space on it?)

**FIGURE 4.2** Using a protective device

Even if your data isn't particularly secret, you'll suffer the consequences if it's destroyed or modified. Some of these consequences have readily calculable costs: once data is lost, it is costly to have it reconstructed. If you were planning to sell that data in some form, you'll have lost sales regardless of whether the data is something you sell directly, the designs from which you build things, or the code for a software product. Intangible costs are also associated with security incidents. The most serious is the loss of confidence (user confidence, customer confidence, investor confidence, staff confidence, student confidence, and public confidence) in your systems and data. Consequently, this results in a loss of confidence in your organization.

Computer security incidents are different from many other types of crimes because detection is unusually difficult. It may take a long time to find out that someone has broken into your site—sometimes you'll never know. Even if somebody breaks in but doesn't actually do anything to your system or data, you'll probably lose time (hours or days) while you verify that the intruder did not do anything. In a lot of ways, a brute-force “trash-everything” attack is easier to manage than an attack that doesn't appear to damage your system. If the intruder destroys everything, you restore from backups and start over. But if the intruder doesn't appear to have done anything, you spend a lot of time second-guessing yourself, wondering what he or she might have done to the system or data. The intruder almost certainly has done something—most intruders start attacks by making sure that they have a way to get back in before they do anything else.

Although this book is primarily about preventing security incidents, it also includes responding to security incidents, and supplies some general guidelines for detecting, investigating, and recovering from security incidents.

4.2.2 Resources

Even if you have data you don't care about (or, perhaps you enjoy reinstalling your operating system every week), if other people are going to use your computers, you probably would like to benefit from this use in some way. Most people want to use their own computers or they want to charge other people for using them. Even people who give away computer time and disk space usually expect to get good publicity and good will; they aren't going to get it from intruders. Since you spend time and money on your computing resources, it is your right to determine how they are used.

Intruders often argue that they are using only excess resources; as a consequence, their intrusions don't cost their victims anything. There are two problems with this argument.

First, it's impossible for an intruder to determine successfully what resources are excess and use only those. It may look as if a system has a significant amount of empty disk space and hours of unused computing time. In fact, though, the user might be just about to start computing animation sequences that are going to use every bit and every microsecond. An intruder cannot restore resources when the user wants them, either. (Here is another way to think about this: I don't ordinarily use my car between midnight and 6 a.m. However, that doesn't mean I am willing to lend it to you without being asked. What if I have an early morning flight the next day, or what if I'm called out to deal with an emergency?)

Second, it is the computer user's right to use their resources the way they want to. That may mean that a significant amount of disk space remains empty and unused.

4.2.3 Reputation

When an intruder appears on the Internet with a stolen identity, anything he or she does is attributed to their victim. What are the consequences of this type of action?

Most of the time, the consequences involve other sites or law enforcement agencies trying find out why the intruder is breaking into these systems. This is not as rare an occurrence as it may seem. One site got serious about security when its system administration staff added a line item to the company's time cards after a conversation with the FBI about break-in attempts originating from the company's site.

Sometimes, such imposters cost more than lost time. An intruder who actively dislikes someone or takes pleasure in making life difficult for others

may change a Website, send electronic mail, or post new messages that purposely claim to come from a company or individual. Generally, the people who choose to do this are doing it for spite, rather than believability. However, even if only a few people believe these messages, the recovery process can be long and humiliating. Anything even remotely believable can do permanent damage to a reputation.

For example, an impostor posing as a Texas A&M professor sent out hate email containing racist comments to thousands of recipients. The impostor was never found, but the professor is still dealing with the repercussions of the forged message. In another case, a student at Dartmouth sent out an email using the signature of a professor late one night during exams. The email claimed the professor had a family emergency: The forged email canceled the next day's exam, and only a few students showed up.

General Network Layout with a Firewall

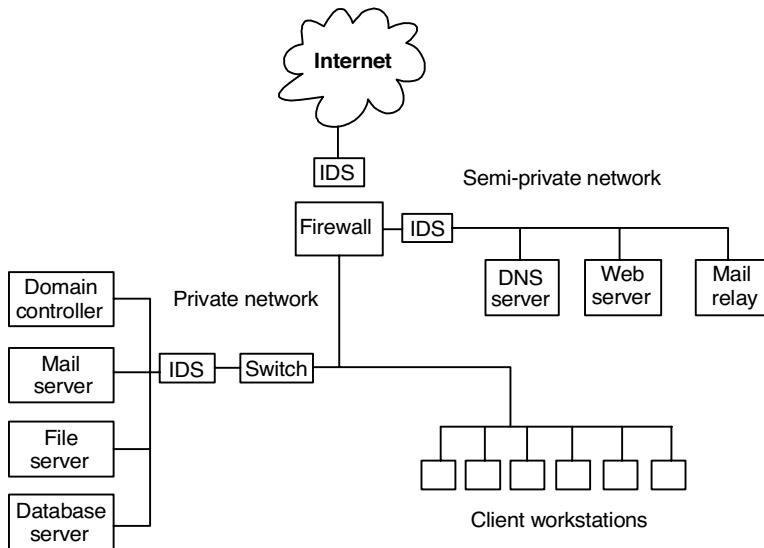


FIGURE 4.3 General network layout with a firewall

It's possible to forge electronic mail or news without gaining access to a Website, but it's much easier to show that a message is a forgery if it's generated from outside the forged site. The message coming from an intruder who has gained access to your site will look exactly like yours because they are pretending to be you. An intruder will also have access to details that an external forger won't. For example, an intruder who has all of your mailing lists available and knows exactly who you send mail to has inside information.

Currently, attacks that replace Websites are very popular; one list shows more than 160 successful attacks simply replaced the sites via boosting by the attackers, but a significant portion of the attacks were directed at the content of the sites. A site that should have touted Al Gore's suitability for the U.S. presidency was replaced by a similar anti-Gore site. Political movements in Peru, Mexico, and China have been involved in cyberattacks, and many entertainment sites, including those for pop stars, pro Wrestling, and the Boston Lyric Opera, all suffered as well.

Even if an intruder does not steal an identity, a break-in at a site isn't good for a company's reputation. It shakes people's confidence in an organization. In addition, most intruders will attempt to go from one company's machines to other companies' machines, which is going to make their next victims think of the first site as a platform for computer criminals. Many intruders will also use a compromised site as distribution site for pirated software, pornography, and/or other stolen information. It is difficult to recover when a business or person's name is linked to software piracy or pornography.

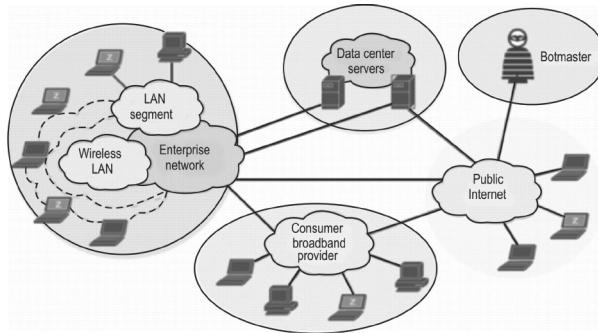
What's out there to worry about? What types of attacks are you likely to face on the Internet, and what types of attackers are likely to be carrying them out? In the sections that follow, we touch on these topics, but don't go into any technical detail. Later chapters describe the kinds of attacks in some detail and explain how firewalls can help protect against them.

4.3 TYPES OF ATTACKS

There are many types of attacks on systems, and various ways of categorizing these attacks. In this section, we break attacks down into three basic categories: intrusion, denial of service, and information theft.

4.3.1 Intrusion

The most common attacks on computer systems are intrusions. With *intrusions*, hackers are able to use someone else's computers. Most attackers want to use these computers as if they were legitimate users. Attackers have dozens of ways to obtain access. They range from social engineering attacks (where a hacker discovers the name of a high-level individual in the company, calls a system administrator claiming to be that person, and then saying their password needs to be changed right now so that they can get important work done) to simple guesswork (the attacker tries account names and password combinations until one works) to intricate ways to get in a system without knowing an account name and a password.

**FIGURE 4.4** Common intrusions

Firewalls help prevent intrusions in a number of ways. Ideally, they block all ways to get into a system without needing an account name and password. Properly configured, they reduce the number of accounts accessible from the outside that are vulnerable to guesswork or social engineering. Most people configure their firewalls to use one-time passwords that prevent guessing attacks. Even if you do not use these passwords and authentication and auditing services, a firewall will give you a controlled place to log attempts to get into your system, and, in this way, they help detect guessing attacks.

4.3.2 Denial of Service

A *denial of service attack* is aimed entirely at preventing users from using their own computers.

In late 1994, writer Josh Quittner and Michelle Slatalla were the target of an “electronic mail bomb.” Apparently in retaliation for an article on the cracker community they’d published in *Wired* magazine, someone broke into IBM, Sprint, and the writers’ network provider, and modified programs so their email and telephone service was disrupted. A flood of emails couldn’t get through; eventually, their Internet connection was shut down entirely. Their phone service also fell victim to the intruders, who reprogrammed the service so that the callers were routed to an out-of-state number where they heard an obscene recording.

Although some cases of electronic sabotage involve the actual destruction or shutting down of equipment or data, more often they follow the pattern of flooding seen in the Quittner-Slatalla case or in the case of 1988 Morris Internet worm. An intruder so floods a system or network—with messages, processes, or network requests—that no real work can be done. The system or network spends all its time responding to messages and requests and can’t satisfy any of them.

While flooding is the simplest and most common way to carry out a denial of service attack, a clever attacker can also disable services, re-route them, or replace them. For example, the phone attack in Quittner-Slatalla case denied phone service by re-routing the phone calls elsewhere. It is possible to mount the same kind of attack against an Internet service.

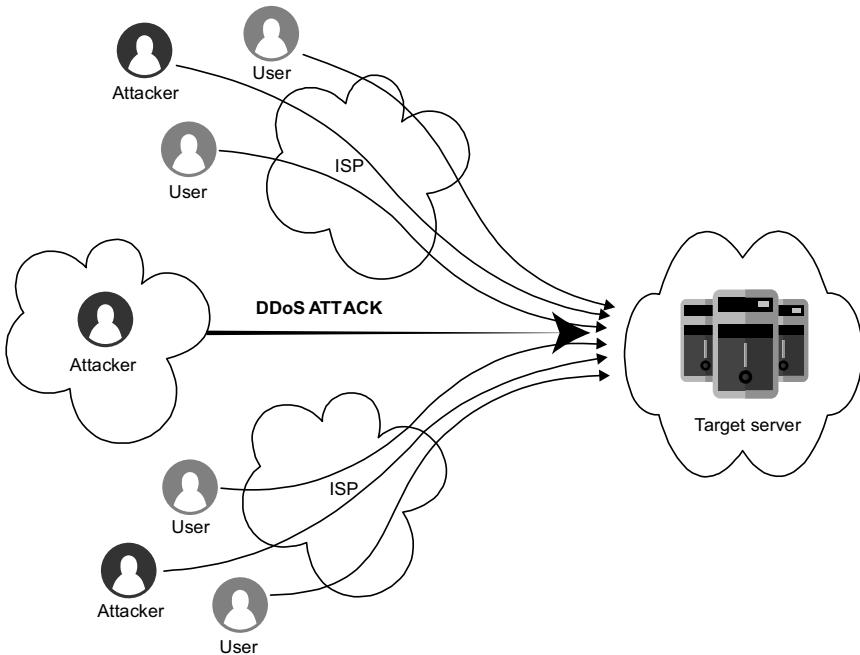


FIGURE 4.5 An example of a distributed denial of service (DDoS) attack

It is almost impossible to avoid all denial of service attacks. For example, many times, administrators set accounts to become unusable after a certain number of failed login attempts. This prevents attackers from simply trying passwords until they find the right one. Unfortunately, this approach provides attackers with an easy way to mount a denial of service attack: they can lock any user's account simply by trying to log in a few times.

Most often, denial of service attacks are un-avoidable. If you accept things from the outside world, be it electronic mail, telephone calls or packages, it is possible to get flooded. The notorious college prank of ordering a pizza or two from every pizzeria in town to be delivered to your least favorite person is a form of denial of service: it's hard to do much else while arguing with 42 pizza deliverers. In the electronic world, denial-of-service is as likely to happen by accident as on purpose (such as a persistent fax machine faxing something to

a voice line). The most important thing is to set up services so that if one of them is flooded, the rest of them can still function while the problem is found and corrected.

Flooding attacks are considered “unsporting” by many attackers because they are not difficult to carry out. For most attackers, they are also pointless, because they do not provide the attackers with the information or the ability to use your computers (the payoff for most other attacks). Intentional flooding attacks are usually the work of people who are angry at a particular person or company, and such people are quite rare.

With the right tools and co-operation, it's fairly easy to trace flood packets back to their source, but that might not help determine who is behind the attacks. The attacks almost always come from machines that have themselves been broken into. Only a really low-level attacker generates an easily traced flood of packets from their own machine. Sometimes flooding attacks are carried out by remote control. Attackers install remotely controlled flooding software on systems that they break into over the course of many weeks or months. This software lies dormant and undiscovered until some later time, when they trigger many of these remotely-controlled installations to simultaneously bombard their victims with massive floods of traffic from many different directions at once. This was the method behind the highly publicized denial of service attacks on Yahoo!, CNN, and other high profile Internet sites early in the year 2000.

Unintentional flooding problems are more common than intentional ones, as we discuss in the “Stupidity and Accidents Section” later in this chapter.

Some denial of service attacks are easy for attackers to carry out, and these are relatively popular. Attacks that involve sending a small amount of data that cause machines to reboot or hang are very popular with the same sort of people who like to set off fire alarms in dormitories in the middle of the night, for much the same reason; with a small investment, an attacker can annoy a very large number of people who are unlikely to be able to find him afterwards. The good news is that most of these attacks are avoidable. A well-designed firewall will usually not be susceptible to them and will prevent them from reaching internal machines that are vulnerable.

4.4 NETWORK TAPS

Some types of attacks allow an attacker to get data without ever having to directly use a system's computers. Usually, these attacks exploit Internet services that are intended to give out information, inducing the services to give

out more information than was intended, or to give it out to the wrong people. Many Internet services are designed for use on local area networks, and don't have the type or degree of security that would allow them to be used safely across the Internet.

Information theft doesn't need to be active or particularly technical. People who want to find out personal information could simply ask (perhaps pretending to be somebody who had a right to know): this is referred to as *active information theft*. They could also tap a phone: this is *passive information theft*. Similarly, people who want to gather electronic information could actively query for it (perhaps pretending to be a machine or a user with valid access) or could passively tap the network and wait for it to flow by.

Most people who steal information try to get access to a user's computers; they are looking for user names and passwords. Fortunately for them, and unfortunately for everybody else, that's the easiest kind of information to get when tapping a network. User name and password information occurs quite predictably at the beginning of any network interaction, and such information can often be reused in the same form.

How would a hacker proceed if they wanted to find out how somebody answers her telephone? Installing a tap would be an easy and reliable way to get that information, and a tap at a central point in the telephone system would yield the telephone greetings of hundreds or thousands of people in a short period of time.

What if the attacker wants to know how somebody spells his or her last name, or what the names and ages of his or her children? In this case, a telephone tap is a slow and unreliable way to get that information. A telephone tap at a central point in the system will probably yield that information about some people. It will certainly gather some secret information that could be used in interesting ways, but the information is going to be buried among the conversations of hundreds of people setting up lunch dates and chatting about the weather.

Similarly, network taps, which are usually called *sniffers*, are very effective at finding password information but are rarely used by attackers to gather other kinds of information. Getting more specific information about a site requires either extreme dedication and patience, or the knowledge that the information will reliably pass through a given place at a given time. For example, if an attacker knows that the victim calls the bank to transfer money between her checking savings accounts at 2 pm every other Friday, it is worth tapping that phone call to find out the person's access codes and account numbers. However, it is probably not worth tapping somebody else's phone, on the off chance that they will do such a transfer because most people don't transfer money over the phone at all.

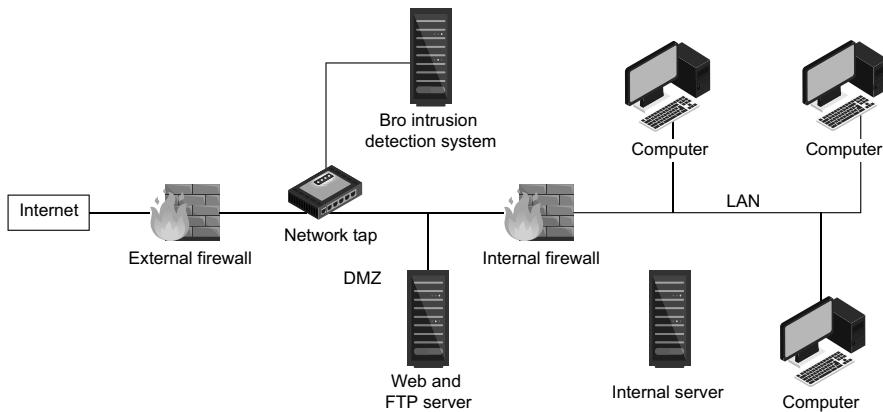


FIGURE 4.6 An illustration of a network tap

Network sniffing is much easier than tapping a telephone line. Historically, the connectors used to hook a computer to an Ethernet network were known as network taps (that's why the term *tapping* is not used for spying on a network), and the connectors behave like taps, too. In most networks, computers can see traffic that is intended for other hosts. Traffic that crosses the Internet may cross any number of local area networks, any one of which can be a point of compromise. Networks service providers and public-access systems are very popular targets for intrusions; sniffers placed there can be extremely successful because so much traffic passes through these networks.

There are several types of protection against information theft. A properly configured firewall will protect the system against people who are trying to obtain more information. Once you have decided to give information out on the Internet, however, it's very difficult to protect against that information reaching an unintended audience, either through mis-authentication (somebody falsely claiming to be authorized) or through sniffing (somebody simply reading information as it crosses a correctly authorized channel). For that matter, once the information is given out to somebody, there is no way to prevent that person from distributing it to other people. These risks are outside of the protection a firewall can provide because they occur once information has intentionally been allowed to go outside a network.

4.5 IP SECURITY FIREWALL

This section very briefly describes the types of attackers who are out there on the Internet. There are many ways to categorize these attackers; we can't really do justice to the many variants of attackers we've seen over the years,

and any quick summary of this kind necessarily presents a rather stereotyped view. Nevertheless, this summary may be useful in distinguishing the main categories of attackers.

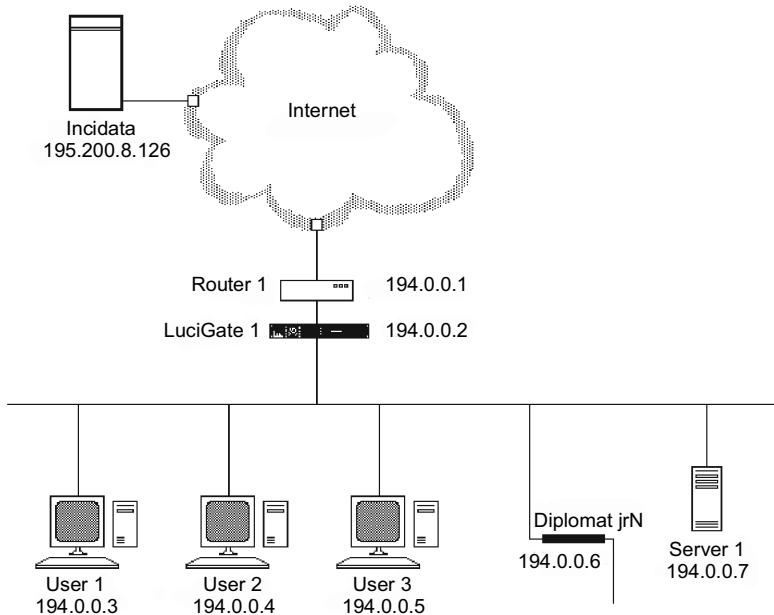


FIGURE 4.7 IP Security firewall

All attackers share certain characteristics. They do not want to be caught, so they try to conceal themselves, their identity, and real geographic location. If they gain access to a system, they will certainly attempt to preserve that access, if possible, by building in extra ways to get access (and they hope no one will notice these access routes, even if the attackers are found). Most of them have some contact with other people who have the same kinds of interests (“the underground” is not hard to find), and most will share the information they get from attacking a system. A second group of attackers may not be as benign.

4.6 JOY RIDERS

Joy riders are bored people looking for amusement. They break in because they think there might be interesting data, or because it would be amusing to use someone else's computers, or because they have nothing better to do.

They might be out to learn about the kind of computer or data a user has. They are curious but not actively malicious. However, they often damage the system through ignorance or in trying to cover their tracks. Joy riders are particularly attracted to well known sites and uncommon computers.

4.7 VANDALS

Vandals are out to do damage, either because they get a thrill from destroying things or because they bear a grudge.

Vandals are a threat to somebody that the Internet underground might think of as “The Enemy” (for example, the phone company or the government) or if it is an entity that tends to annoy people who have computers and time (for example, universities with failing students, a computer company with annoyed customers, or users with an aggressively commercial presence on the Internet). People and businesses can also become a target simply by being large and visible. If a large wall is put up in certain neighborhoods, people will put graffiti on it.

Fortunately, vandals are fairly rare. People don’t like them, even people in the underground who have nothing against breaking into computers in general. Vandals also tend to inspire others to go to great lengths to find them and stop them. Unlike more mundane intruders, vandals have short but splashy careers. Most of them also go for straightforward destructions, which is unpleasant but is relatively easily detected and repaired. In most circumstances, vandals delete data or ruin computer equipment. (Actually, introducing subtle, but significant changes in programs or financial data is much harder to detect and fix.)

Unfortunately, it’s close to impossible to stop a determined vandal; somebody with a true vendetta against a site is going to infiltrate it sooner or later. Certain attacks are attractive to some attackers but not others. For example, denial of service attacks are not attractive to joy riders.

4.8 SCOREKEEPER

Many intruders engage in an updated version of an ancient tradition: They gain “bragging rights” based on the number and types of systems they have broken into.

Like joy riders and vandals, *scorekeepers* prefer certain sites. Breaking into something well known, well defended, or otherwise especially “cool” is usually worth something to them. However, they’ll also attack anything they can get at; they’re going for quantity as well as quality. They don’t have to want anything from a person or business or care in the least about the characteristics of the site. They may or may not do damage on the way through. They’ll certainly gather information and keep it for later use (perhaps using it to barter with other attackers). They’ll probably try to leave themselves a way to get back in later. If at all possible, they’ll use the hacked machines as a platform to attack others.

Scorekeepers are the attackers that are discovered long after they’ve broken into a system. Administrators usually find out about their attack slowly because something’s odd about certain machines or another site or a law enforcement agency calls because the system is being used to attack others. Sometimes, an attacker sends a copy of the private data they found on a cracked system on the other side of the world.

Scorekeepers are what are known as *script kiddies*—attackers who are not themselves technically experts but are using programs or scripts written by other people and following instructions how to use them. Although they do tend to be young, they are called kiddies mostly out of contempt aimed at them by more experienced intruders. Even though these attackers are not innovators, they still pose a real threat to sites that don’t prioritize security. Information spreads very rapidly in the underground, and script kiddies are numerous. Once a script exists, somebody is almost guaranteed to attack a site with it.

These days, some scorekeepers are not even counting machines they have broken into but are keeping score on crashed machines. On the one hand, having a machine crash is generally less destructive than having it broken into; on the other hand, if a particular attack gets into the hands of the script kiddies, and thousands of people use it to crash certain machines, it is a serious problem.

4.9 SPIES: INDUSTRIAL AND OTHERWISE

Most people who break into computers do so for the same reason people climb mountains: because they are there. While these people are not above theft, they usually steal things that are directly convertible into money or

further access (*e.g.*, credit card telephone or network access information). If they find secrets they think they can sell, they may try to do so, but that's not their main business.

As far as anybody knows, serious computer-based espionage is rare outside of traditional espionage circles. (That is, if you're a professional spy, other professional spies are probably watching you and your computers.) Espionage is much more difficult to detect than run-of-the-mill break-ins, however. Information theft need not leave any traces at all, and even intrusions are relatively rarely detected immediately. An attacker who breaks in, copies data, and leaves without disturbing anything is quite likely to get away with it at most sites.

In practical terms, most organizations cannot prevent spies from succeeding. The precautions that governments take to protect sensitive information on computers are complex, expensive, and cumbersome; therefore, they are used on only the most critical resources. These precautions include electromagnetic shielding, careful access controls, and absolutely no connections to unsecured networks.

What can average people do to protect against attackers of this kind? First, ensure that the Internet connection isn't the easiest way for a spy to gather information. Do not make it easy for an attacker to break into computers and find something that immediately appears to be worth trying to sell to spies. It should be expensive and risky to spy on a computer system. Some people say it's unreasonable to protect data from network access when somebody could get it easily by coming to a physical site. We don't agree; physical access is generally more expensive and more risky for an attacker than network access.

4.10 IRRESPONSIBLE MISTAKES AND ACCIDENTS

Most disasters are not caused through ill will; they are accidents or mistakes. One study estimated that 55 percent of all security incidents actually resulted from naïve or untrained users doing things they should not have.

Denial of service incidents, for example, frequently aren't attacks at all. Apple's corporate electronic mail was rendered non-functional for several days (and their network provider was severely inconvenienced) by an accident involving a single mail message sent from a buggy mail server to large mailing list. The mail resulted in a cascade of hundreds of error messages. The only hostile person involved was the system administrator, who wasn't hostile until he had to clean up the resulting mess. Similarly, it's not uncommon for

companies to destroy their own data or release it to the world by accident. Firewalls aren't designed to deal with this kind of problem. In fact, there is no known way to fully protect against accidents or stupidity.

4.11 THEORETICAL ATTACKS

It's relatively easy to determine the risk involved in attacks that are currently under way, but what can be done about attacks that are theoretically possible but have not yet been used? It's very tempting to dismiss them altogether—after all, what matters is not what might happen, but what actually does happen. Why should it be an issue if somebody produces a proof that an attack is possible, but it's so difficult that nobody is actually doing it? There are several reasons, such as

- the limits on what's difficult change rapidly in computing
- problems rarely come in isolation, and one attack that's too difficult may help people find an easier one
- eventually people run out of easier attacks and turn to more difficult to ones
- attacks move almost instantly from "never attempted" to "widely used"

The moment at which an attack is no longer merely theoretical but is actually in use against a site is that time that is technically called "too late." No one should wait until then. Businesses and individuals experience calmer and more peaceful lives if they don't wait until the moment when an attack hits the newspaper headlines. That's where many theoretical attacks suddenly end up.

One computer vendor decided that a certain class of attacks, called stack attacks, were too difficult to exploit, and it was not worth trying to prevent them. These attacks are technically challenging on hardware, and they were more difficult on the vendor's machines. It seemed unlikely that attackers would bother to go to the considerable effort necessary, and preventing the attacks required re-writing fundamental parts of the operating system. Thus, the vendor elected to avoid doing the tedious and dangerous re-writing work to prevent what was then considered a purely theoretical risk. Six months later, an attacker found and exploited one of the vulnerabilities. Once the hard work had been done for one, the rest were easy, so that started a landslide of exploits and bad publicity.

4.12 WHO DO YOU TRUST?

Much of security is about trust. Who is trustworthy? The world does not work unless people are trusted to do some things, and security people sometimes seem to take an overly suspicious attitude, trusting nobody. Why should users not be trusted? Or famous software vendors?

We all know that in day-to-day life, there are various kinds of trust. There are people who would be trusted with a thousand dollars, but not a secret. Some people could be trusted to babysit but not with books. There are even people who are well-loved but not allowed to touch the good china because they break things. The same is true in a computer context. Trusting employees not to steal data and sell it is not the same as trusting them not to give it out by accident. Trusting a software vendor not to sell software designed to destroy computers is not at all the same as trusting the same vendor not to let other people destroy your computer.

You don't need to believe that world is full of horrible, malicious people who are trying to attack you. You do need to live that the world has some horrible, malicious people who are trying to attack you, and is full of really nice people who don't always pay attention to what they are doing.

When you give somebody private information, you are trusting them in two ways. First, you are trusting them not to do anything bad with it. Second, you are trusting them not to let anybody else steal it. Most of the time, people worry about the first problem. In the computer context, you need to explicitly remember to think about the second problem. If you give somebody a credit card number on paper, you have a good idea what procedures are used to protect it, and you can influence them. If carbon sheets are used to make copies, you can destroy them. If you give somebody a credit card electronically, you trust not only their honesty, but also their skill in computer security. It's perfectly reasonable to worry about the latter even if the former is in impeccable.

If the people who use your computers and who write your software are all trustworthy computer security experts—great; but if they are not, decide whether you trust their expertise separately from deciding whether you trust their honesty.

What approaches can you take to protect against the kinds of attacks we've outlined in this chapter? People choose a variety of security models, or approaches, ranging from no security at all, to what's called “security through obscurity” and host security, to network security.

4.12.1 No Security

The simplest possible approach is to put no effort at all into security, and run with whatever minimal security the vendor provides by default. If you're reading this book, you've probably already rejected this model.

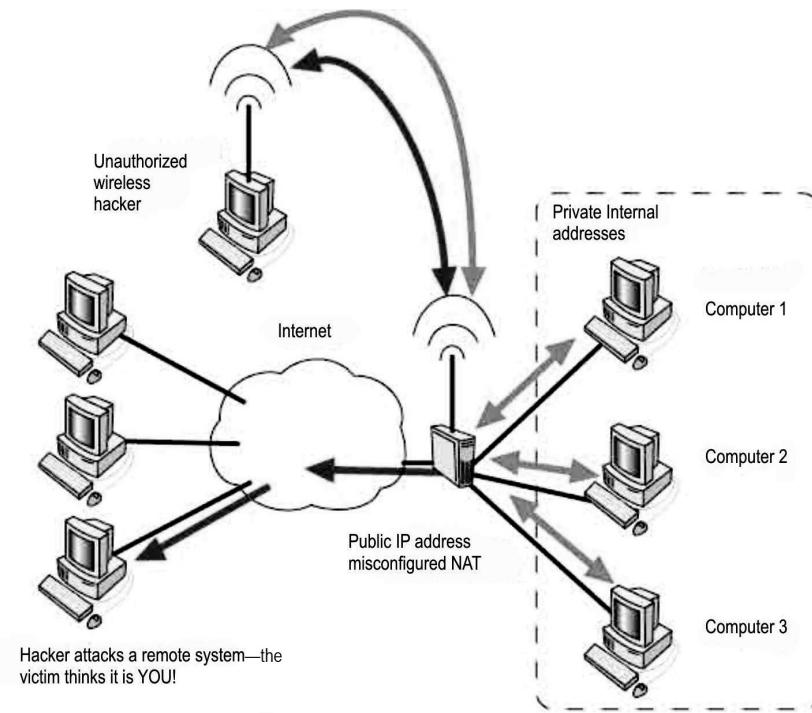
4.13 SECURITY THROUGH OBSCURITY

Another possible security model is commonly referred to as *security through obscurity*. With this model, a system is presumed to be secure simply because (supposedly) nobody knows about its existence, contents, security measures, or anything else. This approach seldom works for long; there are too many ways to find an attractive target. One of the authors had a system that had been connected to the Internet for only about an hour before someone attempted to break in. Luckily, the operating system that was in the process of being installed detected, denied, and logged the access attempts.

Many people assume that even though attackers can find them, the attackers won't bother to. They figure that a small company or a home machine just isn't going to be of interest to intruders. In fact, many intruders aren't aiming at particular targets; they just want to break into as many machines as possible. To them, small companies and home machines simply look like easy targets. They probably won't stay long, but they will attempt to break in, and they may do considerable damage. They may also use compromised machines as platforms to attack other sites.

To function on any network, the Internet included, a site has to have at least a minimal amount of registration, and much of this registration information is available to anyone, just for the asking. Every time a site uses services on the network, someone (at the very least, whoever is providing the service) will know they are there. Intruders watch for new connections in the hope that these sites won't yet have security measures in place. Some sites have reported automated probes apparently based on new site registrations.

There are many different ways someone can determine security-sensitive information from a site. For example, knowing what hardware and software a site has and what version of the operating system it is running gives intruders important clues about what security holes they might try. They can often get this information from your home registration, or by trying to connect to your computer. Many computers disclose their type of operating system in the greeting you get before log in, so an intruder doesn't need access to get it.



NATs security through obscurity:

- Misconfigured wireless NAT's allows hackers to access Internet with your connection.
- All actions appear to come from your internet connection.
- Hackers may also be able to access computers inside Internal network.

— Hacker accesses Internet with your public Internet address

— Hacker controls internal computers

FIGURE 4.8 An attack on a network using the security through obscurity approach

In addition, you send out all sorts of information when you deal with other sites on the Internet. Whenever you visit a website, you tell that site what kind of browser you are running, and reveal the kind of machine you are using. Some email programs include this information in every piece of mail sent out.

Even if you manage to suppress all of these visible sources of information, intruders have scripts as programs that let them use much subtler clues. Although the Internet operates according to standards, there are always loopholes and questionable situations. Different computers do different things when presented with exceptional situations, and intruders can learn much by creating these situations to observe what happens. Sometimes, it's possible to figure out what kind of machine you're dealing with just by watching the sizes and timing it uses to send out data packets.

If all of that fails, intruders have a lot of time on their hands and can often avoid having to figure out obscure facts by simply trying all the possibilities. In the long run, relying on obscurity is not a smart security choice.

4.14 HOST SECURITY

The most common model for computer security is host security. With this model, the security of each host machine is enforced separately, and every effort is made to avoid or alleviate all the known security problems that might affect that particular host. What's wrong with host security? It's not that it doesn't work on any individual machine—the problem is that it doesn't scale to a large number of machines.

The major impediment to effective host security in the modern computing environment is the complexity and diversity of those environments. Most modern environments include machines from multiple vendors, each with its own operating system, and each with its own security problems. Even if the site has machines from only one vendor, different releases of the same operating system often have significantly different security problems. Even if all these machines are from a single vendor and run a single release of the operating system, different configuration (different services enabled, and so on) can bring different subsystems into play (and into conflict) and lead to different sets of security. Even with all that work done correctly, host security still often fails due to bugs in the vendor's software or due to a lack of suitably secure software for some required functions.

Host security also relies on the good intentions and the skill of everyone who has privileged access to any machine. As the number of machines increases, the number of privileged users generally increases as well. Securing a machine is much more difficult than attaching it to a network, so insecure machines may appear on a network as unexpected surprises. The mere fact that it is not supposed to be possible to buy or connect machines without consulting tech support is immaterial; people develop truly innovative purchasing and network-connection schemes if they feel the need.

A host security model may be appropriate for small sites or sites with extreme security requirements. Indeed, all sites should include some level of host security in their overall security plans. Even if you adopt a network security model as we describe in the next section, certain systems in your configuration will benefit from the strongest host security. For example, even if you have built a firewall around your internal network and systems, certain

systems exposed to the outside world will need host security. (We discuss this in detail in Chapter 10, “*Bastion Hosts*.”) However, the host security model alone just isn’t cost-effective for any but small or simple sites; making it work requires too many restrictions and too many people.

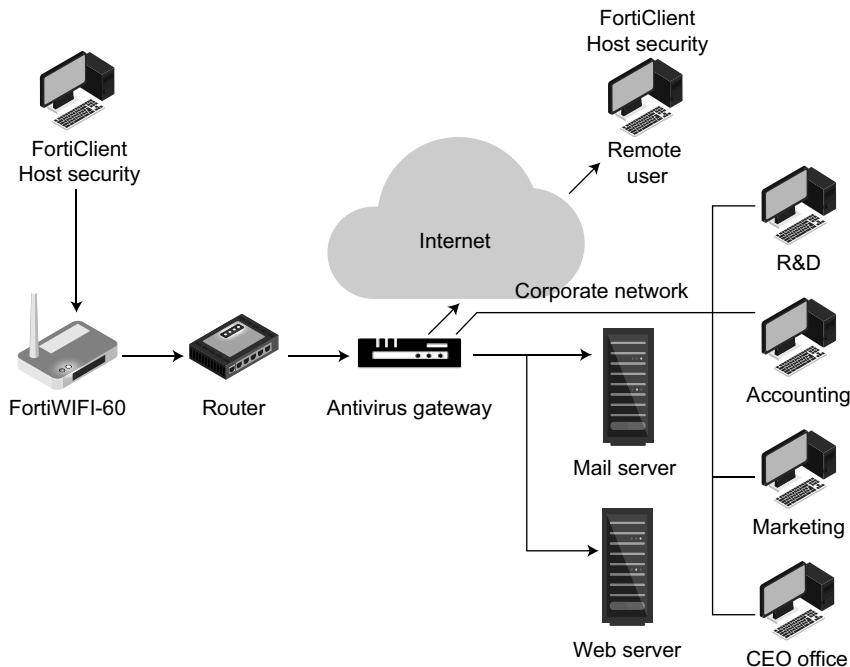


FIGURE 4.9 An example of host security

4.15 NETWORK SECURITY MODEL

As environments grow larger and more diverse, and as securing them on a host-by-host basis grows more difficult, more sites are turning to a network security model. With a network security model, you concentrate on controlling network access to your various hosts and the services they offer, rather than on securing them one by one. Network security approaches include building firewalls to protect your internal systems and networks and using encryption to protect particularly sensitive data as it transits the network.

A site can get tremendous leverage from its security efforts by using a network security model. For example, a single network firewall of the type we

discuss in this book can protect hundreds, thousands, or even tens of thousands of machines against attack from networks beyond the firewall, regardless of the level of host security of the individual machines.

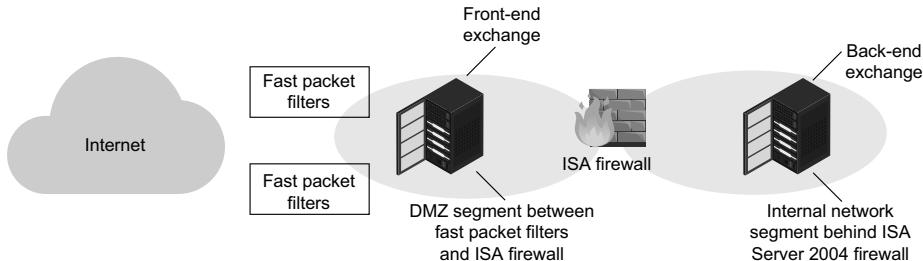


FIGURE 4.10 VAN with firewall security

This kind of leverage depends on the ability to control the access points to the network. At sites that are very large or very distributed, it may be impossible for one group of people to even identify all those access points, much less control them. At that point, the network security model is no longer sufficient, and it's necessary to use layered security, combining a variety of different security approaches.

Although this book concentrates on network security, please note that we are not suggesting you ignore host security. As mentioned previously, you should apply the strongest possible host security measures to your most important machine, especially to those machines that are directly connected to the Internet (this is discussed in more detail in Chapter 10). You will also want to consider using host security internal machine in general, to address security problems other than attacks from the Internet.

4.15.1 No Security Model Can Do It All

No security model can solve all your problems. No security model—short of a maximum security prison—can prevent a person with legitimate access from purposefully damaging your site or taking confidential information from it. To get around powerful host and network security measures, legitimate users can simply use physical methods. These may range from pouring soda into a computer to sending sensitive memos home. You can protect yourself from accidents and ignorance internally, and from malicious external acts, but you cannot protect yourself from your legitimate users without severely damaging

their ability to use their computer. Spies succeed in breaching government security with the ability to use their computers despite regulations and precautions well beyond the resources and tolerance of civilians.

No security model can take care of management problems; computer security will not keep people from wasting time, annoying each other, or embarrassing you. Sites often get trapped into trying to make security protect against these things. When people are wasting time surfing the web, annoying each other by playing tricks with window systems, and embarrassing the company with horrible emails, computer security looks like a promising technological solution that avoids difficult issues. However tempting this may be, a security model wouldn't work here. It is expensive and difficult to even try to solve these problems with computer security, and you are once again in the impossible situation of trying to protect yourself from legitimate users.

No security model provides perfect protection. You can expect to make break-ins rare, brief, and inexpensive, but you can't expect to avoid them altogether. Even the most secure and dedicated sites expect to have a security incident every few years.

You can impress a security expert by saying you have been broken into only once in the last five years; if you say you have never had a breach, they stop being impressed and decide that either you can't detect break-ins or you haven't been around long enough for anyone to seriously try. Security may not prevent every single incident, but it can keep an incident from seriously damaging or even shutting down your business. At one high profile company with multiple computer facilities, a manager complained that his computer facility was supposed to be the most secure, but it got broken into along with several others. The difference was that the break-in was the first one that year for his facility; the intruder was present for only 8 minutes and the computer facility was off the Internet for only 12 hours (from 6 pm to 6 am), after which it resumed business as usual with no visible interruption in service to the company's customers. For one of the other facilities, it was the fourth time; the intruder was present for months before being detected; recovery required taking the facility down for four days; and they had to inform customers that they had shipped them tapes containing possibly contaminated software. Proper security made the difference between an annoying occurrence and a devastating one.

4.15.2 Internet Firewalls

Firewalls are a very effective type of network security. This section briefly describes what Internet firewalls can do for your overall site security. A

firewall is designed to keep a fire from spreading from one part of the building to another. In theory, an Internet firewall serves a similar purpose: It prevents the dangers of the Internet from spreading to your internal network. In practice, a firewall is more like a moat of a medieval castle than a firewall in a modern building. It serves multiple purposes:

- It restricts people to entering at a carefully controlled point.
- It prevents attackers from getting close to your other defenses.
- It restricts people to working at a carefully controlled point.

A firewall is most often installed at the point where your protected internal network connects to the Internet, as shown in Figure 4.11.

All traffic coming from the Internet or going out from your internal network passes through the firewall. Because the traffic passes through it, the firewall has the opportunity to make sure that this traffic is acceptable.

What does “acceptable” means to the firewall? It means that whatever is being done—email, file transfers, remote logins, or any kinds of specific interactions between specific systems—conforms to the security policy of the site.

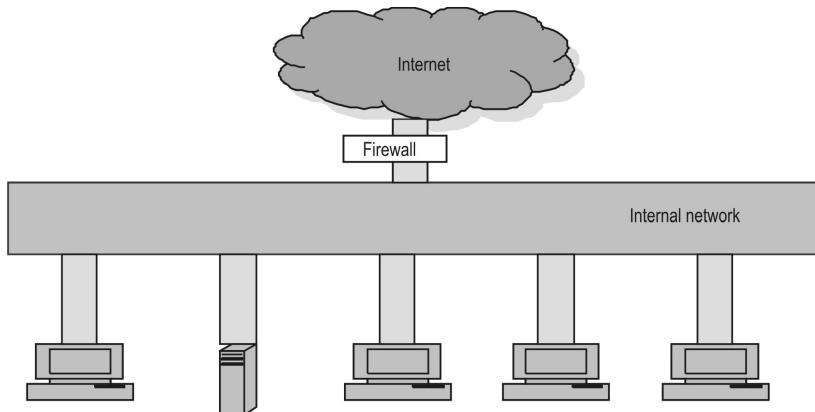


FIGURE 4.11 A firewall usually separates an internal network from the Internet

Logically, a firewall is a separator, a restrictor, and an analyzer. The physical implementation of a firewall varies from site to site. Most often, a firewall is a set of hardware components: a router, a host computer, or some combination of routers, computers, and networks with the appropriate software. There are various ways to configure this equipment; the configuration will depend upon a site's particular security policy, budget, and overall operations.

A firewall is very rarely a single physical object, although some commercial products attempt to put everything into the same box. Usually, a firewall

has multiple parts, and some of these parts may do other tasks besides function as part of the firewall. Your Internet connection is almost always part of your firewall. Even if you have a firewall in a box, it is not going to be neatly separable from the rest of your site; it's not something you can just drop in.

We have compared a firewall to the moat of a medieval castle, and like a moat, a firewall is not invulnerable. It doesn't protect against people who are already inside. It works best if coupled with internal defenses. Even if you stock your moat with alligators, people sometimes manage to swim across. A firewall is also not without its drawbacks; building one requires significant expense and effort, and the restrictions it places on insiders can be a major annoyance.

Given the limitations and drawbacks of firewalls, why would anybody bother to install one? Because a firewall is the most effective way to connect a network to the Internet and still protect that network. The Internet presents marvelous opportunities. Millions of people are out there exchanging information. The benefits are obvious: the chances for publicity, customer service, and information gathering. The popularity of the information superhighway is increasing everybody's desire to be on it. The risks should also be obvious: anytime you get millions of people together, you get crime; it's true in a city, and it's true on the Internet. Any superhighway is fun only while you are in a car. If you have to live or work by the highway, it's loud, smelly, and dangerous.

How can you benefit from the good parts of the Internet without being overwhelmed by the bad? Just as you would like to drive on a highway without suffering the nasty effects of putting a freeway off-ramp into your living room, you need to carefully control the contact that your network has to the Internet. A firewall is a tool for doing that, and in most situations, it's the single most effective tool for doing that.

There are other uses of firewalls. For example, they can be used to divide parts of a site from each other when these parts have distinct security needs (and we'll discuss these uses, as appropriate). The focus of this book, however, is on firewalls as they are used between a site and the Internet.

Firewalls offer significant benefits, but they can't solve every security problem. The following sections briefly summarize what firewalls can and cannot do to protect your systems and your data.

4.15.2.1 What Can a Firewall Do?

Firewalls can do a lot for a site's security. Intact, some advantages of using firewalls extend even beyond security, as described in the sections that follow. A firewall is the focus for security decisions. Think of a firewall as a choke point. All traffic in and out must pass through this single, narrow choke point.

A firewall gives you an enormous amount of leverage for network security because it lets you concentrate your security measures on this choke point: the point where your network connects to the Internet.

Focusing on security in this way is far more efficient than spreading security decisions and technologies around, trying to cover all the bases in a piecemeal fashion. Although firewalls can cost tens of thousands of dollars to implement, most sites find that concentrating on the most defective security hardware and software at the firewall is less expensive and more effective than other security measures—and certainly less expensive than having inadequate security. Firewalls can enforce a security policy.

Many of the services that people want from the Internet are inherently insecure. The firewall is the traffic cop for these services. It enforces the site's security policy, allowing only "approved" services to pass through and those only within the rules set up for them.

For example, one site's management may decide that certain services are simply too risky to be used across the firewall, no matter what system tries to run them or what user wants them. The firewall will keep potentially dangerous services strictly inside the firewall. (There, they can still be used for insiders to attack each other, but that's outside of the firewall's control.) Another site might decide that only one internal system can communicate with the outside world. Still another site might decide to allow access from all systems of a certain type or belonging to a certain group. The variations in site security policies are endless.

A firewall may be called upon to help enforce more complicated policies. For example, perhaps only certain systems within the firewall are allowed to transfer files to and from the Internet; by using other mechanisms to control which users have access to those systems, you can control which users have these capabilities. Depending on the technologies you choose to implement your firewall, a firewall may have a greater or lesser ability to enforce such policies.

4.16 A FIREWALL CAN LOG INTERNET ACTIVITY EFFICIENTLY

Because all traffic passes through the firewall, the firewall provides a good place to collect information about the system and network use and misuse. As a single point of access, the firewall can record what occurs between the protected network and the external network.

4.17 A FIREWALL LIMITS YOUR EXPOSURE

Although this point is most relevant to the use of internal firewalls, which we describe in “Firewall Architectures,” it’s worth mentioning here. Sometimes, a firewall will be used to keep problems that impact one section from spreading through the entire network. In some cases, you’ll do this because one section is more trusted than another—in other cases, because one section is more sensitive than another. Whatever the reason, the existence of the firewall limits the damage that a network security problem can do to the overall network.

WHAT CAN’T A FIREWALL DO?

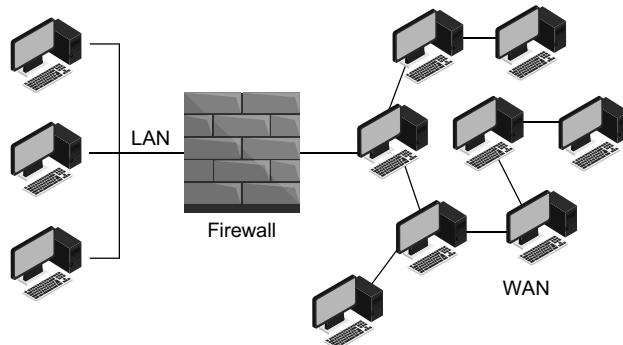


FIGURE 4.12 A firewall limits your network’s exposure

Firewalls offer excellent protection against network threats, but they aren’t a complete security solution. Certain threats are outside the control of the firewall. You need to figure out other ways to protect against these threats by incorporating physical security, host security, and user education into your overall security plan. Some of the weaknesses of firewalls are discussed in the sections that follow.

4.18 A FIREWALL CAN’T PROTECT AGAINST MALICIOUS INSIDERS

A firewall might keep a system user from being able to send proprietary information out of an organization over a network connection; so would simply not have a network connection. But that same user could copy the data onto disk, tape, or paper and carry it out of the building in his or her briefcase.

If the attacker is already inside the firewall—if the fox is inside the henhouse—a firewall can do virtually nothing. Inside users can steal data, damage hardware and software, and subtly modify programs without ever coming near the firewall. Insider threats require internal security measures, such as host security and user education.

4.19 A FIREWALL CAN'T PROTECT CONNECTIONS THAT DON'T GO THROUGH IT

A firewall can effectively control the traffic that passes through it; however, there is nothing a firewall can do about traffic that doesn't pass through it. For example, what if the site allows dial-in access to internal systems behind the firewall? The firewall has absolutely no way of preventing an intruder from getting in through such a modem.

Sometimes, technically expert users or system administrators set up their own “back doors” into the network (such as a dial-up modem connection), either temporarily or permanently, because they chafe at the restrictions that the firewall places upon them and their systems. The firewall can do nothing about this. It's really a people-management problem, not a technical problem.

4.20 A FIREWALL CAN'T PROTECT AGAINST NEW THREATS

A firewall is designed to protect against known threats. A well-designed one may also protect against some new threats. (For example, by denying any but a few trusted services, a firewall will prevent people from setting up new and insecure services.) However, no firewall can automatically defend against every new threat that arises. People continuously discover new ways to attack, using previously trustworthy services, or using attacks that simply had not occur to anyone before. You can't set up a firewall once and expect it to protect your systems forever.

4.21 A FIREWALL CAN'T FULLY PROTECT AGAINST VIRUSES

Firewalls can't keep computer viruses out of a network. It's true that all firewalls scan incoming traffic to some degree, and some firewalls even offer virus protection. However, firewalls don't offer very good virus protection.

Detecting a virus in a random packet of data passing through a firewall is very difficult, and it requires

- recognizing that the packet is part of a program
- determining what the program should look like
- determining that a change in the program is because of a virus

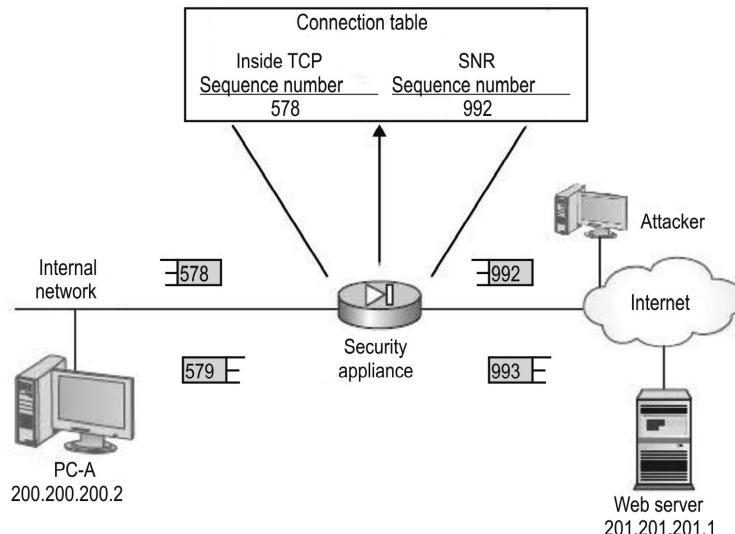


FIGURE 4.13 Firewalls can't fully protect against viruses

Even the first of these is a challenge. Most firewalls are protecting machines of multiple types with different executable formats. A program may be a compiled executable or a script (*e.g.*, a Unix shell script or a Microsoft batch file), and many machines support multiple, compiled executable types. Furthermore, most programs are packaged for transport and are often compressed as well. Packages being transferred via e-mail or Usenet news are also encoded into ASCII in different ways.

For all of these reasons, users may end up bringing viruses behind the firewall, no matter how secure that firewall is. Even if you could do a perfect job of blocking viruses at the firewall, however, you still haven't addressed the virus problem. You have done nothing about the other sources of viruses: Software downloaded from dial-up bulletin-board systems, software brought in on floppies from home or other sites, and even software that comes pre-infected from manufacturers are just as common as virus-infected software on the Internet. Whatever you do to address those threats will also address the problem of software transfer through the firewall.

The most practical way to address the virus problem is through host-based virus protection software and user education concerning the dangers of viruses and precautions to take against them. Virus filtering on the firewall may be a useful adjunct to other precautions, but it will never completely solve the problem.

4.22 A FIREWALL CAN'T SET ITSELF UP CORRECTLY

Every firewall needs some amount of configuration. Every site is slightly different, and it is just not possible for a firewall to magically work correctly when you use it out of the box. The correct configuration is absolutely essential. A misconfigured firewall may provide only the illusion of security. There's nothing wrong with illusions, as long as they are confusing to other sites. A burglar alarm system that consists entirely of some impressive warning stickers and a flashing red light can actually be effective, as long as you don't believe that there's anything else going on. But you know better than to use it on network security, where the warning stickers and the flashing red light are going to be invisible. Unfortunately, many people have firewalls that are no more effective than that, because they have been configured with fundamental problems. A firewall is not a magical protection device that will fix your network security problems no matter what you do with it, and treating it as if it is such a device will merely increase your risk.

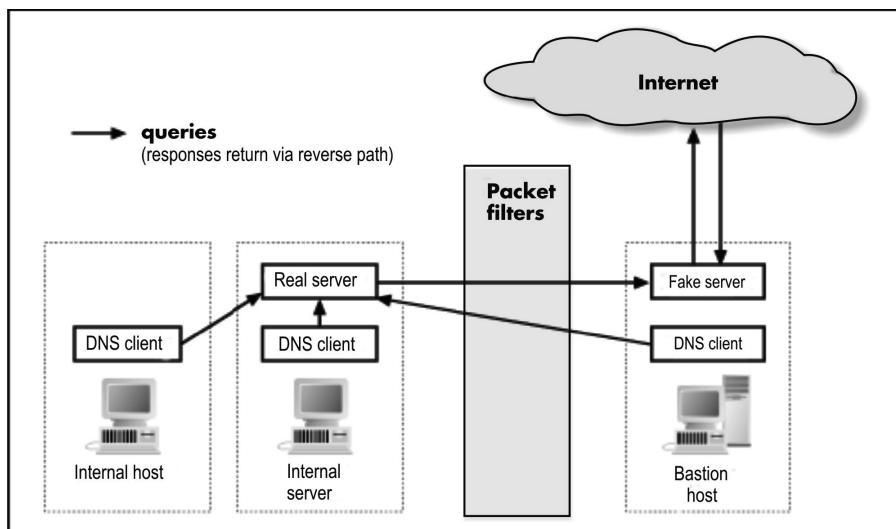


FIGURE 4.14 A firewall needs to be set up correctly to be effective

4.22.1 What's Wrong with Firewalls?

There are two main arguments against using firewalls:

- Firewalls interfere with the way the Internet is supposed to work, introducing all sorts of problems, annoying users, and slowing down the introduction of new Internet services.
- The problems firewalls don't deal with (internal threats and external connections that don't go through the firewall) are more important than the problems they do deal with.

4.23 FIREWALLS INTERFERE WITH THE INTERNET

The Internet is based on a model of end-to-end communication, where individual hosts talk to each other. Firewalls interrupt that end-to-end communication in a variety of ways. Most of the problems that are introduced are the same sorts of problems that are introduced by any security measure. Things are slowed: things that you want to get through can't and it's hard to introduce changes. Having badge readers on doors introduces the same sorts of problems. (You have to swipe the badge and wait for the door to open; when your friends come to meet you they can't get in; and new employees have to get badges.) The difference is that on the Internet there's political and emotional attachment to the idea that information is supposed to flow freely and change is supposed to move rapidly. People are much less willing to accept the sorts of restrictions that they're accustomed to in other environments.

Furthermore, it's truly very annoying to have side effects. There are a number of ways of doing things that provide real advantages and are limited in their spread by firewalls, despite the fact that they aren't security problems. For instance, broadcasting audio and video over the Internet is much easier if you can get precise information about the capabilities of the destination host and the links between you and it. However, firewalls have difficulty managing the connections, and they unintentionally destroy other information. If you're trying to develop new ways of interacting over the Internet, firewalls are incredibly frustrating; everywhere you turn, there's something "cool" that TCP/IP is supposed to be able to do that just doesn't work in the real world. It's no wonder that application developers hate firewalls.

Unfortunately, they don't have any better suggestions for how to keep attackers out. Think how many marvelous things you could have if you didn't

have to lock your front door to keep strangers out; you wouldn't have to sit at home waiting for the repairman or for a package to be delivered. The need for security is unavoidable in our world, and it limits what we can do. The development of the Internet has not changed human nature.

4.24 FIREWALLS DON'T DEAL WITH THE REAL PROBLEM

Some people say that firewalls are passé because they don't deal with real problems. It's true that firewall or no firewall, intruders get in, data goes out, and bad things happen. At sites with really good firewalls, these things occur by avoiding the firewalls. At sites that don't have really good firewalls, these things may go on through the firewalls. Either way, you can argue that this shows that firewalls don't solve the problems.

It's perfectly true; firewalls won't solve all of your security problems. However, the people who point this out don't really have anything better to offer. Protecting individual hosts works for some sites and will help the firewall almost anywhere; detecting and dealing with attacks via network monitoring, once again, will work for some problems and will help a firewall almost anywhere. That's basically the entire list of available alternatives. If you look closely at most of the things promoted as being "better than firewalls," you will discover that they are lightly disguised firewalls marketed by people with restrictive definitions of what a firewall is.

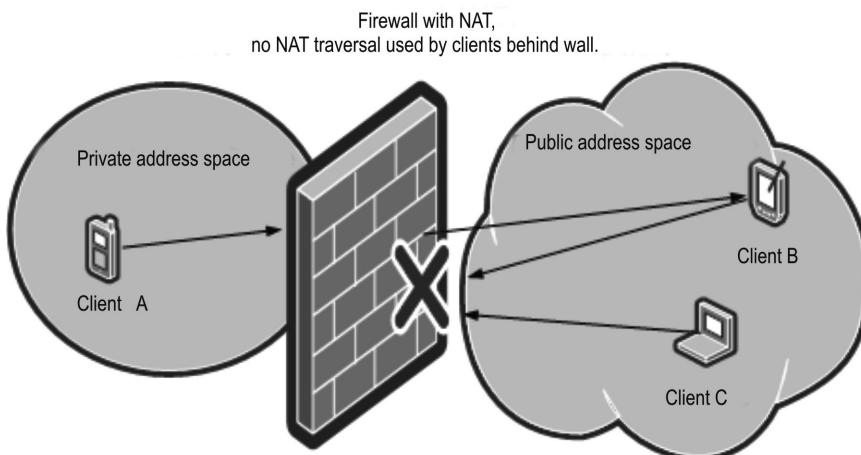


FIGURE 4.15 Firewalls can't solve all security problems

4.24.1 Philosophical Arguments

The world is full of philosophical debates on which people hold strong and divisive beliefs. Firewalls are no exception to this rule.

4.25 BUYING VERSUS BUILDING

Initially, if a site needed a firewall, the administrator had little choice but to design and build it themselves (perhaps with their own staff, or perhaps by hiring a consultant or contractor). Over the years, however, more commercial firewalls offerings have reached the market. These products continue to grow in number and functionality at an astounding rate, and many sites may find that one of these products suits their needs. Most sites find that commercial products are at least a valuable component of their firewall solution.

In deciding whether or not a particular commercial firewall product will meet your needs, you have to understand what your needs are. Even if you decide to buy a firewall, you still need to understand a fair bit about how they're built and how much or more effort evaluating commercial firewall products as they would building their own firewall.

We're not saying that no one should buy a firewall or that everyone should build their own. It's just not necessarily any easier to buy a firewall than it is to build one. It all depends on your particular situation and what resources you have at your disposal. Sites with money to spend but little staff time or expertise available often find buying an attractive solution, while sites with expertise and time, but little money often find building more attractive.

Just what expertise do you need to design and build your own firewall? Like everything else, it depends; it depends on what services you want to provide, what platforms you're using, and what your security concerns are. To install most of the tools described in this book, you need basic Internet skills to obtain the tools and basic system administration skills to configure, compile, and install them. If you don't know what those skills are, you probably don't have them; you can obtain them, but that's beyond the scope of this book.

Some people feel uncomfortable using software that's freely available on the Internet, particularly for security-critical applications. We feel that the advantages outweigh the disadvantages. You may not have the "guarantees" offered by vendors, but you have the ability to inspect the source code and share information with the large community that helps to maintain the

software. In practice, vendors come and go, but the community endures. The packages we discuss in this book are widely used; many of the largest sites on the Internet waste their firewalls on them. These packages reflect years of real-life experience with the Internet and its risks.

Other people feel uncomfortable using commercial software for security-critical applications, feeling that you can't trust software unless you can read the code. While there are real advantages to having code available, auditing code is difficult, and few people can do an adequate job on a package of any significant size. Commercial software has its own advantages. When you buy software, you have a legal contract with a vendor, which may give you some recourse if things go wrong.

Frequently, people argue that open source software is more risky than commercial software because attackers have access to the source code. In practice, the attackers have access to all the source code they need, including commercial source code. If it is not given to them, they steal or reverse-engineer it, they have the motivation and time, and they don't have ethical constraints. There is no distinction between programs on this point.

While it's possible to build a firewall consisting solely of freely available software or solely of commercial software, there's no reason to feel that it's all or nothing; freely available tools provide a valuable complement to purchased solutions. Buying a firewall shouldn't make you reluctant to supplement with fairly available tools, and building one shouldn't make you reluctant to supplement with purchased tools. Don't rule out a product just because it's commercial or just because it's freely available. Truly excellent products with great support appear in both categories, as do poorly thought out products with no support.

Software, Freedom, and Money

A number of terms are used for various kind of software that you may (or may not) be able to use without paying money to anybody:

Free Software

This term is unfortunately ambiguous; sometimes it means software that you don't have to pay for and sometimes it refers to software that has been liberated from certain points of constraints by very carefully tying it up with others. In practice, you can't be sure that it means anything at all, although it strongly implies that you will be able to use the software without paying for it (but not necessarily resell it in any form).

Freely Available Software

This term clearly means software that you don't have to pay for, although it is sometimes used for software that only some classes of users have to pay for (for instance, software that is free to individuals but costs money for corporations).

Public Domain Software

Although this term is often carelessly used, it has a specific legal meaning and refers to software that is free of copyright restrictions and may be used in any way what so ever without the permission of the author. Software is public domain only if it is clearly marked as such; software that contains a copyright notice or use restrictions is not public domain. You may copy public domain software without paying for it.

EXERCISES

1. What is a firewall?
2. Explain how a firewall works.
3. Explain how firewalls interfere with the Internet.
4. What is a protective device?
5. What is a scorekeeper?
6. Explain security through obscurity.

CHAPTER 5

PUBLIC KEY CERTIFICATES

Chapter Goals

- Understand the types of attacks that may be used by hackers to undermine network security.
- Understand the types of vulnerabilities that may be present in your network.
- Learn to classify the different types of networks and users that may interact with your own, and evaluate their risk factors.
- Learn to evaluate your network topology and requirements, and develop a suitable security policy for implementation.
- Become familiar with the tools available for protecting confidential information and your network.

5.1 SECURITY OBJECTIVES

With the rapid growth of interest in the Internet, network security has become a major concern to companies throughout the world. The fact that the information and tools needed to penetrate the security of corporate networks are widely available has increased that concern.

Because of this increased focus on network security, network administrators often spend more effort protecting their networks than on actual network set-up and administration. Tools that probe for system vulnerabilities, such as the Security Administrator Tool for Analyzing Networks (SATAN), and some of the newly available scanning and intrusion detection packages

and appliances, assist in these efforts, but these tools only point out areas of weakness and may not provide a means to protect networks from all possible attacks. Thus, as a network administrator, you must constantly try to keep abreast of the large number of security issues confronting you in today's world. This chapter describes many of the security issues that arise when connecting a private network to the Internet.

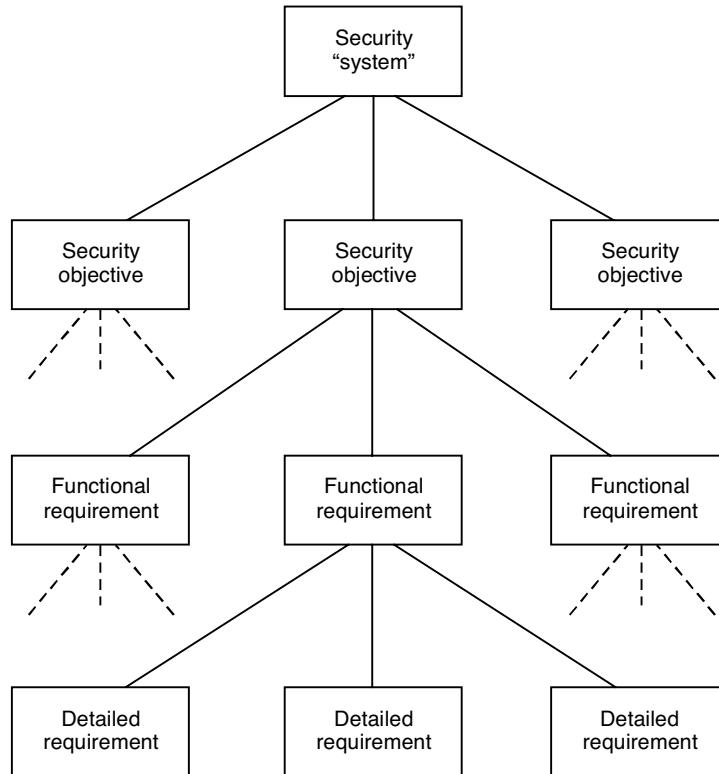


FIGURE 5.1 Security administrator tool for analyzing networks

5.1.1 Security Issues when Connecting to the Internet

When you connect your private network to the Internet, you are physically connecting your network to more than 50,000 unknown networks and all their users. Although such connections open the door to many useful applications and provide great opportunities for information sharing, most private networks contain some information that should not be shared with outside users on the Internet. In addition, not all Internet users are involved in lawful

activities. These two statements foreshadow the key questions behind most security issues on the Internet:

- How do you protect confidential information from those who do not explicitly need to access it?
- How do you protect your network and its resources from malicious users and accidents that originate outside your network?

The Collaborative Internet Security Network

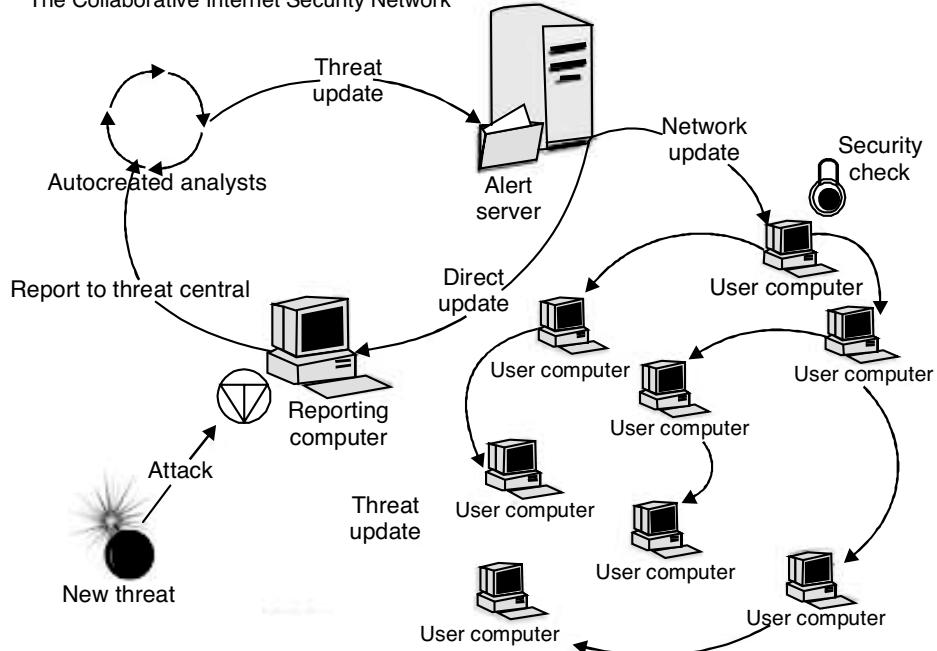
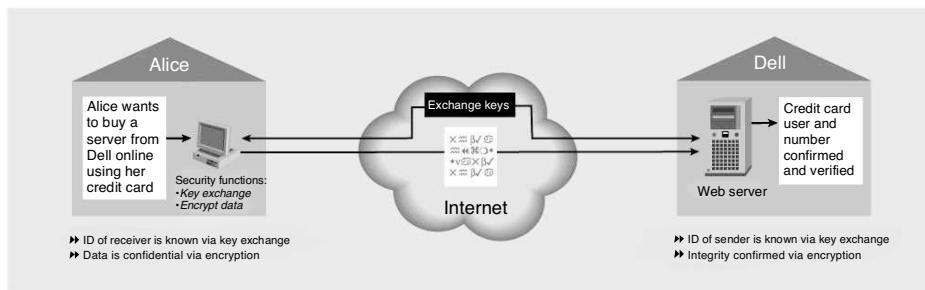


FIGURE 5.2 Security issues when connecting to the Internet

5.1.2 Protecting Confidential Information

Confidential information can reside in two states on a network. It can reside on physical storage media, such as a hard drive or memory, or it can reside in transit across the physical network wire in the form of packets. These two information states present multiple opportunities for attacks from users on your internal network, as well as those users on the Internet. We are primarily concerned with the second state, which involves network security issues. The following are five common methods of attack that present opportunities to compromise the information on your network:

**FIGURE 5.3** Protecting confidential information

- Network packet sniffers
- IP spoofing and denial-of-service attacks
- Password attacks
- Distribution of sensitive internal information to external sources
- Man-in-the-middle attacks

When protecting your information from these attacks, your concern is to prevent the theft, destruction, corruption, and introduction of information that can cause irreparable damage to sensitive and confidential data. This section describes these common methods of attack and provides examples of how your information can be compromised.

Network Packet Sniffers

Because networked computers communicate serially (one information piece is sent after another), large information pieces are broken into smaller pieces. (The information stream would be broken into smaller pieces even if networks communicated in parallel. The overriding reason for breaking streams into network packets is that computers have limited intermediate buffers.) These smaller pieces are called *network packets*. Several network applications distribute network packets in clear text; that is, the information sent across the network is not encrypted. (Encryption is the transformation, or scrambling, of a message into an unreadable format by using a mathematical algorithm). Because the network packets are not encrypted, they can be processed and understood by any application that can pick them up off the network and process them.

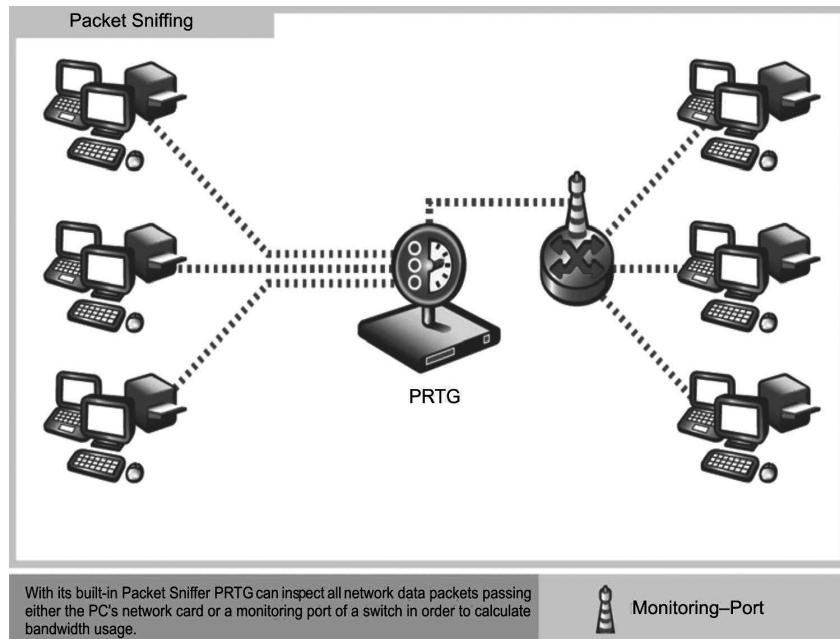


FIGURE 5.4 Network packet sniffers

A network protocol specifies how packets are identified and labeled, which enables a computer to determine whether a packet is intended for it. Because the specifications for network protocols, such as TCP/IP, are widely published, a third party can easily interpret the network packets and develop a packet sniffer. (The real threat today results from the numerous freeware and shareware packet sniffers that are available, which do not require the user to understand anything about the underlying protocols.) A *packet sniffer* is a software application that uses a network adapter card in *promiscuous mode* (a mode in which the network adapter card sends all packets received on the physical network wire to an application for processing) to capture all network packets that are sent across a local-area network.

Because several network applications distribute network packets in clear text, a packet sniffer can provide its user with meaningful and often sensitive information, such as user account names and passwords. If you use networked databases, a packet sniffer can provide an attacker with information that is queried from the database, as well as the user account names and passwords used to access the database. One serious problem with acquiring user account names and passwords is that users often reuse their login names and passwords across multiple applications.

In addition, many network administrators use packet sniffers to diagnose and fix network-related problems. In the course of their usual duties, these network administrators (such as those in the Payroll Department) work during regular employee hours, and so they can potentially examine sensitive information distributed across the network.

Many users employ a single password for access to all accounts and applications. If an application is run in client/server mode and authentication information is sent across the network in clear text, this same authentication information likely can be used to gain access to other corporate resources. Because attackers know and use human characteristics (attack methods known collectively as *social engineering attacks*), such as using a single password for multiple accounts, they are often successful in gaining access to sensitive information.

IP Spoofing and Denial-of-Service Attacks

An *IP spoofing attack* occurs when an attacker outside your network pretends to be a trusted computer. This is facilitated either by using an IP address that is within the range of IP addresses for your network, or by using an authorized external IP address that you trust and to which you want to provide access to specified resources on your network.

Normally, an IP spoofing attack is limited to the injection of data or commands into an existing stream of data passed between a client and server application or a peer-to-peer network connection. To enable bidirectional communication, the attacker must change all routing tables to point to the spoofed IP address.

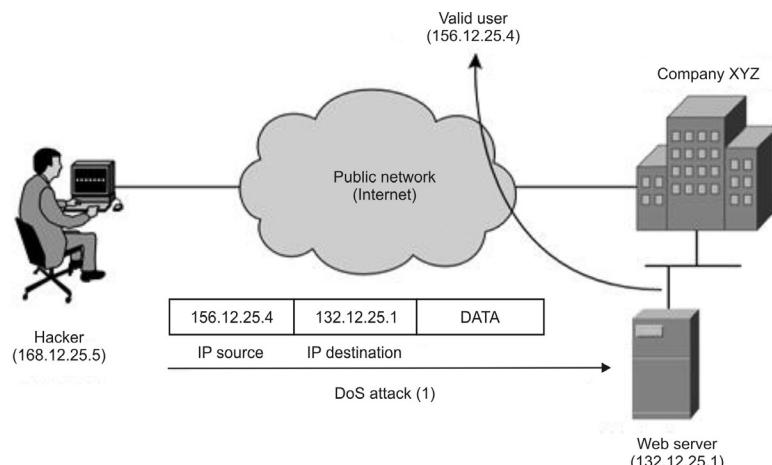


FIGURE 5.5 IP Spoofing and denial-of-service attacks

However, if an attacker manages to change the routing tables to point to the spoofed IP address, he can receive all the network packets that are addressed to the spoofed address and can reply just as any trusted user can.

Another approach that the attacker could take is to not worry about receiving any response from the targeted host. This is called a *Denial-of-Service (DOS)* attack. The denial-of-service occurs because the system receiving the requests becomes busy trying to establish a return communications path with the initiator (which may or may not be using a valid IP address). In more technical terms, the targeted host receives a TCP SYN and returns a SYN-ACK. It then remains in a wait state, anticipating the completion of the TCP handshake that never happens. Each wait state uses system resources until, eventually, the host cannot respond to other legitimate requests.

Like packet sniffers, IP spoofing and DOS attacks are not restricted to people who are external to the network.

Password Attacks

Password attacks can be implemented using several different methods, including brute-force attacks, Trojan horse programs (discussed later in the chapter), IP spoofing, and packet sniffers. Although packet sniffers and IP spoofing can yield user accounts and passwords, password attacks usually refer to repeated attempts to identify a user account and/or password; these repeated attempts are called *brute-force attacks*.



Password attacks (3)

- Dictionary attacks (UNIX Crack, L0pht Crack for Windows NT)

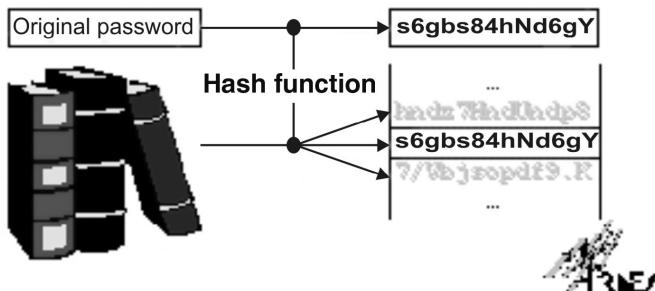


FIGURE 5.6 Password attacks

Often, a brute-force attack is performed using a dictionary program that runs across the network and attempts to log in to a shared resource, such as a server. When an attacker successfully gains access to a resource, that person has the same rights as the user whose account has been compromised to gain access to that resource. If this account has sufficient privileges, the attacker can create a back door for future access, without concern for any status and password changes to the compromised user account.

Distribution of Sensitive Internal Information to External Sources

Controlling the distribution of sensitive information is at the core of a network security policy. Although such an attack may not seem obvious to you, the majority of computer break-ins that organizations suffer are at the hands of disgruntled present or former employees. At the core of these security breaches is the distribution of sensitive information to competitors or others that will use it to your disadvantage. An outside intruder can use password and IP spoofing attacks to copy information, and an internal user can easily place sensitive information on an external computer or share a drive on the network with other users.

For example, an internal user could place a file on an external FTP server without ever leaving his or her desk. The user could also E-mail an attachment that contains sensitive information to an external user.

Man-in-the-Middle Attacks

A man-in-the-middle attack requires that the attacker has access to the network packets that come across the networks. An example of such a configuration could be someone who is working for your Internet Service Provider (ISP), who can gain access to all network packets transferred between your network and any other network. Such attacks are often implemented using network packet sniffers and routing and transport protocols. The possible uses of such attacks are the theft of information, hijacking of an ongoing session to gain access to your internal network resources, traffic analysis to derive information about your network and its users, denial-of-service, corruption of transmitted data, and introduction of new information into network sessions.

5.2 PROTECTING YOUR NETWORK: MAINTAINING INTERNAL NETWORK SYSTEM INTEGRITY

Although protecting your information may be your highest priority, protecting the integrity of your network is critical to protect the information it contains. A breach in the integrity of your network can be extremely costly in terms of

time and effort, and it can open multiple avenues for continued attacks. This section covers the five methods of attack that are commonly used to compromise the integrity of your network:

- Network packet sniffers
- IP spoofing
- Password attacks
- Denial-of-service attacks
- Application layer attacks

When considering what to protect within your network, you are concerned with maintaining the integrity of the physical network, your network software, any other network resources, and your reputation. This integrity involves the verifiable identity of computers and users, proper operation of the services that your network provides, and optimal network performance; all these concerns are important in maintaining a productive network environment. This section provides some examples of the attacks described previously and explains how they can be used to compromise your network's integrity.

5.2.1 Network Packet Sniffers

As mentioned earlier, network packet sniffers can yield critical system information, such as user account information and passwords. When an attacker obtains the correct account information, he or she has the run of your network. In a worst-case scenario, an attacker gains access to a system-level user account, which the attacker uses to create a new account that can be used at any time as a back door to get into your network and its resources. The attacker can modify system-critical files, such as the password for the system administrator account, the list of services and permissions on file servers, and the login details for other computers that contain confidential information.

Packet sniffers provide information about the topology of your network that many attackers find useful. This information, such as what computers run which services, how many computers are on your network, which computers have access to others, and so on, can be deduced from the information contained within the packets that are distributed across your network as part of necessary daily operations.

In addition, a network packet sniffer can be modified to interject new information or change existing information in a packet. By doing so, the attacker can cause network connections to shut down prematurely, as well

as change critical information within the packet. Imagine what could happen if an attacker modified the information being transmitted to your accounting system. The effects of such attacks can be difficult to detect and very costly to correct.

5.2.2 IP Spoofing

IP spoofing can yield access to user accounts and passwords, and it can also be used in other ways. For example, an attacker can emulate one of your internal users in ways that prove embarrassing for your organization; the attacker could send email messages to business partners that appear to have originated from someone within your organization. Such attacks are easier when an attacker has a user account and password, but they are possible by combining simple spoofing attacks with knowledge of messaging protocols. For example, Telnetting directly to the SMTP port on a system allows the attacker to insert bogus sender information.

5.2.3 Password Attacks

Just as with packet sniffers and IP spoofing attacks, a brute-force password attack can provide access to accounts that can be used to modify critical network files and services. An example of an attack that compromises your network's integrity occurs when an attacker modifies the routing tables for your network. By doing so, the attacker ensures that all the network packets are routed to him or her before they are transmitted to their final destination. In such a case, an attacker can monitor all network traffic, effectively becoming a man in the middle.

5.2.4 Denial-of-Service Attacks

Denial-of-service attacks are different from most other attacks because they are not targeted at gaining access to your network or the information on your network. These attacks focus on making a service unavailable for normal use, which is typically accomplished by exhausting some resource limitation on the network or within an operating system or application.

When involving specific network server applications, such as a Hypertext Transfer Protocol (HTTP) server or a File Transfer Protocol (FTP) server, these attacks can focus on acquiring and keeping open all the available connections supported by that server, effectively locking out valid users of the server or service. Denial-of-service attacks can also be implemented using common Internet protocols, such as TCP and Internet Control Message Protocol (ICMP). Most denial-of-service attacks exploit a weakness in the overall

architecture of the system being attacked rather than a software bug or security hole. However, some attacks compromise the performance of your network by flooding the network with undesired and often useless network packets and by providing false information about the status of network resources.

5.2.5 Application Layer Attacks

Application layer attacks can be implemented using several different methods. One of the most common methods is exploiting well-known weaknesses in software commonly found on servers, such as send mail, Post Script, and FTP. By exploiting these weaknesses, attackers can gain access to a computer with the permissions of the account running the application, which is usually a privileged system-level account.

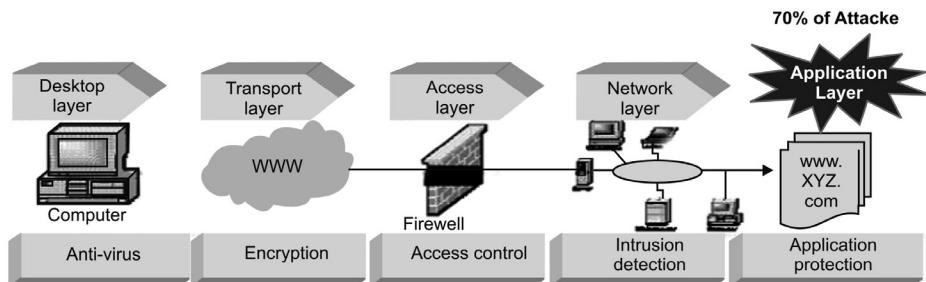


FIGURE 5.7 Application layer attacks

Trojan horse attacks are implemented using bogus programs that an attacker substitutes for common programs. These programs may provide all the functionality that the normal application or service provides, but they also include other features that are known to the attacker, such as monitoring login attempts to capture user account and password information. These programs can capture sensitive information and send it back to the attacker. They can also modify application functionality, such as applying a blind carbon copy to all email messages so that the attacker can read all of an organization's email.

One of the oldest forms of application layer attacks is a Trojan horse program that displays a screen, banner, or prompt that the user believes is the valid login sequence. The program then captures the information that the user types in and stores or emails it to the attacker. Next, the program either forwards the information to the normal login process (normally impossible on modern systems) or simply sends an expected error to the user (for example, bad username/password combination), exits, and starts the normal login sequence. The user, believing that he or she has incorrectly entered the password (a common mistake), retypes the information and is allowed access.

One of the newest forms of application layer attacks exploits the openness of several new technologies: the Hyper Text Markup Language (HTML) specification, web browser functionality, and HTTP. These attacks, which include Java applets and ActiveX controls, involve passing harmful programs across the network and loading them through a user's browser.

Users of ActiveX controls may be lulled into a false sense of security by the Authenticode technology promoted by Microsoft. However, attackers have already discovered how to utilize properly signed and bug-free ActiveX controls to make them act as Trojan horses. This technique uses VBScript to direct the controls to perform their dirty work, such as overwriting files and executing other programs.

These new forms of attack are different in two respects:

- They are initiated not by the attacker, but by the user, who selects the HTML page that contains the harmful applet or script stored using the <OBJECT>, <APPLET>, or <SCRIPT> tags.
- Their attacks are no longer restricted to certain hardware platforms and operating systems because of the portability of the programming languages involved.

5.3 TRUSTED, UNTRUSTED, AND UNKNOWN NETWORKS

As a network manager creates a network security policy, each network that makes up the topology must be classified as one of three types of networks:

- Trusted networks
- Untrusted networks
- Unknown networks

5.3.1 Trusted Networks

Trusted networks are the networks inside your network security perimeter. These networks are the ones that you are trying to protect. Often you or someone in your organization administers the computers that comprise these networks, and your organization controls their security measures. Usually, trusted networks are within the security perimeter.

When you set up the firewall server, you explicitly identify the type of networks that are attached to the firewall server through network adapter cards.

After the initial configuration, the trusted networks include the firewall server and all networks behind it.

One exception to this general rule is the inclusion of virtual private networks (VPNs), which are trusted networks that transmit data across an untrusted network infrastructure. For the purposes of our discussion, the network packets that originate on a VPN are considered to originate from within your internal perimeter network. This origin is logical because of how VPNs are established. For communications that originate on a VPN, security mechanisms must exist by which the firewall server can authenticate the origin, data integrity, and other security principles contained within the network traffic according to the same security principles enforced on your trusted networks.

5.3.2 Untrusted Networks

Untrusted networks are the networks that are known to be outside your security perimeter. They are untrusted because they are outside your control. You have no control over the administration or security policies for these sites. They are the private, shared networks from which you are trying to protect your network. However, you still need and want to communicate with these networks although they are untrusted.

When you set-up the firewall server, you explicitly identify the untrusted networks from which that firewall can accept requests. Untrusted networks are outside the security perimeter and are external to the firewall server.

5.3.3 Unknown Networks

Unknown networks are networks that are neither trusted nor untrusted. They are unknown quantities to the firewall because you cannot explicitly tell the firewall server that the network is a trusted or an untrusted network. Unknown networks exist outside your security perimeter. By default, all non-trusted networks are considered unknown networks, and the firewall applies the security policy that is applied to the Internet node in the user interface, which represents all unknown networks. However, you can identify unknown networks below the Internet node and apply more specialized policies to those untrusted networks.

5.4 ESTABLISHING A SECURITY PERIMETER

When you define a network security policy, you must define procedures to safeguard your network and its contents and users against loss and damage. From this perspective, a network security policy plays a role in enforcing the overall security policy defined by an organization.

A critical part of an overall security solution is a network firewall, which monitors traffic crossing network perimeters and imposes restrictions according to security policy. Perimeter routers are found at any network boundary, such as between private networks, intranets, extranets, or the Internet. Firewalls most commonly separate internal (private) and external (public) networks.

A network security policy focuses on controlling the network traffic and usage. It identifies a network's resources and threats, defines network use and responsibilities, and details action plans for when the security policy is violated. When you deploy a network security policy, you want it to be strategically enforced at defensible boundaries within your network. These strategic boundaries are called *perimeter networks*.

5.5 PERIMETER NETWORKS

To establish your collection of perimeter networks, you must designate the networks of computers that you wish to protect and define the network security mechanisms that protect them. To have a successful network security perimeter, the firewall server must be the gateway for all communications between trusted networks and untrusted and unknown networks.

Each network can contain multiple perimeter networks. When describing how perimeter networks are positioned relative to each other, three types of perimeter networks are present: the outermost perimeter, internal perimeters, and the innermost perimeter. Note that the multiple internal perimeters are relative to a particular asset, such as the internal perimeter that is just inside the firewall server.

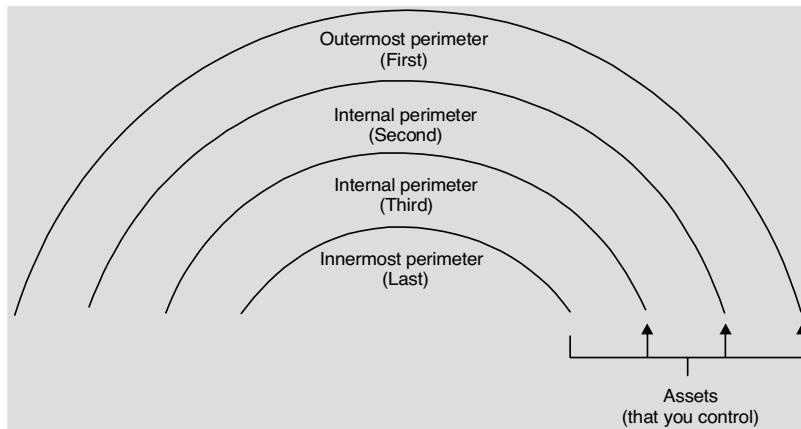


FIGURE 5.8 Three types of perimeter networks exist: outermost, internal, and innermost

The outermost perimeter network identifies the separation point between the assets that you control and the assets that you do not control—usually, this point is the router that you use to separate your network from your ISP's network. Internal perimeter networks represent additional boundaries where you have other security mechanisms in place, such as Intranet firewalls and filtering routers.

Figure 5.9 depicts two perimeter networks (an outermost perimeter network and an internal perimeter network) defined by the placement of the internal and external routers and the firewall server.

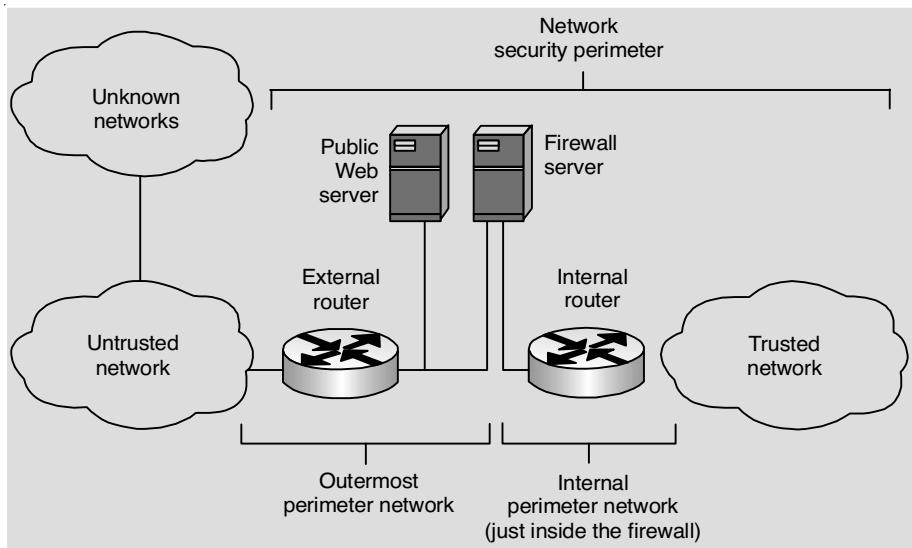


FIGURE 5.9 An example of a two-perimeter network security design

Positioning your firewall between an internal and external router provides little additional protection from attacks on either side, but it greatly reduces the amount of traffic that the firewall server must evaluate, which can increase the firewall's performance. From the perspective of users on an external network, the firewall server represents all accessible computers on the trusted network. It defines the point of focus, or choke point, through which all communications between the two networks must pass.

The outermost perimeter network is the most insecure area of your network infrastructure. Normally, this area is reserved for routers, firewall servers, and public Internet servers, such as HTTP, FTP, and Gopher servers. This area of the network is the easiest area to gain access to and, therefore, is the most frequently attacked, usually in an attempt to gain access to the

internal networks. Sensitive company information that is for internal use only should not be placed on the outermost perimeter network. Following this precaution helps avoid having your sensitive information stolen or damaged.

5.6 DEVELOPING YOUR SECURITY DESIGN

The design of the perimeter network and security policies require the following subjects to be addressed.

5.6.1 Know Your Enemy

Knowing your enemy means knowing attackers or intruders. Consider who might want to circumvent your security measures and identify their motivations. Determine what they might want to do and the damage that they could cause to your network.

Security measures can never make it impossible for a user to perform unauthorized tasks with a computer system; they can only make it harder. The goal is to make sure that the network security controls are beyond the attacker's ability or motivation.

5.6.2 Count the Cost

Security measures usually reduce convenience, especially for sophisticated users. Security can delay work and can create expensive administrative and educational overhead. Security can use significant computing resources and require dedicated hardware.

When you design your security measures, understand their costs and weigh those costs against the potential benefits. To do that, you must understand the costs of the measures themselves and the costs and likelihood of security breaches. If you incur security costs out of proportion to the actual dangers, you have done yourself a disservice.

5.6.3 Identify Any Assumptions

Every security system has underlying assumptions. For example, you might assume that your network is not tapped, that attackers know less than you do, that they are using standard software, or that a locked room is safe. Be sure to examine and justify your assumptions. Any hidden assumption is a potential security hole.

5.6.4 Control Your Secrets

Most security is based on secrets. Passwords and encryption keys, for example, are secrets. Too often, though, the secrets are not all that secret. The most important part of keeping secrets is in knowing the areas that you need to protect. What knowledge would enable someone to circumvent your system? You should jealously guard that knowledge and assume that everything else is known to your adversaries. The more secrets you have, the harder it will be to keep them all. Security systems should be designed so that only a limited number of secrets need to be kept.

5.6.5 Human Factors

Many security procedures fail because their designers do not consider how users will react to them. For example, because they can be difficult to remember, automatically generated nonsense passwords often are written on the undersides of keyboards. For convenience, a secure door that leads to the system's only tape drive is sometimes propped open. For expediency, unauthorized modems are often connected to a network to avoid onerous dial-in security measures.

If your security measures interfere with essential use of the system, those measures will be resisted and perhaps circumvented. To get compliance, you must make sure that users can get their work done, and you must sell your security measures to users. Users must understand and accept the need for security.

Any user can compromise system security, at least to some degree. For instance, passwords can often be found simply by calling legitimate users on the telephone, claiming to be a system administrator, and asking for them. If your users understand security issues, and if they understand the reasons for your security measures, they are far less likely to make an intruder's life easier.

At a minimum, users should be taught never to release passwords or other secrets over unsecured telephone lines (especially cellular telephones) or email. Users should be wary of people who call them on the telephone and ask questions. Some companies have implemented formalized network security training so that employees are not allowed access to the Internet until they have completed a formal training program.

5.6.6 Know Your Weaknesses

Every security system has vulnerabilities. You should understand your system's weak points and know how they could be exploited. You should also know the areas that present the greatest danger and should prevent access

to them immediately. Understanding the weak points is the first step toward turning them into secure areas.

5.6.7 Limit the Scope of Access

You should create appropriate barriers in your system so that if intruders access one part of the system, they do not automatically have access to the rest of the system. The security of a system is only as good as the weakest security level of any single host in the system.

5.6.8 Understand Your Environment

Understanding how your system normally functions, knowing what is expected and what is unexpected, and being familiar with how devices are usually used will help you detect security problems. Noticing unusual events can help you catch intruders before they can damage the system. Auditing tools can help you detect those unusual events.

5.6.9 Limit Your Trust

You should know exactly which software you rely on, and your security system should not have to rely on the assumption that all software is bug-free.

5.6.10 Remember Physical Security

Physical access to a computer (or a router) usually gives a sufficiently sophisticated user total control over that computer. Physical access to a network link usually allows a person to tap that link, jam it, or inject traffic into it. It makes no sense to install complicated software security measures when access to the hardware is not controlled.

5.6.11 Make Security Pervasive

Almost any change that you make in your system may have security effects. This is especially true when new services are created. Administrators, programmers, and users should consider the security implications of every change they make. Understanding the security implications of a change takes practice; it requires lateral thinking and a willingness to explore every way that a service could potentially be manipulated.

5.7 SECURE SOCKETS LAYER

SSL was developed by Netscape to provide security when transmitting information on the Internet. Netscape recognized the need to develop a process that would ensure confidentiality when entering and transmitting information on the Web. Without such a process, very few individuals would feel comfortable entering information like credit card numbers on a website. Netscape recognized that ecommerce on the Web would never get off the ground without consumer confidence. As a result, SSL was developed to address the security needs of Web surfers.

It is somewhat ironic that we require such a high level of security for transactions on the Web. Most knowledgeable individuals would never enter their Visa or Mastercard number on a site that did not employ SSL for fear of having the information intercepted. However, those same individuals would not hesitate to give that same information over the phone to an unknown person when ordering flowers, nor would they fear giving their credit cards to a waiter at a restaurant. Consider that this involves handing a card over to someone you have never met who inevitably disappears for 10 minutes. Where is the security in that exchange? For some reason we hold transactions on the Web to a higher standard of security than we do most other types of transactions. The risk that a credit card number will be stolen in transit on the Internet is very small. A greater risk is that the credit card number will be stolen from a system on which it is stored. That is precisely what happened to me: A while back, I received an email from my ISP informing me that a computer, which had been stolen from the ISP, may have contained credit card information for a number of its customers. The email went on to state that it was possible that my credit card information was on the stolen machine. The company said that the file containing the credit card numbers was encrypted, so it did not believe that there was any real risk. Nevertheless, the firm said that it was advising its customers of this incident so they could take appropriate action. The original transaction with the ISP in which I gave the company my credit card information was not over the Internet. It was a traditional low-tech transaction. Like most companies, the ISP stored the user account information, including credit card numbers, in a database on a network. That is where the real risk lies. SSL utilizes both asymmetric and symmetric key encryption to set-up and transfer data in a secure mode over an unsecured network. When used with a browser client, SSL establishes a secure connection between the client browser and the server. Usually, it's the HTTP over SSL (HTTPS). It sets up an encrypted tunnel between a browser

and a Web server over which data packets can travel. No one tapping into the connection between the browser and the server can decipher the information passing between the two. Integrity of the information is established by hashing algorithms. Confidentiality of the information is ensured with encryption.

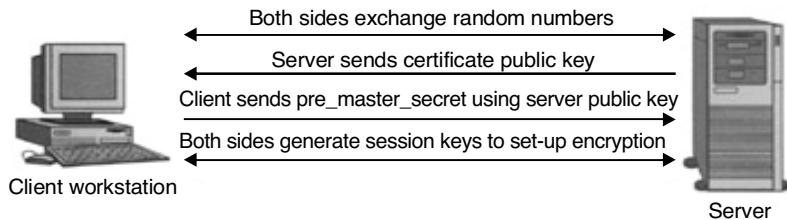


FIGURE 5.10 SSL session hand shake

To set up an SSL session, both sides exchange random numbers. The server sends its public key with a digital certificate signed by a recognized CA attesting to the authenticity of the sender's identity and binding the sender to the public key. The server also sends a session ID. The browser client creates a pre_master_secret key. The client browser encrypts the pre_master_secret key using the server's public key and transmits the encrypted pre_master_secret key to the server. Then, both sides generate a session key using the pre_master_secret and random numbers. The SSL session set-up begins with asymmetric encryption. The server presents the browser client with its public key, which the client uses to encrypt the pre_master_secret. However, once the client sends the encrypted pre_master_secret key back to the server, it employs a session key to establish a secure connection. The initial set-up uses asymmetric encryption, but the two parties switch over to symmetric encryption. This is done because symmetric encryption creates much less overhead. Less overhead means better throughput and a faster response time. Asymmetric cryptosystems are much more CPU-intensive and would significantly slow the exchange of information. As a result, for spontaneous exchanges, asymmetric encryption is used initially to establish a secure connection and to authenticate identities (using digital certificates). Once identities are established and public keys are exchanged, the communicating entities switch to symmetric encryption for efficiency. Even with the use of symmetric encryption, network throughput is significantly diminished with SSL. Cryptographic processing is extremely CPU-intensive. Web servers that would normally be able to handle hundreds of connections may only be able to handle a fraction of that when employing SSL. In 1999, *Internet Week* reported on a test of a Sun 450 server and the effects of SSL. At full capacity, the server could handle

about 500 connections per second of normal HTTP traffic. However, the same server could only handle about three connections per second when the connections employed SSL. The fact that SSL can have such a hindering effect on network performance has to be included in any capacity planning for e-commerce sites. There are SSL accelerators available that can enhance the performance of Web servers that employ SSL. Products from Hewlett-Packard, Compaq, n Cipher, and others offer solutions that speed up the cryptographic processing. Usually, these products are separate boxes that interface with a server and off-load the SSL process from the server's CPU. They can also take the form of accelerator boards that are installed in the server.

5.8 EMAIL SECURITY

The most important thing to remember about standard email is that it is very secure. Email is almost always vulnerable to disclosure in one way or another. Very often, email, by necessity, must traverse many networks to reach its destination. During transit, an email message may pass through many mail servers. As a result, it is vulnerable to interception, replication, disclosure, or modification anywhere along its prescribed path. It can be copied and stored to be retrieved at a later date. In fact, each mail server an email message passes through may be making a copy of the message before it is forwarded. Whether you use your corporation's email system or an email account provided by an ISP, your email messages reside on an email server. Even if you download your email to your local disk drive, those messages probably have already been backed up and stored on some other media. If your corporation uses an email portal to the Internet, then your emails are being copied there before being forwarded onto the appropriate Internet address. The point is that if you think your email is secure and confidential, then you are greatly mistaken. In addition to disclosure, email messages are vulnerable to alteration. Anywhere along the path, an email message can be intercepted and modified before being forwarded. Common email provides no method of detecting messages that have been modified in transit. Messages that have been copied and stored can also be modified and retransmitted at a later time. The vulnerability lies in the fact that email identities are very easy to forge. With common email, there is no built-in process to ensure that the sender of a message is who he or she claims to be. Email headers are easy enough to spoof, and most individuals do not know what to look for when receiving an email. One solution to these

problems is to use secure email. The basic requirements of secure email are described as follows:

- **Non-Disclosure of the Contents of the Email Message:** This is usually achieved by employing some encryption technology.
- **Message Integrity:** In other words, secure email ensures that the message has not been altered during transit and provides a method to certify the message's integrity. This is usually achieved by employing some hashing or message digest algorithm.
- **Verification of Sender:** Secure email provides some method to ensure the identity of the sender with a high degree of confidence. This is usually achieved by employing digital signature technology.
- **Verification of Recipient:** This can be achieved by employing public key encryption. The following sections offer a very brief overview of just some of the options available for secure email. Some are more popular than others. In fact, some are rather obscure. Most use encryption of some kind to varying degrees to provide message confidentiality. They also use different methods to provide sender authentication. Some use an informal process, while others depend on a formal hierarchy that does not exist yet and may never exist.

5.9 SECURE EMAIL PROTOCOLS

There is no lack of secure email standards: In fact, that is the problem. There are several competing standards and products from which you can choose. Some of the standards and products that are available are

- PEM
- Secure Multipurpose Internet Mail Extension (MIME) (S/MIME)
- MIME Object Security Service (MOSS)
- Message Security Protocol (MSP)

These competing standards and products are one of the primary reasons that secure email has not been widely implemented. The standards are not interoperable. If you use PGP to send someone a secure email, but the recipient employs S/MIME, then the recipient will not be able to open and read the message, let alone authenticate the sender of the message.

5.9.1 Pretty Good Privacy (PGP)

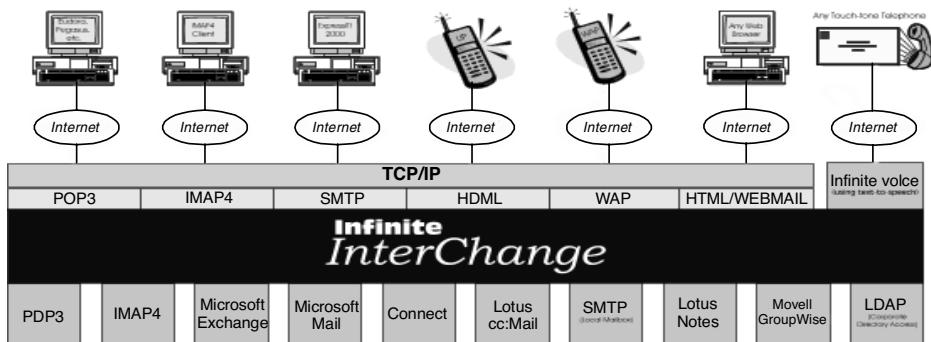


FIGURE 5.11 Secure email protocols

PGP is an encryption program that was developed by Phil Zimmerman in the early 1990s. It can be used to encrypt not only email but also files. There is even a special version of PGP available for encrypting telephone conversations. There is also a commercial PGP VPN client available from Network Associates, which is advertised as being fully IPsec compliant and supporting X.509 digital certificates. Basically, Zimmerman's idea behind PGP was to offer encryption capabilities to the masses. When Zimmerman made the PGP program available free of charge on the Internet for anyone to download, he got into a great deal of trouble with the U.S. government for violating restrictions on the export of encryption technology. For a while, it appeared that the government was going to press its case against Zimmerman, but reason eventually prevailed and an agreement was reached. There was also some trouble in that the earlier version of PGP used proprietary technology owned by RSA. As a result, later versions had to be modified so that the RSA copyright was not affected, and some earlier versions of PGP may not be compatible with the later versions. PGP uses public key cryptography to ensure confidentiality. It also uses digital signatures to authenticate the sender's identity, ensure message integrity, and provide non-repudiation. In addition, PGP can be used to encrypt files on storage media to ensure the confidentiality of stored files. Network Associates markets PGP. It is available for purchase, but in keeping with Zimmerman's original intent, PGP is also available as freeware. It can be downloaded at a number of sites. One of the best resources for PGP is MIT's Website.

5.9.2 Privacy-Enhanced Mail (PEM)

PEM is a proposed standard that defines the use of public key encryption to secure email for transmission on the Internet. The standard also provides for authentication. PEM uses a hierarchical organization for authentication and the distribution of keys. For a key to be valid, it must be signed by a CA. The hierarchy needed for authentication does not yet exist. This lack of the hierarchy infrastructure is probably one of the main reasons PEM has not been widely deployed. However, PEM should be able to employ X.509 digital certificates for the purpose of authenticating the identity of the sender.

5.9.3 PGP Versus PEM

There are several major differences between PGP and PEM. First, PGP is a product, while PEM is a standard. In addition, PEM does not allow for anonymous messages. When a message is signed with PEM, the signature can be reviewed by anyone. Even those who cannot decrypt the message can determine the signer of the message. The PEM standard relies on a formal hierarchical public key system, which has yet to be implemented. PGP relies on a web of trust, a more informal process that utilizes key rings maintained by various organizations. For this reason, PGP is not well suited for commercial use. In fact, you may find that many corporate IT departments do not allow it to be loaded on company-owned systems.

5.9.4 Secure MIME (S/MIME)

MIME is an extension of the original Internet email protocol standard that specifies how messages must be formatted, so they can be exchanged between different email systems.

S/MIME is a standard proposed by RSA that describes a secure method of sending email that uses the RSA encryption system. S/MIME is supported by Microsoft and Netscape in the latest versions of their Web browsers and has been endorsed by other vendors that make messaging products. RSA has proposed S/MIME as a standard to the Internet Engineering Task Force (IETF). An alternative to S/MIME is PGP/MIME, which has also been proposed as a standard. In the example above, we reviewed how we could use Eudora and PGP for secure e-mail. PGP uses public key cryptography and an informal organization of key rings to provide message confidentiality, integrity, and sender authentication. S/MIME employs digital certificates with digital signatures for sender authentication. It also uses hashing to ensure message

integrity and a combination of secret key (symmetric) and public key (asymmetric) encryption to ensure confidentiality.

We can illustrate how S/MIME functions by using Microsoft's Outlook Express. MS Outlook is designed to support S/MIME. The first step is to install a personal *digital certificate* in Outlook Express. A digital certificate is often called a *digital ID*. The two terms are synonymous. Installing the digital certificate is easily done by going to the "options" menu and clicking on the "security" tab. There you will find the option get digital ID.

5.10 WEB-BASED EMAIL SERVICES

Many people use free Web-based email services from companies such as Yahoo, Microsoft, or AOL. For example, Microsoft's Hotmail service has approximately 40 million subscribers. In general, Web-based email services should be considered very unsecure. There have been numerous security breaches reported with these email services. In 1999, several vulnerabilities were discovered with Microsoft's Hotmail alone. One reported problem allowed potential hackers full access to emails stored on Hotmail. That same year, a similar vulnerability was reported with Network Solutions' free Web-based email service. Another reported vulnerability involved the way Hotmail handled Java script code. This vulnerability enabled hackers to present a bogus password entry screen. By doing so, hackers could steal the passwords to accounts, thereby gaining full access to the accounts.



FIGURE 5.12 Web-based email services

Security of Stored Messages

Generally, most discussions regarding email security cover the security during the transmission of the email and the authentication process. As we have seen, encryption is useful for ensuring confidentiality and integrity during transmission of a message and for authenticating the sender's identity. However, many people overlook the risks associated with stored email. We have already discussed examples of Web-based email services that have been compromised, with associated email accounts exposed. However, even company email is at risk to disclosure. Email is certainly at risk of disclosure if it is stored on a centralized service. Email stored locally is somewhat more secure, but there may still be copies of messages stored on central servers. On more than one occasion, an email message has come back to haunt its sender in a court of law. One famous case is an email message sent by Bill Gates that included embarrassing information about Microsoft's business practices. This email surfaced in the antitrust case brought against Microsoft. Even messages that have been deleted have been recovered to the consternation of the person who believed that he or she had deleted the message. Some encryption packages, such as PGP, allow users to store files in an encrypted format. To decrypt the files requires entering in a password that the user assigns. However, even if email is stored in an encrypted format, you can still be legally compelled to reveal the contents. A company called Global Markets purports to offer a solution to this problem. Global Markets (www.1on1mail.com) has an email service that reportedly not only encrypts email but also has a self-destruct feature. This is another free Web-based email service. The software deletes the email two hours after receipt and overwrites the file space on the disk.

5.11 CERTIFICATION AUTHORITY HIERARCHIES

The Microsoft Public Key Infrastructure (PKI) supports a hierarchical certification authority (CA) model. A certification hierarchy provides scalability, ease of administration, and consistency with a growing number of commercial and other CA products.

In its simplest form, a certification hierarchy consists of a single CA. However, in general, a hierarchy contains multiple CAs with clearly defined parent-child relationships. In this model, the child subordinate certification authorities are certified by their parent CA-issued certificates, which bind a certification authority's public key to its identity. The CA at the top of a

hierarchy is referred to as the *root authority*, or root CA. The child CAs of the root CAs are called *subordinate Certification Authorities* (CAs).

In Windows XP and the Windows Server 2003 family, if you trust a root CA (by having its certificate in your Trusted Root Certification Authorities certificate store), you trust every subordinate CA in the hierarchy, unless a subordinate CA has had its certificate revoked by the issuing CA or has an expired certificate. Thus, any root CA is a very important point of trust in an organization and should be secured and maintained accordingly.

Verification of certificates thus requires trust in only a small number of root CAs. This provides flexibility in the number of certificate-issuing subordinate CAs. There are several practical reasons for supporting multiple subordinate CAs, including

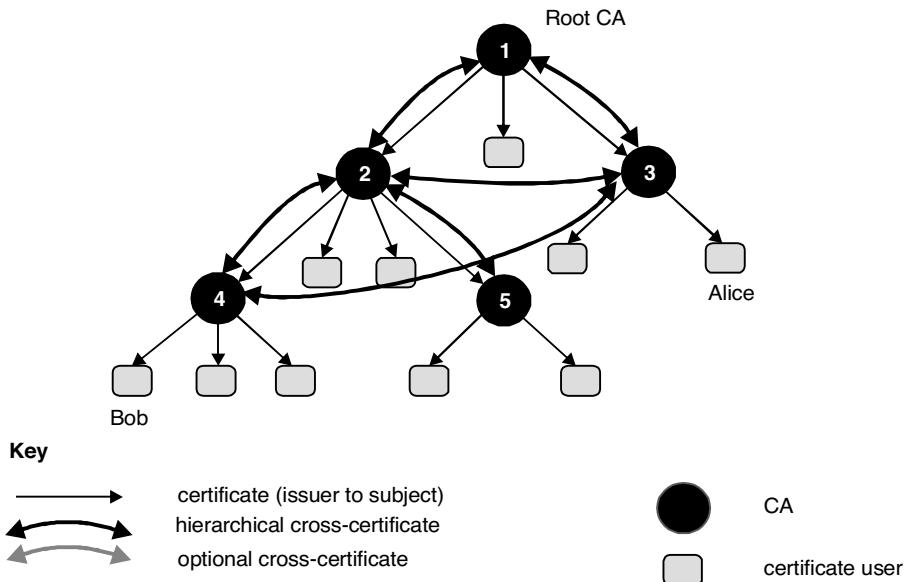


FIGURE 5.13 Key distribution

- **Usage:** Certificates may be issued for a number of purposes, such as secure email and network authentication. The issuing policy for these uses may be distinct, and separation provides a basis for administering these policies.
- **Organizational Divisions:** There may be different policies for issuing certificates, depending upon an entity's role in the organization. Again, you can create subordinate CAs to separate and administer these policies.

- **Geographic Divisions:** Organizations may have entities at multiple physical sites. Network connectivity between these sites may dictate a requirement for multiple subordinate CAs to meet usability requirements.
- **Load Balancing:** If your PKI is going to support the issuing of a large number of certificates, having only one certification authority issue and manage all of these certificates can result in considerable network load for that single certification authority. Using multiple subordinate certification authorities to issue the same kind of certificates divides the network load between certification authorities.
- **Backup and Fault Tolerance:** Multiple certification authorities increase the possibility that your network will always have operational certification authorities available to service users.

Such a certification authority hierarchy also provides administrative benefits, including

- Flexible configuration of the CA security environment to tailor the balance between security and usability, such as key strength, physical protection, and protection against network attacks. For example, you may choose to employ special purpose cryptographic hardware on a root CA, operate it in a physically secure area, or operate it offline. These may be unacceptable for subordinate CAs, due to cost or usability considerations.
- The ability to “turn off” a specific portion of the CA hierarchy without affecting the established trust relationships. For example, you can easily shut down and revoke an issuing CA certificate that is associated with a specific geographic site without affecting other parts of the organization.

5.12 KEY RECOVERY AND ESCROWED ENCRYPTION

In an ideal setting, parties involved in an encrypted message exchange would never lose their passwords (keys), nor would they ever leave. In fact, both of these happen all the time. Consider the following scenario: A customer deals with a sales representative from a company. The customer sends an encrypted order to the sales representative. Before the order is completed, the sales representative leaves the company. In order for the transaction to be completed without requesting a repeat order from the customer, the sales representative’s supervisor or replacement must be able to read the message sent by the customer. This can only be done if one of these parties was included in the original transaction, or if they can recover the encryption key used.

5.12.1 Key Recovery Methodologies

There are a number of algorithms, proposals, and commercial products available that describe or implement key recovery using key escrow. Most key escrow systems have three main components: a User Security Component (USC), a Key Escrow Component (KEC), and a Data Recovery Component (DRC). The USC encrypts and decrypts the data. It often incorporates a Data Recovery Field (DRF) in the message to facilitate key recovery. The KEC stores the data recovery key(s). The DRC performs the recovery of the plaintext from the encrypted text using the piece(s) from the KEC and the DRF set-up by the USC.

Key recovery requirements desired by governments include access without knowledge or consent of the users involved, omnipresent adoption, and speedy retrieval. Law enforcement demands access to the encrypted communication data as well—something not inherent for any commercial needs. These requirements are at odds with some of the goals of the original encryption. The following sections outline techniques that are representative of the different methodologies currently available.

5.12.2 Key Recovery Entry

The *Key Recovery Entry* scheme is a key recovery system that adds a small field to a message when it is transmitted. It also relies on a Certificate Authority (CA) body to authenticate requests and provide a key that opens the extra field. The author breaks his approach into four cases: no key recovery, recovery of session key, recovery of private keys, and PGP key recovery. For the first case, the normal encryption and decryption steps are performed for the algorithm used. A session key is used, along with private keys for each user. Both the private key and the session key are needed to decrypt the message. If either key is forgotten, the message cannot be recovered. For the second case, a CA is used to uniquely identify individuals. These certificates are retained in a database. The CA stores the answers to a series of questions, such as mother's maiden name. The public key is stored with the answers. When an encrypted message is sent, an additional field is added to the message. This is the Key Recovery Entry (KRE). It includes the session key and the public key of the recipient, as stored with the CA. If the recipient has lost his private key, the message may still be decrypted by contacting the CA and responding to the stored questions. If the answers match, the session key is provided. For the third case, a session key does not exist that will allow the decryption of the message. Whenever a private key is created or updated using a key package, a key recovery file is also created. This file contains the private key,

secured with a password created by a hash of the verification answers stored with the CA. The CA holds the key for decrypting this file. If the private key is forgotten, the recovery file is used to retrieve the key. In the final case, there exists no CA. In this case, a Lightweight Certificate Authority (LCA) must be added to the system. It is responsible for storing the public key and verification answers.

5.12.3 Key Escrow

The first part of the demonstration involves setting up the key escrow. AES Encryption should be selected. The key to be used is entered in the edit box for Client A (shown as “the key is 88888”). It needs to be 16 characters long. The button “Send A Key to Trustee” is pressed. This allows the application to send the key to Client A. Client A will also escrow its key with the trustee. Next, the key to be used for Client B is entered in that edit box. It also needs to be 16 characters long, and should be the same value as the key used for Client A. The button “Send B Key to Trustee” is pressed. This allows the application to send the key to Client B. Client B will also escrow its key with the Trustee.

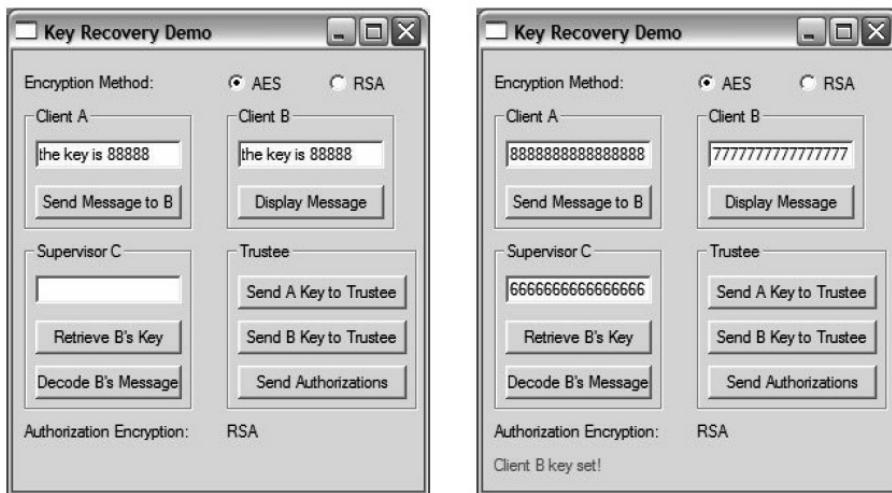


FIGURE 5.14 Set-up key for Client A and Client B

For the RSA portion of the key escrow, the clients exchange keys when they are setup.

It is not part of the key escrow demonstration. All three clients also need to set up their digital signatures (or authorization codes) with the trustee. This is used to verify the identity of any specific client. It is compared when a client requests the key of another client from the trustee process. The authorization code for each client should be entered in the appropriate edit boxes. (The authorization codes must be all numeric.) When there is an authorization code in all three edit boxes, the button “Send Authorizations” is pressed. Each client will store its authorization code and contact the Trustee to save the code for later authentication. The authorization codes are encrypted so the Trustee does not actually see them. In this example, Client A is identified by the signature “8888888888888888,” Client B by the signature “7777777777777777,” and Client C by the signature “6666666666666666.”

5.13 STRONG AND WEAK CRYPTOGRAPHY

The phrase *cryptographically strong* is often used to describe an encryption algorithm, and implies, in comparison to some other algorithm (which is thus cryptographically weak), greater resistance to attack. But it can also be used to describe hashing and unique identifier and filename creation algorithms.

An encryption algorithm is intended to be unbreakable (in which case, it is as strong as it can ever be), but it might be breakable (in which case, it is as weak as it can ever be) so there is not, in principle, a continuum of strength as the idiom would seem to imply: Algorithm A is stronger than Algorithm B which is stronger than Algorithm C, and so on. The situation is made more complex, and less subsumable into a single strength metric, by the fact that there are many types of cryptanalytic attack and that any given algorithm is likely to force the attacker to do more work to break it when using one attack than another.

The usual sense, in which this term is (loosely) used, is in reference to a particular attack, brute force key search—especially in explanations for newcomers to the field. Indeed, with this attack (always assuming keys were randomly chosen), there is a continuum of resistance depending on the length of the key used. But even so there are two major problems: many algorithms allow use of different length keys at different times, and any algorithm can forgo use of the full key length possible. Thus, Blowfish and RC5 are block cipher algorithms whose design specifically allowed for several key lengths, and who cannot therefore be said to have any particular strength with respect to brute force key search. Furthermore, U.S. export regulations restrict the

key length for exportable crypto products. In several cases in the 1980s and 1990s, only partial keys were used, decreasing the strength against brute force attacks for the exported versions. The same thing happened outside the U.S. as well, as for example, in the case of more than one of the crypto algorithms in the GSM cellular telephone standard.

The term is commonly used to convey the idea that an algorithm is suitable for a task in cryptography or information security, but also resists cryptanalysis and has no, or few, security weaknesses. Tasks are varied, and might include

- generating randomness
- encrypting data
- providing a method to ensure data integrity

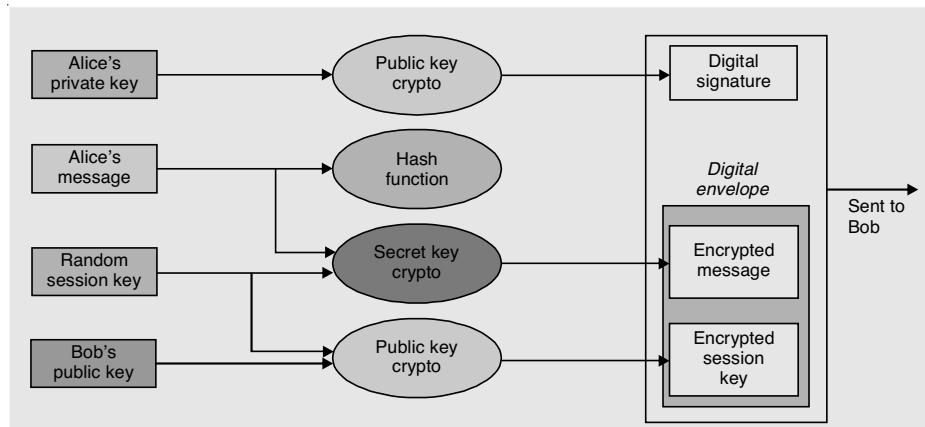


FIGURE 5.15 Cryptographic algorithms require their encapsulation in a cryptosystem

Cryptographically strong indicates that the described method is mature, perhaps even approved for use against different kinds of systematic attacks in theory and/or practice. It signals that the method may resist those attacks long enough to protect the information carried (and what stands behind the information) for a useful length of time. But due to the complexity and subtlety of the field, neither is almost ever the case. Such assurances are not actually available in real practice.

There will be always uncertainty as advances (*e.g.*, in cryptanalytic theory or merely affordable computer capacity) may reduce the effort needed to successfully use some attack method against an algorithm.

In addition, actual use of cryptographic algorithms requires their encapsulation in a cryptosystem, and doing so often introduces vulnerabilities that are not due to faults in an algorithm. For example, essentially all algorithms require a random choice of keys, and any cryptosystem that does not provide such keys will be subject to attack, regardless of any attack resistant qualities of the encryption algorithm(s) used.

5.14 SECURITY ALTERNATIVES FOR WEB FORMS

Virtually all businesses, most government agencies, and many individuals now have Websites. The number of individuals and companies with Internet access is expanding rapidly, and all of these have graphical Web browsers. As a result, businesses are enthusiastic about setting up facilities on the Web for ecommerce. But the reality is that the Internet and the Web are extremely vulnerable to attacks of various sorts. As businesses wake up to this reality, the demand for secure Web services grows.

The topic of Web security is a broad one and can easily fill a book (several are recommended at the end of this chapter). In this chapter, we begin with a discussion of the general requirements for Web security and then focus on two standardized schemes that are becoming increasingly important as part of Web commerce: SSL/TLS and SET.

5.14.1 Web Security Considerations

The World Wide Web is fundamentally a client/server application running over the Internet and TCP/IP intranets. As such, the security tools and approaches discussed so far in this book are relevant to the issue of Web security. But the Web presents new challenges not generally appreciated in the context of computer and network security:

- The Internet is two-way. Unlike traditional publishing environments, even electronic publishing systems involving teletext, voice response, or fax-back, the Web is vulnerable to attacks on the Web servers over the Internet.
- The Web increasingly serves as a highly visible outlet for corporate and product information and as the platform for business transactions. Reputations can be damaged and money can be lost if the Web servers are subverted.

- Although Web browsers are very easy to use, Web servers are relatively easy to configure and manage, and Web content is easy to develop, the underlying software is extraordinarily complex. This complex software may hide many potential security flaws. The short history of the Web is filled with examples of new and upgraded systems, properly installed, that are vulnerable to a variety of security attacks.
- A Web server can be exploited as a launchpad into a corporation's or agency's entire computer complex. Once the Web server is subverted, an attacker may be able to gain access to data and systems not part of the Web itself but connected to the server at the local site.
- Casual and untrained (in security matters) users are common clients for Web-based services. Such users are not necessarily aware of the security risks that exist and do not have the tools or knowledge to take effective countermeasures.

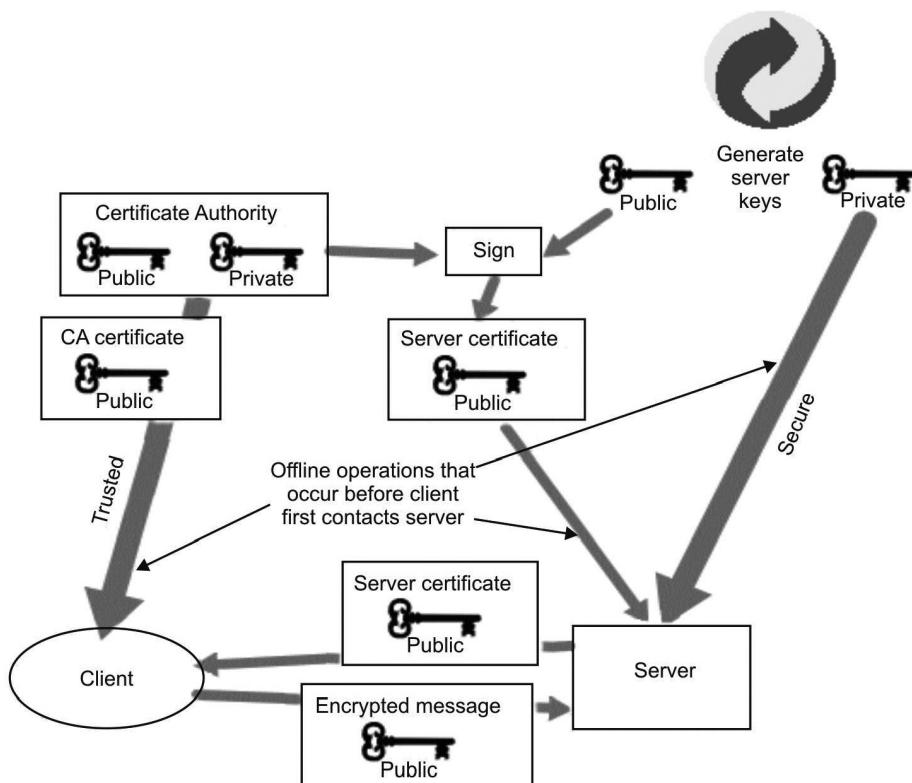


FIGURE 5.16 Web security threats

Table 5.1 Comparison of Threats on the Web.

Threats	Consequences	Countermeasures	
Integrity	<ul style="list-style-type: none"> • Modification of user data • Trojan horse browser • Modification of memory • Modification of message traffic in transit 	<ul style="list-style-type: none"> • Loss of information • Compromise of machine • Vulnerability to all other threats 	Cryptographic checksums
Confidentiality	<ul style="list-style-type: none"> • Eavesdropping on the Net • Theft of info from server • Theft of data from client • Info about network configuration • Info about which client talks to server 	<ul style="list-style-type: none"> • Loss of information • Loss of privacy 	Encryption, Web proxies
Denial-of-Service	<ul style="list-style-type: none"> • Killing of user threads • Flooding machine with bogus threats • Filling up disk or memory • Isolating machine by DNS attacks 	<ul style="list-style-type: none"> • Disruptive • Annoying • Prevent user from getting work done 	Difficult to prevent
Authentication	<ul style="list-style-type: none"> • Impersonation of legitimate users • Data forgery 	<ul style="list-style-type: none"> • Misrepresentation of user • Belief that false information is valid 	Cryptographic techniques

Table 5.1 provides a summary of the types of security threats faced in using the Web. One way to group these threats is in terms of passive and active attacks. *Passive attacks* include eavesdropping on network traffic between browser and server and gaining access to information on a website that is supposed to be restricted. *Active attacks* include impersonating another user, altering messages in transit between client and server, and altering information on a Web site.

Another way to classify Web security threats is in terms of the location of the threat: Web server, Web browser, and network traffic between browser and server. Issues of server and browser security fall into the category of computer system security. Part four of this book addresses the issue of system security in general but is also applicable to Web system security. Issues of traffic security fall into the category of network security and are addressed in this chapter.

5.15 WEB TRAFFIC SECURITY APPROACHES

A number of approaches to providing Web security are possible. The various approaches that have been considered are similar in the services they provide and, to some extent, in the mechanisms that they use, but they differ with respect to their scope of applicability and their relative location within the TCP/IP protocol stack.

Another relatively general-purpose solution is to implement security just above TCP. The foremost example of this approach is the Secure Sockets Layer (SSL) and the follow-on Internet standard of SSL known as Transport Layer Security (TLS). At this level, there are two implementation choices. For full generality, SSL (or TLS) could be provided as part of the underlying protocol suite and therefore be transparent to applications. Alternatively, SSL can be embedded in specific packages. For example, Netscape and Microsoft Explorer browsers come equipped with SSL, and most Web servers have implemented the protocol.

Application-specific security services are embedded within the particular application. Figure 5.16 shows examples of this architecture. The advantage of this approach is that the service can be tailored to the specific needs of a given application. In the context of Web security, an important example of this approach is the Secure Electronic Transaction (SET).

EXERCISES

1. Name three common network attacks used to undermine network security.
2. What are the three main types of networks that must be considered when defining a security policy?
3. List some of the areas of possible vulnerability in your own network.
4. What tools and applications are available to help monitor and test for system and network vulnerabilities?
5. List five important considerations to address when defining a security policy.

SECURITY AT THE IP LAYER

Chapter Goals

- Securing information on a network is cryptography.
- Plaintext to produce a stream of secrete
- Block ciphers
- Plaintext attack
- Public key cryptosystems
- Symmetric key encryption
- Data Encryption Standard (DES)
- Secrete key exchange

6.1 CRYPTOGRAPHY

For the exchange of information and commerce to be secure on any network, a system or process must be put in place that satisfies the need for confidentiality, access control, authentication, integrity, and non-repudiation. The key to the securing information on a network is cryptography. Cryptography can be used as a tool to provide privacy, to authenticate the identities of communicating parties, and to ensure message integrity. Its components include Confidentiality, access control, authentication, integrity, and non-repudiation.

Confidentiality: The ability to encrypt or encode a message to be transmitted over an insecure network

Access Control: The ability to control the level of access that an individual or entity can have to a network or system and how much information they can receive

Authentication: The ability to verify the identity of individuals or entity on the network

Integrity: The ability to ensure that a message or data has not been altered in transit from the sender to the recipient

Non-Repudiation: The ability to prevent individuals or entities from denying that they sent or received a file, when in fact they did.

Traditionally, cryptography conjures up thoughts of spies and secret codes. In reality, cryptography and encryption have found broad application in society. Every time you use an ATM machine to get cash or a point-of-sale machine to make a purchase, you are using encryption. *Encryption* is the process of scrambling the contents of a file or message to make it unintelligible to anyone not in possession of the “key” required to unscramble it. Civilizations have been using various cryptosystems for at least 4,000 years. A cryptosystem or algorithm is the process or procedure used to turn plaintext into crypto text. A crypto algorithm is also known as a “cipher.” There are several key elements that go into making an effective cryptosystem. First, it must be reversible. A crypto algorithm is of no practical use if once you have scrambled your information, you cannot unscramble it. The security of the cryptosystem should be dependent on the secrecy and length of the key, not on the details of the algorithm. In other words, knowing the algorithm should not make it significantly easier to crack the code (restricted versus unrestricted). If security is dependent on keeping the algorithm secret, then it is considered a *restricted* algorithm. It is also important that the algorithm has been subjected to substantial cryptanalysis. Only those algorithms that have been analyzed completely and at length are trustworthy. The algorithm should contain no serious or exploitable weakness. Theoretically, all algorithms can be broken by one method or another. However, an algorithm should not contain an inherent weakness that an attacker can easily exploit. Below is an example of a cipher; to scramble a message with this cipher, simply match each letter in a message to the first row and convert it into the number or letter in the second row. To unscramble a message, match each letter or number in a message to the corresponding number or letter in the second row and convert it into the letter in the first row.



FIGURE 6.1 To unscramble a message, match each letter or number in a message to the corresponding number

To illustrate how this works, see the following where the cipher is used to scramble the message “Little green apples.”

- Secrete: FCNNF5 AL55H 1JJF5M
- Clear text: LITTLE GREEN APPLES

This rudimentary cipher would not be effective at keeping a message secret for long. It does not comply with one of the qualities of a truly effective cipher, where knowing the algorithm should not make it significantly easier to crack the code. This is an example of a restricted algorithm. In this case, the cipher is the code. Once you know the cipher, you can unscramble any message. Ciphers usually fall into one of two categories: block ciphers or stream ciphers.

6.2 STREAM CIPHERS

Stream cipher algorithms process plaintext to produce a stream of secrete. The cipher inputs the plaintext in a stream and outputs a stream of secrete.

- Plaintext: Let us talk one to one.
- Ciphertext: F5n om nlfe ih5 ni ih5.

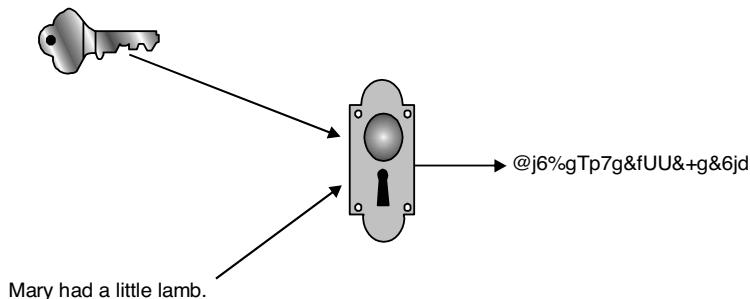


FIGURE 6.2 Stream cipher

Stream ciphers have several weaknesses. The most crucial shortcoming of stream ciphers is the fact that patterns in the plaintext can be reflected in the ciphertext. To illustrate this weakness, we can use the rudimentary cipher introduced earlier in the chapter. Below, we have scrambled the plaintext message “Let us talk one to one” into ciphertext to compare the two patterns.

- Plaintext: Let us talk one to one.
- Ciphertext: F5n om n1fe ih5 ni ih5.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	2	3	4	5	6	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T

FIGURE 6.3 To unscramble a message, match each letter or number in a message to the corresponding number

Patterns in the plaintext are reflected in the ciphertext. Words and letters that are repeated in the plaintext are also repeated in the ciphertext. Knowing that certain words repeat makes breaking the code easier. In addition, certain words in the English language appear with predictable regularity. Letters of the alphabet also appear in predictable regularity. The most commonly used letters of the alphabet in the English language are E, T, A, O, N, and I. The least commonly used letters in the English language are J, K, X, Q, and Z. The most common combination of letters in the English language is “th.” As a result, if a code breaker is able to find a “t” in a code, it doesn’t take long to find an “h.” It is not hard for a trained code breaker to break this type of code. Another weakness of stream ciphers is that they can be susceptible to a substitution attack, even without breaking the code. This is a type of replay attack where someone can simply copy a section of an old message and insert it into a new message. You don’t need to break the code to insert the old section into a new message. Examples of stream ciphers include the Vernam cipher, Rivest cipher #4 (RC4), and one-time pads.

6.3 BLOCK CIPHERS

Block ciphers differ from stream ciphers in that they encrypt and decrypt information in fixed size blocks rather than encrypting and decrypting each letter or word individually. A block cipher passes a block of data or plaintext through its algorithm to generate a block of cipher text. Ideally, a block cipher should generate cipher text roughly equivalent in size (in terms of number

of blocks) to the cleartext. A cipher that generates a block of ciphertext that is significantly larger than the information it is trying to protect is of little practical value. Think about it in terms of network bandwidth: If the cipher text block was twice the size of the plaintext, the net effect is that your bandwidth would be cut in half. This would also have an impact on files stored in an encrypted format. An unencrypted file 10 MB in size would be 20 MB in size when encrypted. Another requirement of block ciphers is that the cipher text should contain no detectable pattern. Examples of well-known block ciphers include the Data Encryption Standard (DES), the International Data Encryption Algorithm (IDEA), and SKIPJAK.

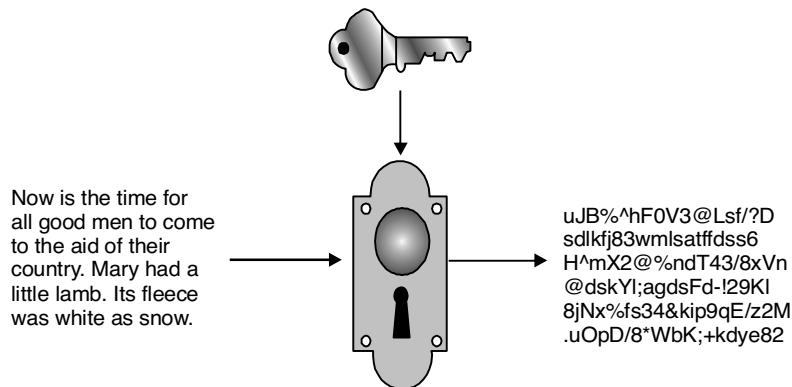


FIGURE 6.4 Block cipher

6.3.1 Breaking Ciphers

For as long as ciphers have existed, there have been people trying to break them. There are many methods employed to break ciphers. Some methods are ingenious. Some are sophisticated and technical in nature, while others are cruder in nature. The following sections describe some of the more widely used techniques employed in breaking ciphers.

6.4 KNOWN PLAINTEXT ATTACK

This method relies on the code breaker knowing in advance the plaintext content of a cipher text message. Having both the plaintext and the ciphertext, the code breaker reengineers the cipher and the key used to create the ciphertext.

6.4.1 Chosen Plaintext Attack

This method relies on the ability of the code breaker to somehow get a chosen plaintext message encrypted. During World War II, the United States used a variation of this method to ascertain the plans of the Japanese navy in the Pacific. Right after Pearl Harbor, the U.S. Pacific Fleet was forced to fight what was primarily a defensive war. The U.S. Pacific Fleet had been devastated by the Japanese surprise attack on Pearl Harbor, and all that was left of the fleet were three aircraft carriers and a handful of supporting ships. The United States had some success in breaking the Japanese codes. The U.S. Navy had determined that the Japanese were planning to attack a location referred to in their transmissions as “AF.” The United States suspected that site AF was Midway Island. To determine if AF was, in fact, Midway, the United States ordered that a message be transmitted from Midway stating that the island’s water condenser had broken down. The message was to be sent in the clear so that there would be no chance that the Japanese could not intercept it. Sure enough, the Japanese took the bait. A few days later, the United States intercepted a Japanese coded message stating that AF’s water condenser had failed. From that message, the United States knew that the Japanese were going to attack Midway. As a result, the United States was able to send what was left of the Pacific Fleet to Midway where they ambushed the Japanese carrier task force. The United States sank four of the Japanese’ frontline aircraft carriers. It was a strategic victory for the United States in the Pacific from which the Japanese navy never recovered. From that point on, it was the Japanese Navy that was forced to fight a defensive war.

6.5 CRYPTANALYSIS

Technically, any method employed to break a cipher or code is cryptanalysis. However, cryptanalysis here specifically refers to employing mathematical analysis to break a code. This method requires a high level of skill and sophistication. It is usually only employed by academics and governments. Today, it relies very heavily on the use of ultrafast super computers. Probably the most active and successful organization in the world dedicated to breaking codes is the National Security Agency (NSA). This is the largest spy agency in the United States. It is sometimes referred to as the Puzzle Palace, because the group spends so much time and energy on codes and cipher. The NSA employs tens of thousands of people. The only comparable organization in the world ever to have existed in terms of size is the former Soviet Union’s KGB. But with the breakup of the Soviet Union, the NSA is now left without peer.

6.6 BRUTE FORCE

The brute force method tries every possible combination of keys or algorithms to break a cipher. Doing so can require tremendous resources. Usually, this type of attack requires computer assistance. If the algorithm is simple or the key is small, then the CPU resources required could be provided by a simple PC. If the algorithm is sophisticated or the key is large, then advanced computing power might be required.

6.6.1 Social Engineering

This method relies on breaking a cipher by getting someone knowledgeable about the cipher to reveal information on how to break it. Bribing someone, tricking him or her into divulging information, or threatening him or her with harm can reveal information. When the threat of harm is employed, it is sometimes referred to as *rubber-hose cryptanalysis*.

6.6.2 Other Types of Attacks

Some other types of attacks are as follows:

- **Substitution:** This is a type of replay attack where a previous message, in part or in whole, is inserted into a legitimate message. An attacker does not need to break the cipher for this type of attack to be effective.
- **Timing attacks:** Some cryptosystems can be broken if an outsider is able to accurately measure the time required to perform the encryption and decryption of a known ciphertext. The known ciphertext and the timing provide enough information to deduce fixed exponents and factors of some systems. This vulnerability is mostly theoretical. If an attacker has enough access to a network to be able to accurately measure the time required to encrypt and decrypt information, then you have other and bigger problems to worry about.

6.7 ENCRYPTION

Encryption is the process of scrambling the contents of a file or message to make it unintelligible to anyone not in possession of the “key” required to unscramble the file or message. There are two types of encryption: symmetric (private/secret) key and asymmetric (public) key encryption.

6.8 SYMMETRIC KEY ENCRYPTION

When most people think of encryption, it is symmetric key cryptosystems that they think of. The *symmetric key*, also referred to as *private key* or *secret key*, is based on a single key and algorithm being shared between the parties who are exchanging encrypted information. The same key both encrypts and decrypts messages.

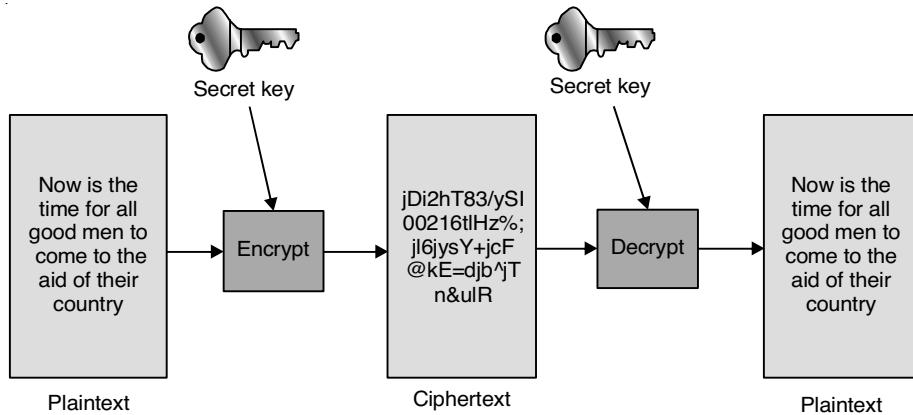


FIGURE 6.5 Symmetric key encryption

The strength of the scheme is largely dependent on the size of the key and on keeping it secret. Generally, the larger the key, the more secure the scheme. In addition, symmetric key encryption is relatively fast. The main weakness of the system is that the key or algorithm has to be shared. You can't share the key information over an unsecured network without compromising the key. As a result, private key cryptosystems are not well suited for spontaneous communication over open and unsecured networks. In addition, symmetric key provides no process for authentication or non-repudiation. Remember, non-repudiation is the ability to prevent individuals or entities from denying (repudiating) that a message was sent or received or that a file was accessed or altered, when in fact it was. This ability is particularly important when conducting ecommerce. Table 6.1 lists the advantages and disadvantages of symmetric key cryptosystems. Examples of widely deployed symmetric key cryptosystems include DES, IDEA, Blowfish, RC4, CAST, and SKIPJACK.

Table 6.1 The Advantages and Disadvantages of Symmetric Key Cryptography.

Advantages	Disadvantages
Fast	Requires secret sharing
Relatively secure	Complex administration
Widely understood	No authentication No non-repudiation

6.9 DATA ENCRYPTION STANDARD (DES)

DES is one of the oldest and most widely used algorithms. DES was developed by IBM with the encouragement of the NSA. It was originally deployed in the mid 1970s. DES consists of an algorithm and a key. The key is a sequence of eight bytes, each containing eight bits for a 64-bit key. Since each byte contains one parity bit, the key is actually 56 bits in length. According to author James Bamford in his book *The Puzzle Palace*, IBM originally intended to release the DES algorithm with a 128-bit key, but the NSA convinced IBM to release it with the 56-bit key instead. Supposedly this was done to make it easier for the NSA to decrypt covertly intercepted messages. DES is widely used in Automated Teller Machine (ATM) and Point-of-Sale (POS) networks, so if you use an ATM or debit card you are using DES. DES has been enhanced with the development of triple DES. However, DES has been broken. It is gradually being phased out of use.

Graphic courtesy of Charles Breed

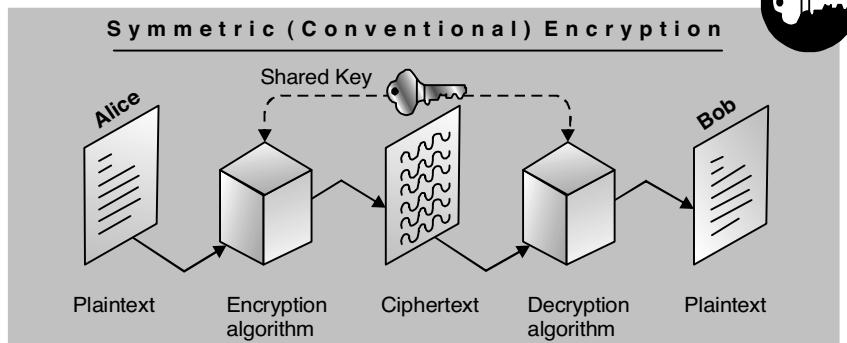


FIGURE 6.6 Data encryption standard

6.9.1 International Data Encryption Algorithm (IDEA)

IDEA is a symmetric key block cipher developed at the Swiss Federal Institute in the early 1990s. IDEA utilizes a 128-bit key. Supposedly, it is more efficient to implement in software than DES. Since it was not developed in the United States, it is not subject to U.S. export restrictions.

6.9.2 CAST

The CAST algorithm supports variable key lengths, anywhere from 40 bits to 256 bits in length. CAST uses a 64-bit block size, which is the same as the DES, making it a suitable drop-in replacement. CAST has been reported to be two to three times faster than a typical implementation of DES and six to nine times faster than a typical implementation of triple DES. The CAST algorithm was developed by Carlisle Adams and Stafford Travares and patented by Entrust Technologies, but a version of the CAST algorithm is available for free commercial and non-commercial use. CAST is employed in Pretty Good Privacy (PGP).

6.9.3 Rivest Cipher #4 (RC4)

Developed by Ron Rivest of RSA fame, RC4 is a stream cipher that uses a variable size key. However, when used with a key of 128 bits, it can be very effective. Until recently, the approved export version only used a 40-bit key. RC4 is used in Netscape Navigator and Internet Explorer.

6.10 ASYMMETRIC KEY ENCRYPTION

For centuries, all cryptography was based on the symmetric key cryptosystems. Then in 1976, two computer scientists, Whitfield Diffie and Martin Hellman of Stanford University, introduced the concept of asymmetric cryptography. Asymmetric cryptography is also known as *public key cryptography*. Public key cryptography uses two keys as opposed to one key for a symmetric system. With public key cryptography there is a public key and a private key. The keys' names describe their function. One key is kept private, and the other key is made public. Knowing the public key does not reveal the private key. A message encrypted by the private key can only be decrypted by the corresponding public key. Conversely, a message encrypted by the public key can only be decrypted by the private key.

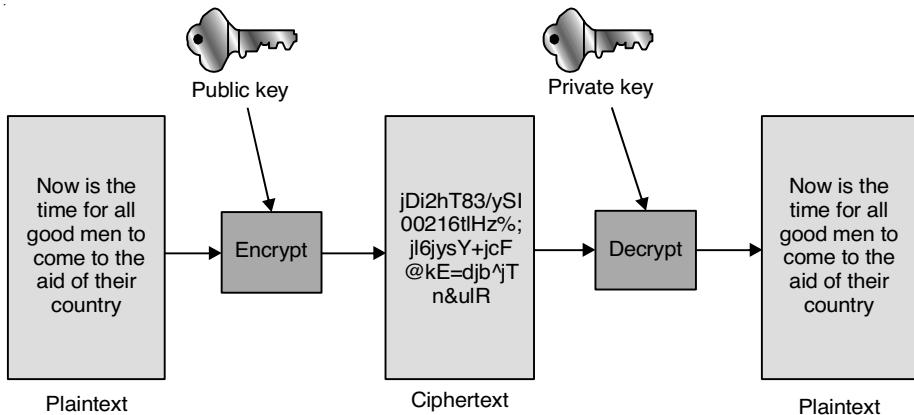


FIGURE 6.7 Asymmetric key encryption

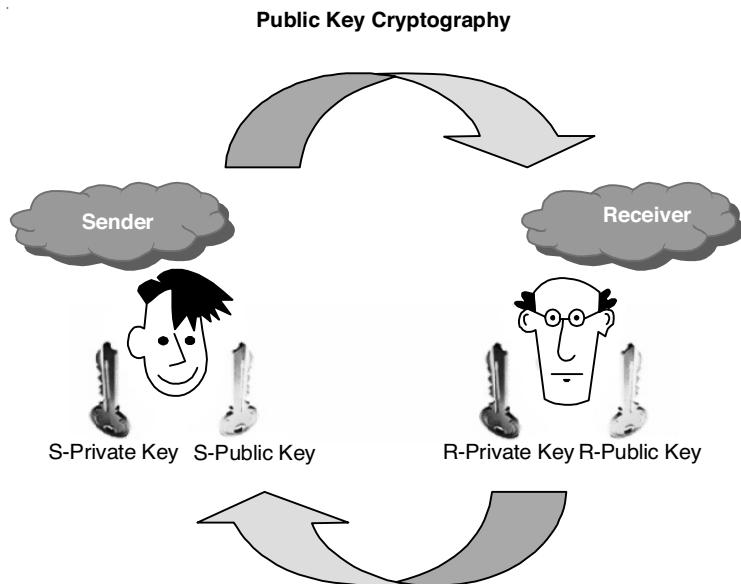
With the aid of public key cryptography, it is possible to establish secure communications with any individual or entity when using a compatible software or hardware device. For example, if Alice wishes to communicate in a secure manner with Bob, a stranger with whom she has never communicated before, Alice can give Bob her public key. Bob can encrypt his outgoing transmissions to Alice with Alice's public key. Alice can then decrypt the transmissions using her private key when she receives them. Only Alice's private key can decrypt a message encrypted with her public key. If Bob transmits to Alice his public key, then Alice can transmit secure encrypted data back to Bob that only Bob can decrypt. It does not matter that they exchanged public keys on an unsecured network. Knowing an individual's public key tells you nothing about his or her private key. Only an individual's private key can decrypt a message encrypted with his or her public key. The security breaks down if either of the parties' private keys is compromised. While symmetric key cryptosystems are limited to securing the privacy of information, asymmetric or public key cryptography is much more versatile. Public key cryptosystems can provide a means of authentication and can support digital certificates. With digital certificates, public key cryptosystems can provide enforcement of non-repudiation. Unlike symmetric key cryptosystems, public key allows for secure spontaneous communication over an open network. In addition, it is more scalable for very large systems (tens of millions) than symmetric key cryptosystems. With symmetric key cryptosystems, the key administration for large networks is very complex.

Table 6.2 The Advantages and Disadvantages of Public Key Cryptography.

Advantages	Disadvantages
No secret sharing necessary Authentication supported Provides non-repudiation Scalable	Slower or computationally intensive Certificate authority required

6.11 PUBLIC KEY CRYPTOSYSTEMS

There are three public key algorithms in wide use today—Diffie-Hellman, RSA, and the Digital Signature Algorithm (DSA). They are described in the following sections.

**FIGURE 6.8** Public key cryptosystems

6.11.1 Diffie-Hellman

The Diffie-Hellman algorithm was developed by Whitfield Diffie and Martin Hellman at Stanford University. It was the first usable public key algorithm. Diffie-Hellman is based on the difficulty of computing discrete logarithms. It

can be used to establish a shared secret key that can be used by two parties for symmetric encryption. Diffie-Hellman is often used for IPSEC key management protocols. For spontaneous communications with Diffie-Hellman, two communicating entities would each generate a random number that is used as their private keys. They exchange public keys. They each apply their private keys to the other's public key to compute identical values (shared secret key). They then use the shared secret key to encrypt and exchange information.

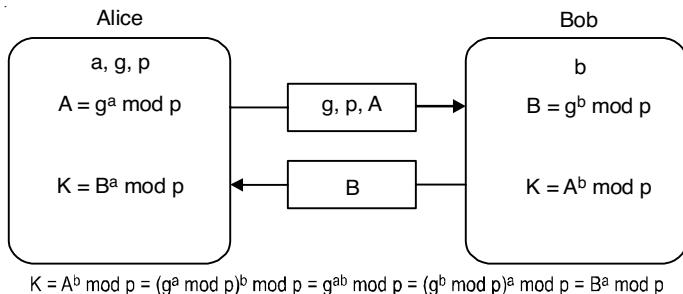


FIGURE 6.9 The Diffie-Hellman algorithm

6.11.2 Message Integrity

To attain a high level of confidence in the integrity of a message or data, a process must be put in place to prevent or detect alterations during transit. One technique employed is called a hash function. A *hash function* takes a message of any length and computes a product value of fixed length. The product is referred to as a *hash value*. The length of the original message does not alter the length of the hash value. Hash functions are used to ensure the integrity of a message or file. Using the actual message or file, a hash function computes a hash value, which is a cryptographic checksum of the message. This checksum can be thought of as a fingerprint for that message. The hash value can be used to determine if the message or file has been altered since the value was originally computed. Let's take email as an example: The hash value for a message is computed at both the sending and receiving ends. If the message is modified in anyway during transit, the hash value computed at the receiving end will not match the value computed at the sending end. Hash functions must be one-way only. In other words, there should be no way to reverse the hash value to obtain information on the message. Obviously, this would represent a risk. Another requirement of an effective one-way hash function is that the possibility of “collisions” is very limited, if nonexistent. A collision occurs when the same hash value is computed for two or more unique messages.

If the messages are different, the hash values should be different. No two unique messages should compute the same hash value.

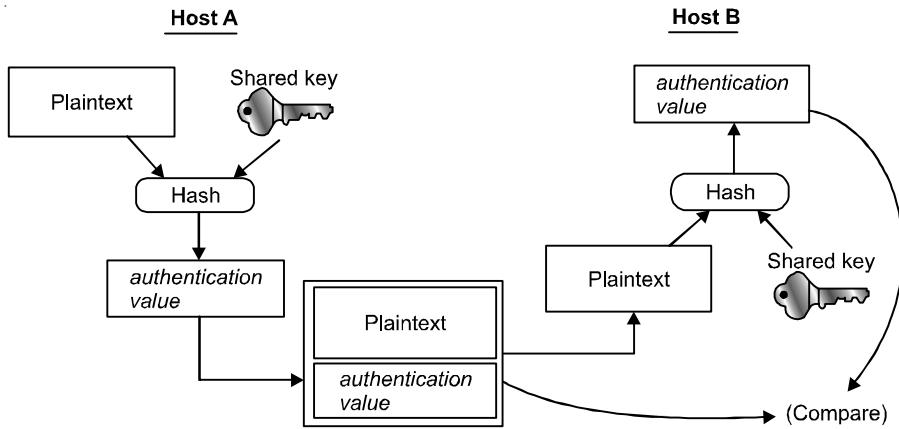


FIGURE 6.10 Ensuring message integrity

Table 6.3 Widely Used Hashing Algorithms.

Message digest #4 (MD4) from RSA
Message digest #5 (MD5) from RSA
Secure hash algorithm-1 (SHA-1)
RACE Integrity Primitives Evaluation (RIPE) MD-160 (RIPEMD-160)

MD4

MD4 was developed by Ron Rivest of RSA. MD4 is a one-way hash function that takes a message of variable length and produces a 128-bit hash value or message digest. MD4 has some weaknesses. Analysis has shown that at least the first two rounds of MD4 are not one-way (there are three rounds in MD4), and that the algorithm is subject to collisions.

MD5

MD5 was also created by Ron Rivest as an improvement on MD4. Like MD4, MD5 creates a unique 128-bit message digest value derived from the contents of a message or file. This value, which is a fingerprint of the message or file content, is used to verify the integrity of the message's or file's contents. If

a message or file is modified in any way, even a single bit, the MD5 cryptographic checksum for the message or file will be different. It is considered very difficult to alter a message or file in a way that will cause MD5 to generate the same result as was obtained for the original file. While MD5 is more secure than MD4, it too has been found to have some weaknesses. Analysis has shown a collision in the compression function of MD5, although not for MD5 itself. Nevertheless, this attack casts doubts on the whether MD5 is truly a collision-resistant hash algorithm. The MD5 algorithm is intended for digital signature applications, where a large file must be “compressed” in a secure manner before being encrypted with a private (secret) key under a public key cryptosystem such as RSA.

6.12 SECURE HASH ALGORITHM-1 (SHA-1)

SHA-1 is a one-way hash algorithm used to create digital signatures. SHA-1 is derived from SHA, which was developed in 1994 by the NIST. SHA-1 is similar to the MD4 and MD5 algorithms developed by Ron Rivest. SHA-1 is slightly slower than MD4 and MD5, but it is reported to be more secure. The SHA-1 hash function produces a 160-bit hash value or message digest. No known cryptographic attacks against SHA-1 have been successful. Since it produces a 160-bit message digest it is more resistant to brute force attacks than MD4 and MD5, which produce a 128-bit message digest.

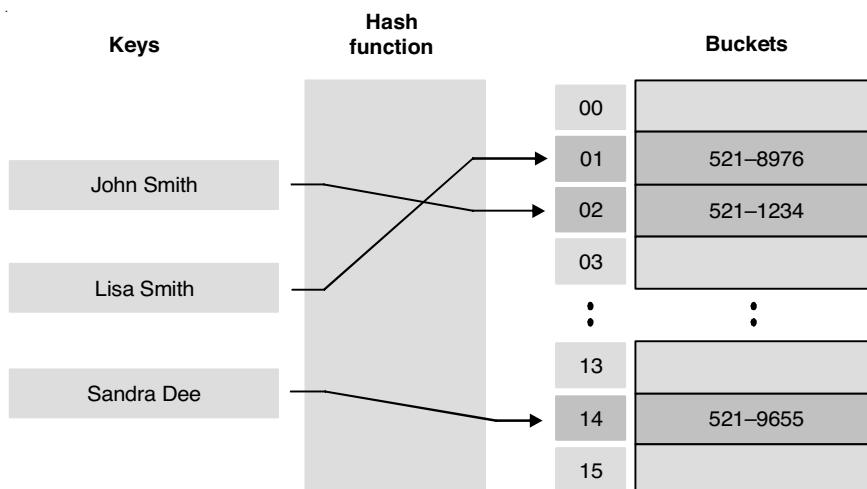


FIGURE 6.11 An example of a hash function

6.12.1 Authentication

To have a high level of confidence and trust in the integrity of information received over a network, the transacting parties need to be able to authenticate each other's identity. In the example involving Alice and Bob, it was demonstrated how they could transmit secure information between each party using encryption by exchanging public keys. While confidentiality was ensured with the use of public key cryptography, there was no authentication of the parties' identities. Bob may not really have been Bob. For that matter, Bob doesn't really know if Alice was Alice. In addition, how does Alice know that when she was sending her public key to Bob, that Jack did not intercept it and use it to send his public key to her and masquerade as Bob? To ensure secure business transactions on unsecured networks like the Internet, both parties need to be able to authenticate their identities. Authentication in a digital setting is a process whereby the receiver of a message can be confident of the identity of the sender. The lack of secure authentication has been a major obstacle in achieving widespread use of the Internet for commerce. One process used to authenticate the identity of an individual or entity involves digital signatures.

6.13 PUBLIC KEY INFRASTRUCTURE

As part of the future implementation of digital certificates, a movement is under way to develop a PKI. The infrastructure will be necessary to authenticate digital certificates and CAs. A PKI is a hierarchical network of CAs. A “root certificate” authority certifies subordinate CAs. The hierarchy is recognized as trusted by all entities that trust the hierarchical CA. Not every entity needs to trust the other, just the hierarchy. Some plans envision a hierarchy of CAs, where one CA certifies the identity of the previous CA. The top-level root CA in the United States could be the U.S. government. Others envision a more horizontal scheme of cross-certification with only a few layers. In either case, a certificate-based PKI can provide a process to establish trust relationships.

The difficult part is developing the standards and infrastructure for certifying digital signatures and certificates between organizations using different schemes. The NIST is working on the development of a federal PKI. While there are many challenges to developing a national PKI, the most daunting task is the development of the global infrastructure. When we discuss a global or international PKI, we open a Pandora's box of “national security” issues.

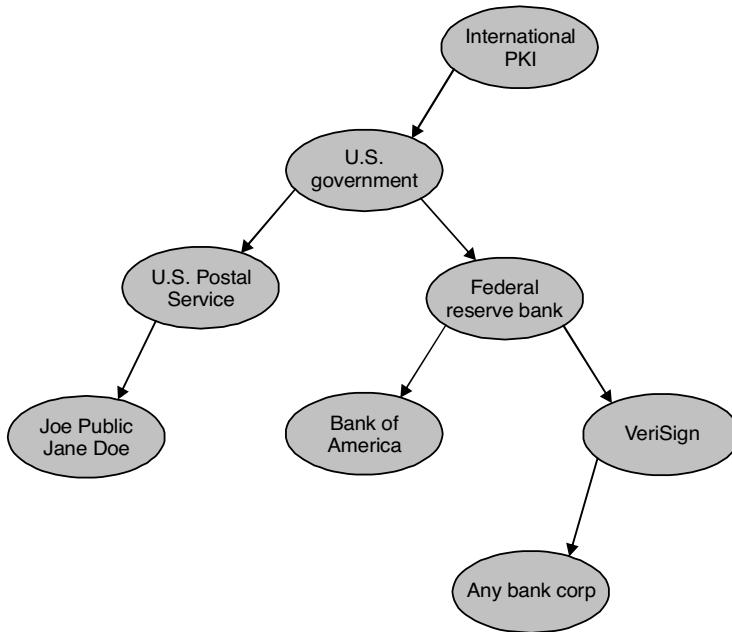


FIGURE 6.12 A theoretical PKI

6.14 SECRETE KEY EXCHANGE

Secrete key exchange utilizes a single central server, referred to as a trusted server, to act as a trusted third party to authenticate users and control access to resources on the network. The basic premise behind Secrete security is that it is not possible to ensure security on all network servers. This concept assumes that server security breaches are inevitable in a distributed computing environment with multiple servers. The premise is that it is impossible to secure all the servers, so one shouldn't even attempt to. The Secrete model proposes, however, that it is possible to truly secure a single server. Therefore, it holds that it is more secure to control all network access from one central secure server. The Secrete key exchange process is really quite simple, but at the same time quite eloquent. Secrete never transmits passwords on the network, regardless of whether they are encrypted or not. Secrete utilizes cryptographic keys referred to as *tickets* to control access to network server resources. Tickets are encrypted passes or files issued by the trusted server to users and processes to determine access level. There are six types of tickets: initial, invalid, pre-authenticated, renewable, forwardable, and postdated.

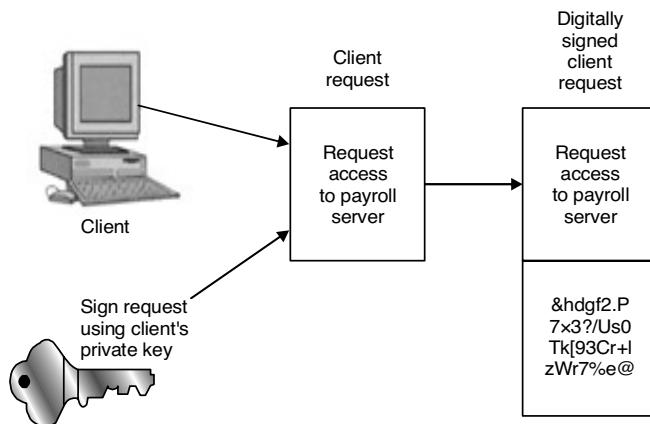


FIGURE 6.13 Secret key exchange, step one

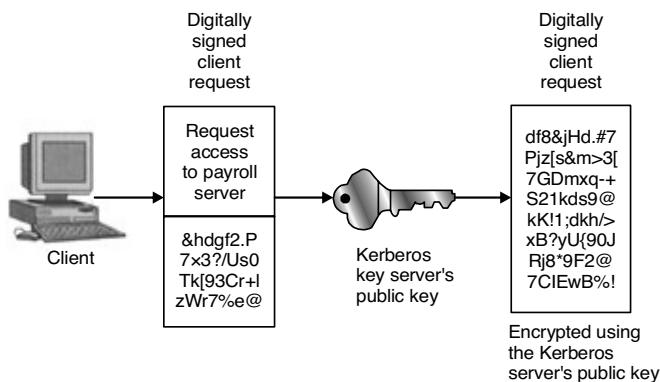


FIGURE 6.14 Secret key exchange, step two

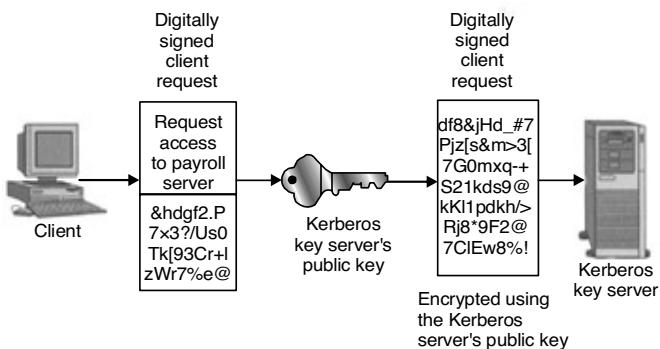


FIGURE 6.15 Secret key exchange, step three

The Secret server decrypts the request using its private key and then authenticates the originator of the request by verifying the digital signature of the sender. The request was digitally signed using the sender's private key, which the Secret server verifies by using the sender's public key. The Secret server maintains a database of all the public keys of authorized users, so it does not have to rely upon the sender or a trusted third party to verify the sender's public key. If the Secret server does not have the sender's public key in its database, then the digital signature cannot be verified. Similarly, if the Secret server does not have the sender's public key, then the sender is not an authorized user of the network, and the request will be denied. Once the Secret server receives the request and authenticates the sender's identity, the server verifies that the client has authorization to access the requested resource. In this example the resource requested is access to the payroll server. If the Secret server determines that the client does have authorization to access the payroll server, the Secret server sends identical session tickets to both the client and the payroll server. To transmit the session ticket to the client, the Secret server encrypts it with the client's public key. To transmit the ticket to the payroll server, the Secret server uses the payroll server's public key.

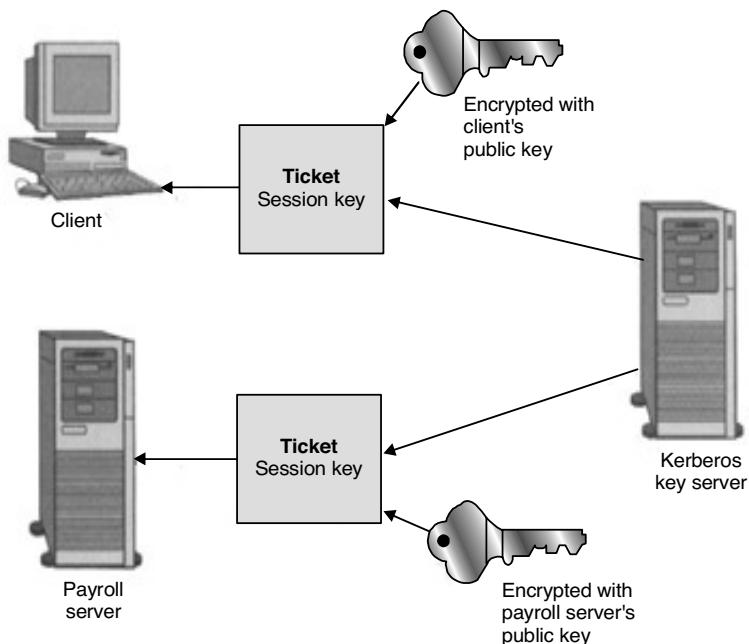


FIGURE 6.16 Secret key exchange, step four

The tickets could also be digitally signed by the Secretre server to avoid the possibility of counterfeit tickets being sent to a client or network resource. The client then sends a copy of its ticket to the payroll server. Before transmitting the ticket, the client encrypts the ticket using the payroll server's public key.

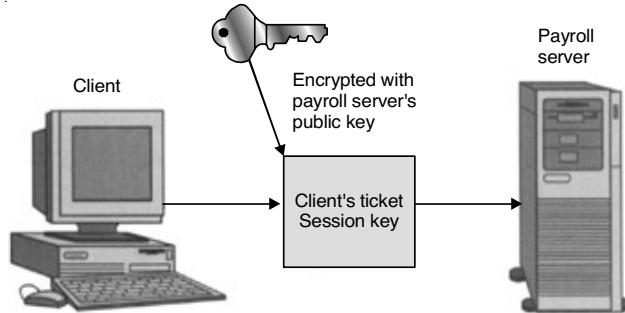


FIGURE 6.17 Secretre key exchange, step five

When the payroll server receives the encrypted ticket from the client, the server decrypts the ticket using the server's own private key. The payroll server then compares the ticket that it received from the client to the ticket that it received from the Secretre server. If the client's ticket matches the server's ticket, then the client will be allowed to connect to the server. If they don't match, the connection is refused. Once the connection is established, the systems can encrypt the communication using either the session key or the client's public key, or they can use no encryption at all.

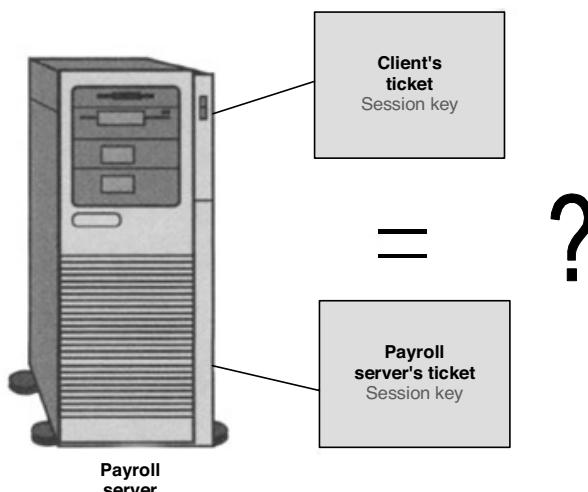


FIGURE 6.18 Secretre key exchange, step six

One advantage that Secrete has over other schemes, such as using digital certificates and a PKI, is that revocation of authorization and authentication can be done immediately. The PKI relies upon CRLs to remove the authorization for an individual or entity. Access to network resources may not be terminated until the CRL works its way through the PKI or the original digital certificate expires. In either case, the original certificate will provide access to network resources long after the time period you want access terminated. With Secrete, every time an individual or entity requests access to a network resource, the Secrete server is queried. As a result, once access is terminated at the Secrete server, the change is effective immediately.

6.15 WEB SECURITY

We have just studied two important areas where security is needed: communications and email. You can think of these as the soup and appetizer. Now it is time for the main course: Web security. The Web is where most of the Troubles (or intruders) hang out nowadays and do their dirty work. In the following sections we will look at some of the problems and issues relating to Web security. Web security can be roughly divided into three parts. First, how are objects and resources named securely? Second, how can secure, authenticated connections be established? Third, what happens when a Web site sends a client a piece of executable code? After looking at some threats, we will examine all these issues.

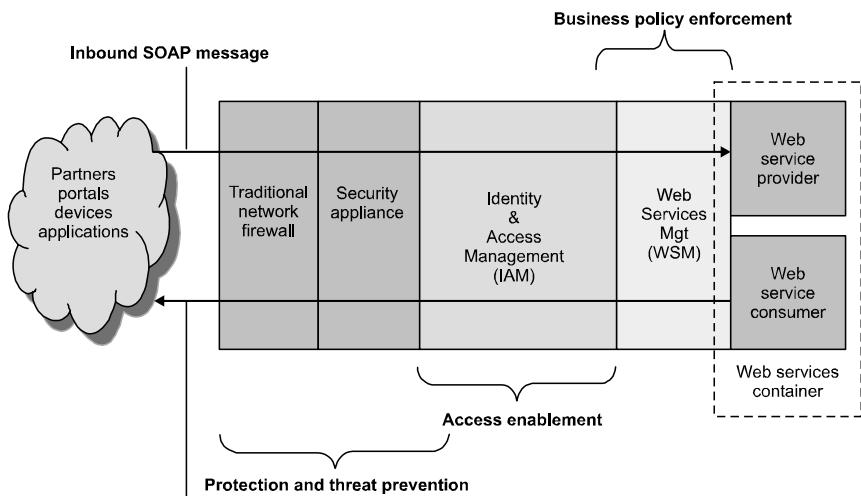


FIGURE 6.19 An illustration of Web security

6.15.1 Threats

One reads about Web site security problems in the newspaper almost weekly. The situation is grim. First, the home pages of numerous organizations have been attacked and replaced by a new home page of the crackers' choosing. (The popular press calls people who break into computers *hackers*, but many programmers reserve that term for great programmers. We prefer to call these people *crackers*.) Sites that have been cracked include Yahoo, the U.S. Army, the CIA, NASA, and the New York Times. In most cases, the crackers just put up some funny text, and the sites were repaired within a few hours. Numerous sites have been brought down by denial-of-service attacks, in which the cracker floods the site with traffic, rendering it unable to respond to legitimate queries. Often the attack is mounted from a large number of machines that the cracker has already broken into (DDoS attacks). These attacks are so common that they do not even make the news any more, but they can cost the attacked site thousands of dollars in lost business. In 1999, a Swedish cracker broke into Microsoft's Hotmail Web site and created a mirror site that allowed anyone to type in the name of a Hotmail user and then read all of the person's current and archived email. In another case, a 19-year-old Russian cracker named Maxim broke into an e-commerce Web site and stole 300,000 credit card numbers. Then he approached the site owners and told them that if they did not pay him \$100,000, he would post all the credit card numbers to the Internet. They did not give in to his blackmail, and he posted the credit card numbers, inflicting great damage to many innocent victims. In a different vein, a 23-year-old California student emailed a press release to a news agency falsely stating that the Emulex Corporation was going to post a large quarterly loss and that the C.E.O. was resigning immediately. Within hours, the company's stock dropped by 60%, causing stockholders to lose over \$2 billion. The perpetrator made a quarter of a million dollars by selling the stock short just before sending the announcement. While this event was not a Web site break-in, it is clear that putting such an announcement on the home page of any big corporation would have a similar effect. We could (unfortunately) go on like this for many pages. But it is now time to examine some of the technical issues related to Web security. There is much more information about security problems available for interested readers (Anderson, 2001; Garfinkel with Spafford, 2002; and Schneier, 2000).

6.15.2 Secure Naming

Let us start with something very basic: Alice wants to visit Bob's Web site. She types Bob's URL into her browser and a few seconds later, a Web page appears. But is it Bob's? Maybe yes and maybe no. Trudy might be up to her

old tricks again. For example, she might be intercepting all of Alice's outgoing packets and examining them. When she captures an HTTP *GET* request headed to Bob's Web site, she could go to Bob's Web site herself to get the page, modify it as she wishes, and return the fake page to Alice. Alice would be none the wiser. Worse yet, Trudy could slash the prices at Bob's estore to make his goods look very attractive, thereby tricking Alice into sending her credit card number to "Bob" to buy some merchandise. One disadvantage to this classic man-in-the-middle attack is that Trudy has to be in a position to intercept Alice's outgoing traffic and forge her incoming traffic. In practice, she has to tap either Alice's phone line or Bob's, since tapping the fiber backbone is fairly difficult. While active wiretapping is certainly possible, it is a certain amount of work, and while Trudy is clever, she is also lazy. Besides, there are easier ways to trick Alice.

6.16 DNS SPOOFING

For example, suppose Trudy is able to crack the DNS system, maybe just the DNS cache at Alice's ISP, and replace Bob's IP address (say, 36.1.2.3) with her (Trudy's) IP address (say, 42.9.9.9). That leads to the following attack. The way it is supposed to work is illustrated in Figure 6.20(a). Here, (1) Alice asks DNS for Bob's IP address, (2) gets it, (3) asks Bob for his home page, and (4) gets that, too. After Trudy has modified Bob's DNS record to contain her own IP address instead of Bob's, we get the situation of Figure 6.20(b). Here, when Alice looks up Bob's IP address, she gets Trudy's, so all her traffic intended for Bob goes to Trudy. Trudy can now mount a man-in-the-middle attack without having to go to the trouble of tapping any phone lines. Instead, she has to break into a DNS server and change one record, a much easier proposition.

How might Trudy fool the DNS? It turns out to be relatively easy. Briefly summarized, Trudy can trick the DNS server at Alice's ISP into sending out a query to look up Bob's address. Unfortunately, since DNS uses UDP, the DNS server has no real way of checking who supplied the answer. Trudy can exploit this property by forging the expected reply and thus injecting a false IP address into the DNS server's cache. For simplicity, we will assume that Alice's ISP does not initially have an entry for Bob's Web site, *bob.com*. If it does, Trudy can wait until it times out and try later (or use other tricks). Trudy starts the attack by sending a lookup request to Alice's ISP asking for the IP address of *bob.com*. Since there is no entry for this DNS

name, the cache server queries the top-level server for the *com* domain to get one. However, Trudy beats the *com* server to the punch and sends back a false reply saying “*bob.com* is 42.9.9.9,” where that IP address is hers. If her false reply gets back to Alice’s ISP first, that one will be cached and the real reply will be rejected as an unsolicited reply to a query no longer outstanding. Tricking a DNS server into installing a false IP address is called *DNS spoofing*. A cache that holds an intentionally false IP address like this is called a *poisoned cache*. Actually, things are not quite that simple. First, Alice’s ISP checks to see that the reply bears the correct IP source address of the top-level server. But since Trudy can put anything she wants in that IP field, she can defeat that test easily because the IP addresses of the top-level servers have to be public. Second, to allow DNS servers to tell which reply goes with which request, all requests carry a sequence number. To spoof Alice’s ISP, Trudy has to know its current sequence number. The easiest way to learn the current sequence number is for Trudy to register a domain herself, say, *trudy-the-intruder.com*. Let us assume its IP address is also 42.9.9.9. She also creates a DNS server for her newly-hatched domain, *dns.trudy-the-intruder.com*. It, too, uses Trudy’s 42.9.9.9 IP address, since Trudy has only one computer. Now she has to make Alice’s ISP aware of her DNS server. That is easy to do. All she has to do is ask Alice’s ISP for *foobar.trudy-the-intruder.com*, which will cause Alice’s ISP to find out who serves Trudy’s new domain by asking the top-level *com* server. With *dns.trudy-the-intruder.com* safely in the cache at Alice’s ISP, the real attack can start. Trudy now queries Alice’s ISP for *www.trudy-the-intruder.com*. The ISP naturally sends Trudy’s DNS server a query asking for it. This query bears the sequence number that Trudy is looking for. Trudy then asks Alice’s ISP to look up Bob. She immediately answers her own question by sending the ISP a forged reply, allegedly from the top-level *com* server saying: “*bob.com* is 42.9.9.9.” This forged reply carries a sequence number one higher than the one she just received. While she is at it, she can also send a second forgery with a sequence number two higher, and maybe a dozen more with increasing sequence numbers. One of them is bound to match. The rest are thrown out. When Alice’s forged reply arrives, it is cached; when the real reply comes in later, it is rejected since no query is then outstanding. Now when Alice looks up *bob.com*, she is told to use 42.9.9.9, Trudy’s address. Trudy has mounted a successful man-in-the-middle attack from the comfort of her own living room.

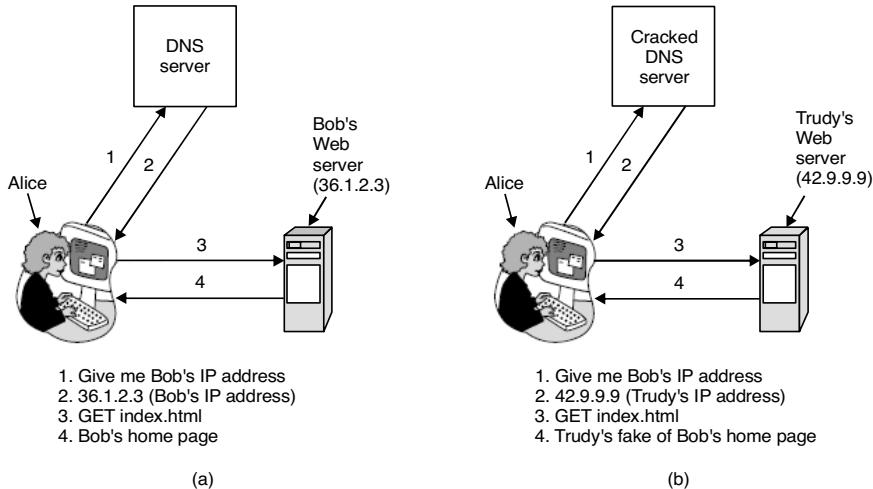


FIGURE 6.20 (a) Normal situation (b) An attack based on breaking into the DNS and modifying Bob's record

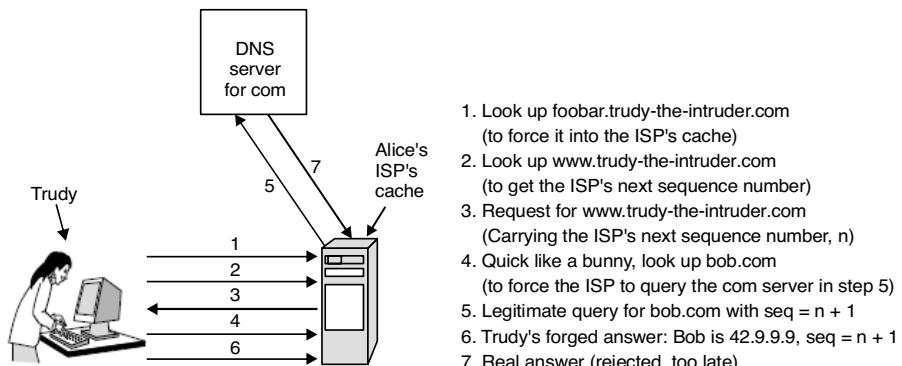


FIGURE 6.21 How Trudy spoofs Alice's ISP

6.16.1 Secure DNS

This one specific attack can be foiled by having DNS servers use random IDs in their queries rather than just counting, but it seems that every time one hole is plugged, another one turns up. The real problem is that the DNS was designed at a time when the Internet was a research facility for a few hundred universities. Security was not an issue then; making the Internet work was the issue. The environment has changed radically over the years, so in 1994,

the IETF set up a working group to make DNS fundamentally secure. This project is known as DNSsec (DNS security); its output is presented in RFC 2535. Unfortunately, DNSsec has not been fully deployed yet, so numerous DNS servers are still vulnerable to spoofing attacks. DNSsec is conceptually extremely simple. It is based on public key cryptography. Every DNS zone has a public/private key pair. All information sent by a DNS server is signed with the originating zone's private key, so the receiver can verify its authenticity. DNSsec offers three fundamental services:

1. Proof of where the data originated
2. Public key distribution
3. Transaction and request authentication

The main service is the first one, which verifies that the data being returned has been approved by the zone's owner. The second one is useful for storing and retrieving public keys securely. The third one is needed to guard against playback and spoofing attacks. Note that secrecy is not an offered service since all the information in DNS is considered public. Since phasing in DNSsec is expected to take several years, the ability for security-aware servers to interwork with security-ignorant servers is essential, which implies that the protocol cannot be changed.

Let us now look at some of the details. DNS records are grouped into sets called *RRSets* (Resource Record Sets), with all the records having the same name, class, and type being lumped together in a set. An RRSet may contain multiple A records, for example, if a DNS name resolves to a primary IP address and a secondary IP address. The RRSets are extended with several new record types (discussed below). Each RRSet is cryptographically hashed (*e.g.*, using MD5 or SHA-1). The hash is signed by the zone's private key (*e.g.*, using RSA). The unit of transmission to clients is the signed RRSet. Upon receipt of a signed RRSet, the client can verify whether it was signed by the private key of the originating zone. If the signature agrees, the data are accepted. Since each RRSet contains its own signature, RRSets can be cached anywhere, even at untrustworthy servers, without endangering the security. DNSsec introduces several new record types. The first of these is the *KEY* record. This record holds the public key of a zone, user, host, or other principal, the cryptographic algorithm used for signing, the protocol used for transmission, and a few other bits. The public key is stored naked. X.509 certificates are not used due to their bulk. The algorithm field holds a 1 for MD5/RSA signatures (the preferred choice), and other values for other

combinations. The protocol field can indicate the use of IPsec or other security protocols, if any. The second new record type is the *SIG* record. It holds the signed hash according to the algorithm specified in the *KEY* record. The signature applies to all the records in the RRSet, including any *KEY* records present, but excluding itself. It also holds the times when the signature begins its period of validity and when it expires, as well as the signer's name and a few other items. The DNSsec design is such that a zone's private key can be kept off-line. Once or twice a day, the contents of a zone's database can be manually transported (*e.g.*, on CD-ROM) to a disconnected machine on which the private key is located. All the RRSets can be signed there, and the *SIG* records thus produced can be conveyed back to the zone's primary server on CD-ROM. In this way, the private key can be stored on a CD-ROM locked in a safe, except when it is inserted into the disconnected machine for signing the day's new RRSets. After signing is completed, all copies of the key are erased from memory and the disk, and the CD-ROMs are returned to the safe. This procedure reduces electronic security to physical security, something people understand how to deal with. This method of presigning RRSets greatly speeds up the process of answering queries because no cryptography has to be done on the fly. The trade-off is that a large amount of disk space is needed to store all the keys and signatures in the DNS databases. Some records will increase tenfold in size due to the signature. When a client process gets a signed RRSet, it must apply the originating zone's public key to decrypt the hash, compute the hash itself, and compare the two values. If they agree, the data are considered valid. However, this procedure begs the question of how the client gets the zone's public key. One way is to acquire it from a trusted server, using a secure connection (*e.g.*, using IPsec). However, in practice, it is expected that clients will be preconfigured with the public keys of all the top-level domains. If Alice now wants to visit Bob's Web site, she can ask DNS for the RRSet of *bob.com*, which will contain his IP address and a *KEY* record containing Bob's public key. This RRSet will be signed by the top-level *com* domain, so Alice can easily verify its validity. An example of what this RRSet might contain. Now armed with a verified copy of Bob's public key, Alice can ask Bob's DNS server (run by Bob) for the IP address of *www.bob.com*. This RRSet will be signed by Bob's private key, so Alice can verify the signature on the RRSet Bob returns. If Trudy somehow manages to inject a false RRSet into any of the caches, Alice can easily detect its lack of authenticity because the *SIG* record contained in it will be incorrect. However, DNSsec also provides a cryptographic mechanism to bind a response to a specific query, to prevent the kind of spoof Trudy managed to pull off.

Table 6.4 An Example RRSet for *bob.com*. The *KEY* Record is Bob's Public One.

Domain name	Time to live	Class	Type	Value
bob.com.	86400	IN	A	36.1.2.3
bob.com.	86400	IN	KEY	3682793A7B73F731029CE2737D...
bob.com.	86400	IN	SIG	86947503A8B848F5272E53930C...

The *SIG* record is the top-level *com* server's signed hash of the *A* and *KEY* records to verify their authenticity. This (optional) anti-spoofing measure adds to the response a hash of the query message signed with the respondent's private key. Since Trudy does not know the private key of the top-level *com* server, she cannot forge a response to a query Alice's ISP sent there. She can certainly get her response back first, but it will be rejected due to its invalid signature over the hashed query. DNSsec also supports a few other record types. For example, the *CERT* record can be used for storing (e.g., X.509) certificates. This record has been provided because some people want to turn DNS into a PKI. Whether this actually happens remains to be seen. We will stop our discussion of DNSsec here. For more details, please consult RFC 2535.

6.16.2 Self-Certifying Names

Secure DNS is not the only possibility for securing names. A completely different approach can be used: the *Secure File System* (Mazieres et al., 1999). For this, the authors designed a secure, scalable, worldwide file system, without modifying the (standard) DNS and without using certificates or assuming the existence of a PKI. In this section, we show how their ideas are applied to the Web. Accordingly, in the description below, we will use Web terminology rather than the file system terminology used in the paper. But to avoid any possible confusion, while this scheme could be applied to the Web to achieve high security, it is not currently in use and would require substantial software changes to introduce it. We start out by assuming that each Web server has a public/private key pair. The essence of the idea is that each URL contains a cryptographic hash of the server's name and public key as part of the URL. We see the URL for Bob's photo. It starts out with the usual *http* scheme followed by the DNS name of the server (*www.bob.com*). Then comes a colon and 32-character hash. At the end is the name of the file, again as usual. Except for

the hash, this is a standard URL. With the hash, it is a *self-certifying URL*. The obvious question is: What is the hash for? The hash is computed by concatenating the DNS name of the server with the server's public key and running the result through the SHA-1 function to get a 160-bit hash.

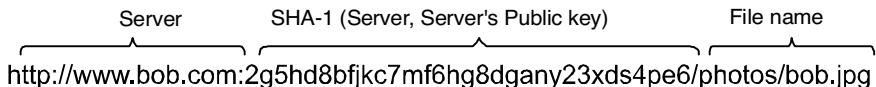


FIGURE 6.22 A self-certifying URL containing a hash of the server's name and public key

In this scheme, the hash is represented as a sequence of 32 digits and lower-case letters, with the exception of the letters “l” and “o” and the digits “1” and “0”, to avoid confusion. This leaves 32 digits and letters. With 32 characters available, each one can encode a 5-bit string. A string of 32 characters can hold the 160-bit SHA-1 hash. Actually, it is not necessary to use a hash; the key itself could be used. The advantage of the hash is to reduce the length of the name. In the simplest (but least convenient) way to see Bob's photo, Alice just types this in. The browser sends a message to Bob's Web site asking him for his public key. When Bob's public key arrives, the browser concatenates the server name and public key and runs the hash algorithm. If the result agrees with the 32-character hash in the secure URL, the browser is sure it has Bob's public key. After all, due to the properties of SHA-1, even if Trudy intercepts the request and forges the reply, she has no way to find a public key that gives the expected hash. Any interference from her will thus be detected. Bob's public key can be cached for future use. Now, Alice has to verify that Bob has the corresponding private key. She constructs a message containing a proposed AES session key, a nonce, and a timestamp. She then encrypts the message with Bob's public key and sends it to him. Since only Bob has the corresponding private key, only Bob is able to decrypt the message and send back the nonce encrypted with the AES key. Upon receiving the correct AES-encrypted nonce, Alice knows she is talking to Bob. Also, Alice and Bob now have an AES session key for subsequent *GET* requests and replies. Once Alice has Bob's photo (or any Web page), she can bookmark it, so she does not have to type in the full URL again. Furthermore, the URLs embedded in Web pages can also be self-certifying, so they can be used by just clicking on them, but with the additional security of knowing that the page returned is the correct one. Other ways to avoid the initial typing of the self-certifying URLs are to get them over a secure connection to a trusted server or have them present in the X.509 certificates signed by CAs.

Another way to get self-certifying URLs would be to connect to a trusted search engine by typing in its self-certifying URL (the first time) and going through the same protocol as described above, leading to a secure, authenticated connection to the trusted search engine. The search engine could then be queried, with the results appearing on a signed page full of self-certifying URLs that could be clicked on without having to type in long strings. Let us now see how well this approach stands up to Trudy's DNS spoofing. If Trudy manages to poison the cache of Alice's ISP, Alice's request may be falsely delivered to Trudy rather than to Bob. But the protocol now requires the recipient of an initial message (*i.e.*, Trudy) to return a public key that produces the correct hash. If Trudy returns her own public key, Alice will detect it immediately because the SHA-1 hash will not match the self-certifying URL. If Trudy returns Bob's public key, Alice will not detect the attack, but Alice will encrypt her next message, using Bob's key. Trudy will get the message, but she will have no way to decrypt it to extract the AES key and nonce. Either way, all spoofing DNS can do is provide a denial-of-service attack.

6.17 THE SECURE SOCKETS LAYER

Secure naming is a good start, but there is much more to Web security. The next step is secure connections. We will now look at how secure connections can be achieved. When the Web burst into public view, it was initially used for just distributing static pages. However, before long, some companies got the idea of using it for financial transactions, such as purchasing merchandise by credit card, online banking, and electronic stock trading. These applications created a demand for secure connections. In 1995, Netscape Communications Corp, the then-dominant browser vendor, responded by introducing a security package called *SSL (Secure Sockets Layer)* to meet this demand. This software and its protocol are now widely used (also by Internet Explorer), so it is worth examining them in some detail. SSL builds a secure connection between two sockets, including

1. Parameter negotiation between client and server
2. Mutual authentication of client and server
3. Secret communication
4. Data integrity protection

We have seen these items before, so there is no need to elaborate on them. The positioning of SSL in the usual protocol stack is as follows. Effectively, it is a new layer interposed between the application layer and the transport layer, accepting requests from the browser and sending them down to TCP for transmission to the server. Once the secure connection has been established, SSL's main job is handling compression and encryption. When HTTP is used over SSL, it is called *HTTPS (Secure HTTP)*, even though it is the standard HTTP protocol. Sometimes it is available at a new port (443) instead of the standard port (80), though. As an aside, SSL is not restricted to being used only with Web browsers, but that is its most common application. The SSL protocol has gone through several versions. Below we will discuss only version 3, which is the most widely used version. SSL supports a variety of different algorithms and options. These options include the presence or absence of compression, the cryptographic algorithms to be used, and some matters relating to export restrictions on cryptography. The last is mainly intended to make sure that serious cryptography is used only when both ends of the connection are in the United States. In other cases, keys are limited to 40 bits, which cryptographers regard as something of a joke. Netscape was forced to put in this restriction in order to get an export license from the U.S. Government. SSL consists of two sub protocols, one for establishing a secure connection and one for using it.

Let us start out by seeing how secure connections are established. The connection establishment sub protocol starts with message 1 when Alice sends a request to Bob to establish a connection. The request specifies the SSL version Alice has and her preferences with respect to compression and cryptographic algorithms. It also contains a nonce, *RA*, to be used later. Now it is Bob's turn. In message 2, Bob makes a choice among the various algorithms that Alice can support and sends his own nonce, *RB*. Then in message 3, he sends a certificate containing his public key. If this certificate is not signed by some well-known authority, he also sends a chain of certificates that can be followed back to one. All browsers, including Alice's, come preloaded with about 100 public keys, so if Bob can establish a chain anchored at one of these, Alice will be able to verify Bob's public key. At this point Bob may send some other messages (such as a request for Alice's public key certificate). When Bob is done, he sends message 4 to tell Alice it is her turn. Alice responds by choosing a random 384-bit *premaster key* and sending it to Bob encrypted with his public key (message 5). The actual session key used for encrypting data is derived from the premaster key combined with both nonces in a complex way. After message 5 has been received, both Alice and Bob are able

to compute the session key. For this reason, Alice tells Bob to switch to the new cipher (message 6) and also that she is finished with the establishment sub protocol (message 7). Bob then acknowledges her (messages 8 and 9). However, although Alice knows who Bob is, Bob does not know who Alice is (unless Alice has a public key and a corresponding certificate for it, an unlikely situation for an individual). Therefore, Bob's first message may well be a request for Alice to log in using a previously established login name and password. The login protocol, however, is outside the scope of SSL. Once it has been accomplished, by whatever means, data transport can begin. As mentioned above, SSL supports multiple cryptographic algorithms. The strongest one uses triple DES with three separate keys for encryption and SHA-1 for message integrity. This combination is relatively slow, so it is mostly used for banking and other applications in which the highest security is required. For ordinary e-commerce applications, RC4 is used with a 128-bit key for encryption and MD5 is used for message authentication. RC4 takes the 128-bit key as a seed and expands it to a much larger number for internal use. Then it uses this internal number to generate a keystream. The keystream is XORed with the plaintext to provide a classical stream cipher, export versions also use RC4 with 128-bit keys, but 88 of the bits are made public to make the cipher easy to break. For actual transport, a second sub-protocol is needed. Messages from the browser are first broken into units of up to 16 KB. If compression is enabled, each unit is then separately compressed. After that, a secret key is derived from the two nonces and the premaster key is concatenated with the compressed text; the result hashed with the agreed-on hashing algorithm (usually MD5). This hash is appended to each fragment as the MAC. The compressed fragment plus MAC is then encrypted with the agreed-on symmetric encryption algorithm (usually by XORing it with the RC4 keystream). Finally, a fragment header is attached and the fragment is transmitted over the TCP connection.

Table 6.5 Layers (and Protocols) for a Home User Browsing with SSL.

Application (HTTP)
Security (SSL)
Transport (TCP)
Network (IP)
Data link (PPP)
Physical (modem, ADSL, cable TV)

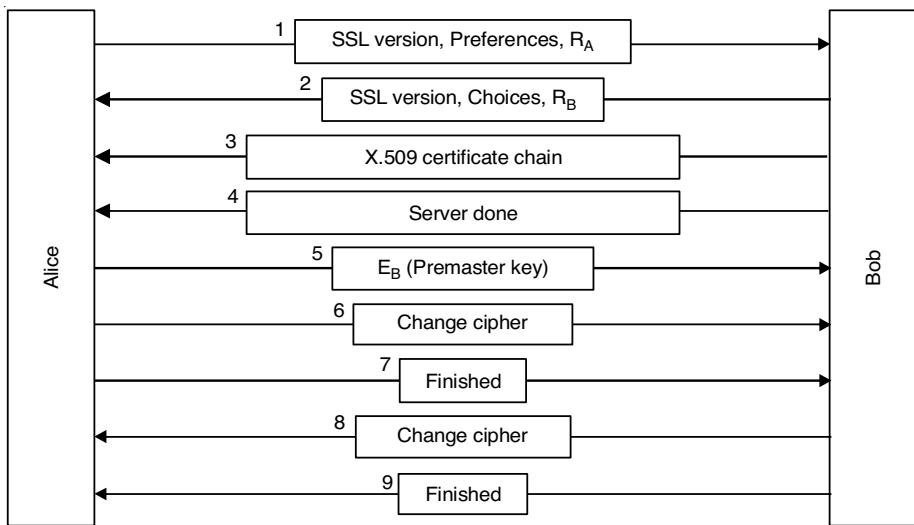


FIGURE 6.23 A simplified version of the SSL connection establishment sub-protocol

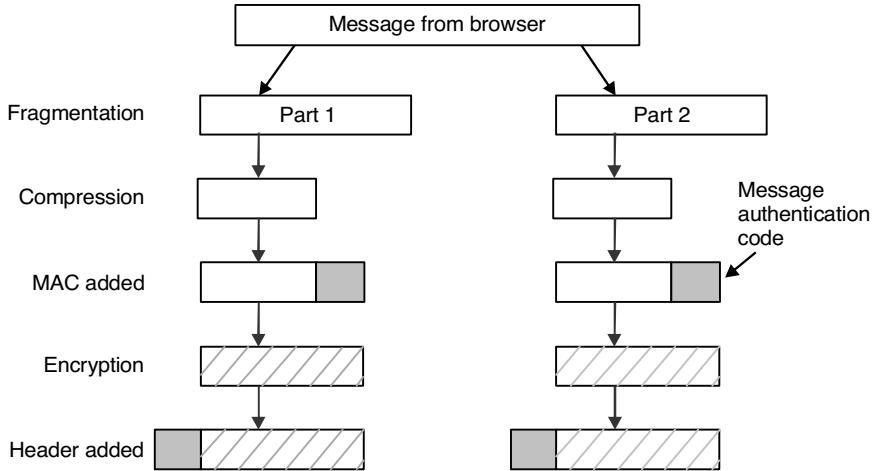


FIGURE 6.24 Data transmission using SSL

A word of caution is in order, however. Since it has been shown that RC4 has some weak keys that can be easily cryptanalyzed, the security of SSL using RC4 is on shaky ground (Fluhrer et al., 2001). Browsers that allow the user to choose the cipher suite should be configured to use triple DES with 168-bit keys and SHA-1 all the time, even though this combination is slower than RC4 and MD5. Another problem with SSL is that the principals may not have

certificates, and even if they do, they do not always verify that the keys being used match them. In 1996, Netscape Communications Corp. turned SSL over to IETF for standardization. The result was *TLS (Transport Layer Security)*. It is described in RFC 2246. The changes made to SSL were relatively small, but just enough that SSL version 3 and TLS cannot interoperate. For example, the way the session key is derived from the premaster key and nonces was changed to make the key stronger (*i.e.*, harder to cryptanalyze). The TLS version is also known as SSL version 3.1. The first implementations appeared in 1999, but it is not clear yet whether TLS will replace SSL in practice, even though it is slightly stronger. The problem with weak RC4 keys remains.

6.18 RSA ALGORITHM

As noted earlier, Diffie and Hellman introduced a new approach to cryptography, and challenged cryptologists to design a general-purpose encryption algorithm that satisfies the public key encryption requirements. One of the first responses to the challenge was developed in 1977 by Ron Rivest, Adi Shamir, and Len Adleman at MIT. Since then, the Rivest-Shamir-Adleman (RSA) scheme has become the most widely accepted and implemented general-purpose approach to public key encryption. Here, we examine the following aspects of the RSA algorithm:

1. Mathematical preliminaries of RSA
2. RSA algorithm description
3. Computational aspects of RSA
4. Threats to RSA
5. Public Key Cryptography Standards (PKCS)

Among these topics, we will focus on the following—how RSA operates, why it would work, and why it is secure. Students are encouraged to read additional books to understand the computational aspect and the security of RSA.

1. Mathematical Preliminaries

In this section, we introduce the mathematical background that helps us understand RSA.

A. Modular Addition

Let's start with one of the simplest ciphers: the general Caesar cipher. Its encryption and decryption operation can be represented using the following mathematical functions.

$$C = (P + K) \bmod 26 \quad \dots(1)$$

$$P = (C - K) \bmod 26 \quad \dots(2)$$

P\K	0	1	2	3	4	5	6	7	8	9
0	0	1	2	3	4	5	6	7	8	9
1	1	2	3	4	5	6	7	8	9	0
2	2	3	4	5	6	7	8	9	0	1
3	3	4	5	6	7	8	9	0	1	2
4	4	5	6	7	8	9	0	1	2	3
5	5	6	7	8	9	0	1	2	3	4
6	6	7	8	9	0	1	2	3	4	5
7	7	8	9	0	1	2	3	4	5	6
8	8	9	0	1	2	3	4	5	6	7
9	9	0	1	2	3	4	5	6	7	8

T

A

B

L

E

I

ADDITION MODULO 10

For simplicity, we replace 26 with 10, and show the general Caesar cipher, which is also the modular addition operation, in Table I. Note that the decryption operation, which subtracts the secret key K from cipher text

C modulo 10, can also be done by adding K' , which is the additive inverse of K modulo 10. An additive modular inverse of K is the number which is added to K to get 0 after modular operation. For example, 4's inverse (modulo 10) is 6, because $(4 + 6) \bmod 10 = 0$. If the secret key were 4, then to encrypt in general Caesar cipher, 4 is added to the plaintext; and to decrypt, 6 is added to the cipher text. Formally, we have

$$C = (P + K) \bmod 26 \quad \dots(3)$$

$$P = (C + K') \bmod 26 \quad \dots(4)$$

where

$$K + K' \bmod 10 = 0 \quad \dots(5)$$

B. Modular Multiplication

Now let's look at the mod 10 multiplication operation as shown in Table II. We note that only when $K = 1, 3, 7, 9$, the modular multiplication operation works as a cipher, because it only performs a one-to-one mapping between the plaintext and the cipher text in these cases. What is special about the numbers $\{1, 3, 7, 9\}$? The answer is that those numbers are all relatively prime to 10. Generally, a number K is relatively prime to n .

P\K	0	1	2	3	4	5	6	7	8	9
0	0	0	0	0	0	0	0	0	0	0
1	0	1	2	3	4	5	6	7	8	9
2	0	2	4	6	8	0	2	4	6	8
3	0	3	6	9	2	5	8	1	4	7
4	0	4	8	2	6	0	4	8	2	6
5	0	5	0	5	0	5	0	5	0	5
6	0	6	2	8	4	0	6	2	8	4
7	0	7	4	1	8	5	2	9	6	3
8	0	8	6	4	2	0	8	6	4	2
9	0	9	8	7	6	5	4	3	2	1

T

A

B

L

E

II

MULTIPLICATION MODULO 10

$$\gcd(K, n) = 1; 1 \leq K < n \quad \dots(6)$$

where \gcd denotes the greatest common divisor.

For decryption, we can look for multiplicative inverse, and undo the multiplication by multiplying the cipher text by the multiplicative inverse of the key. The multiplicative inverse of K , denoted by K^{-1} , is the number by which you'd multiply K to get 1 in mod n . Formally, the cryptosystem can be represented as

$$C = (P \cdot K) \bmod n \quad \dots(7)$$

$$P = (C \cdot K^{-1}) \bmod n \quad \dots(8)$$

where

$$K + K^{-1} \bmod n = 1 \quad \dots(9)$$

algorithm (Euclid's algorithm) to calculate K^{-1} based on K , while in public key encryption model, the private key cannot be derived from knowledge of the public key.

Let's further explore other mathematical functions. Before that, we'd examine the question how many numbers less than n are relatively prime to n ? This number is denoted as $\Phi(n)$, and is called *totient function*. As we will see later, this number is quite important in the design of RSA. It is obvious that,

- when n is a prime, $\Phi(n) = n - 1$;
- when n is the product of two distinct primes p, q , (*i.e.*, $n = p \cdot q$, $p = q$ are primes), $\Phi(n) = (p - 1)(q - 1)$.

C. Modular Exponentiation

Now let's proceed to consider encryption and decryption using modular exponentiation operation.

$$C = (P^K) \bmod n \quad \dots(10)$$

$$P = (C^{K''}) \bmod n \quad \dots(11)$$

where K'' is the exponentiative inverse of K .

Just like multiplicative inverse, we may ask what kind of values of K has the exponentiative inverse? and how is its inverse calculated? The answers to these questions led to the design of RSA. In what follows, we give a description of RSA algorithm first, then discuss how it is related with modular exponentiation.

2. RSA Description

The RSA scheme is a block cipher. Each plaintext block is an integer between 0 and $n - 1$ for some n , which leads to a block size $\leq \log_2(n)$. The typical block size for RSA is 1024 bits. The details of the RSA algorithm are described as follows.

Key Generation

1. Pick two large prime numbers p and q , $p = q$;
2. Calculate $n = p \times q$;
3. Calculate $\Phi(n) = (p - 1)(q - 1)$;
4. Pick e , so that $\gcd(e, \Phi(n)) = 1$, $1 < e < \Phi(n)$;
5. Calculate d , so that $d \cdot e \bmod \Phi(n) = 1$, i.e., d is the multiplicative inverse of e in $\bmod \Phi(n)$;
6. Get public key as $K_U = \{e, n\}$;
7. Get private key as $K_R = \{d, n\}$.
 - **Encryption**
For plaintext block $P < n$, its cipher text $C = P^e \bmod n$.
 - **Decryption**
For cipher text block C , its plaintext is $P = C^d \bmod n$.

A. Why RSA Works

As we have seen from the RSA design, RSA algorithm uses modular exponentiation operation. For $n = p \cdot q$, e which is relatively prime to $\Phi(n)$, has an exponential inverse in mod n . Its exponential inverse d can be calculated as the multiplicative inverse of e in mod $\Phi(n)$. The reason is illustrated as follows.

Based on Euler's theorem, for y which satisfies $y \mod \Phi(n) = 1$, the following equation holds true.

$$x^y \mod n = x \mod n \quad \dots(12)$$

As $d \cdot e \mod \Phi(n) = 1$, we have that $P^{ed} \equiv P \mod n$. So the correctness of RSA cryptosystem is

- **Encryption:** $C = P^e \mod n$;
- **Decryption:** $P = C^d \mod n = (P^e)d \mod n = P^{ed} \mod n = P \mod n = P$.

B. Why RSA is Secure

The premise behind RSA's security is the assumption that factoring a big number (n into p , and q) is hard. And thus it is difficult to determine $\Phi(n)$. Without the knowledge of $\Phi(n)$, it would be hard to derive d based on the knowledge of e .

However factoring n is not the only way to break RSA. Students are encouraged to read the suggested material to find out more threats to RSA.

EXERCISES

1. Explain how securing information on a network involves cryptography.
2. What is the plaintext needed to produce a stream of Secret?
3. What are block ciphers?
4. What is a plaintext attack?
5. Explain public key cryptosystems.
6. Explain symmetric key encryption.
7. Explain the Data Encryption Standard (DES).
8. Discuss the Secret key exchange.
9. What is RSA?
10. What is the Secure Sockets Layer?
11. Explain web security.

REMOTE ACCESS WITH INTERNET PROTOCOL SECURITY

Chapter Goals

- Identify different types of wireless technologies.
- Identify different wireless solutions.
- Introduce quadrature amplitude modulation.
- Explain wireless systems.
- Discuss the benefits of using wireless technologies for communications.

7.1 WIRELESS TECHNOLOGIES

To support networking solutions that consumer electronic devices and appliances can plug into, Microsoft is working on a range of wireless technologies to enable a robust set of user scenarios for Local Area Networks (LANs), Personal Area Networks (PANs), and Wide Area Networks (WANs).

Windows provides extensive Native 802.11 support, which is the widely adopted standard for high-speed networking across wireless Local Area Networks (WLANS).

7.1.1 Types of Wireless Technology

Eighteen major types of wireless technologies exist, containing a large number of subset technologies that range from ATM-protocol based (which sells at approximately \$200,000 per data link) to wireless local-area network (WLAN, which sells at less than \$500,000 per data link). Frequencies of the different technologies travel between several hundred feet (wireless LAN) and 25 miles (MMDS).

The process by which radio waves are propagated through the air, the amount of data carried, immunity to interference from internal and external sources, and a host of other characteristics varies from technology to technology.

Wireless technologies are differentiated by the following:

- **Protocol**—ATM or IP
- **Connection type**—Point-to-Point (P2P) or multipoint (P2MP) connections
- **Spectrum**—Licensed or unlicensed

Table 7.1 The Different Wireless Technologies.

Broadband ¹	Narrowband
WAN	WAN and WLAN
Licensed ²	Unlicensed
Digital	Analog
Line-of-sight ³	Non-line-of-sight
Simplex ⁴	Half-/full-Duplex
Point-to-point	Multipoint

¹Broadband—Data rates that exceed 1.5 Mbps

²Licensed—Granted by or purchased from the FCC

³Line-of-sight—Direct line of sight between two antennae

⁴Simplex—One transmitter

7.2 BASE STATION

The *base station* (also referred to as the hub or the cell site) is the central location that collects all traffic to and from subscribers within a cell. The indoor base station equipment consists of channel groups. The channel groups each

connect to the existing network, typically with a DS-3 with ATM signaling. The function of the channel group is to effectively act as a high-speed radio modem for the DS-3 traffic. The outdoor base station equipment (Tx/Rx node) modules are located on a tower or a rooftop mount and consist of frequency translation hardware and transmitters/receivers. The Tx/Rx node delivers and collects all the traffic to and from subscribers within a cell or a sector. Additionally, the Tx/Rx node equipment translates the channel group output into the appropriate frequency for over-the-air transmission. Multiple channel groups are used in each sector to meet the traffic demands, thus providing a highly scalable architecture.

7.3 TECHNOLOGY OF OFFLINE MESSAGE KEYS

Many modern fixed microwave communication systems are based on Quadrature Amplitude Modulation (QAM). These systems have various levels of complexity.

Simpler systems, such as Phase Shift Keying (PSK), are very robust and easy to implement because they have low data rates. In PSK modulation, the shape of the wave is modified in neither amplitude nor frequency, but rather in phase. The phase can be thought of as a shift in time. In Binary Phase Shift Keying (BPSK), the phases for the sine wave start at either 0 or $1/4$. In BPSK modulation, only 1 bit is transmitted per cycle (called a *symbol*). In more complex modulation schemes, more than 1 bit is transmitted per symbol. The modulation scheme QPSK (quadrature phase shift keying) is similar to the BPSK. However, instead of only two separate phase states, QPSK uses four ($0, 1/2, 1/4$, and $3/2, 1/4$), carrying 2 bits per symbol. Like BPSK, QPSK is used because of its robustness. However, because it modulates only 2 bits per symbol, it still is not very efficient for high-speed communications. Hence, higher bit rates require the use of significant bandwidth.

Even though QPSK uses no state changes in amplitude, it is sometimes referred to as 4-QAM. When four levels of amplitude are combined with the four levels of phase, we get 16-QAM. In 16 QAM, 2 bits are encoded on phase changes, and 2 bits are encoded on amplitude changes, yielding a total of 4 bits per symbol.

In Figure 7.1, each unique phase is spaced equally in both the I and Q coordinates. The angle of rotation indicates the phase, and the distance from the center point indicates the amplitude. This approach to modulation can be expanded out to 64-QAM and 256-QAM or higher. Although 64-QAM

is very popular in both cable and wireless broadband products, 256-QAM is also being tested. The higher the density in QAM, the higher a signal-to-noise (s/n) ratio must be maintained to meet the required Bit-Error Rates (BERs).

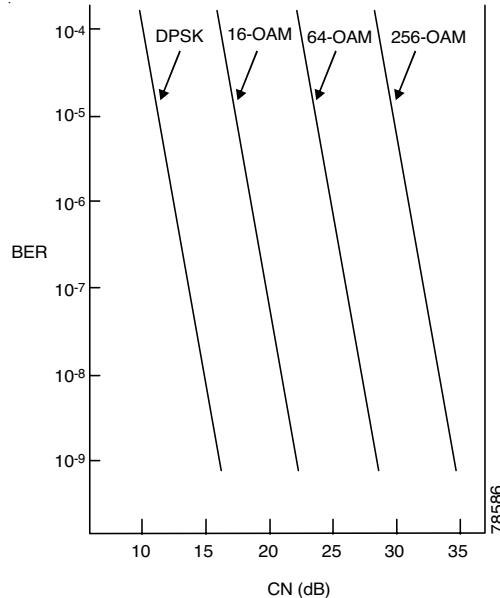


FIGURE 7.1 Error rates for PSK and QAM systems

How the data is encoded also plays an important part in the equation. The data is usually scrambled, and a significant amount of Forward Error Correction (FEC) data is also transmitted. Therefore, the system can recover those bits that are lost because of noise, multipath, and interference. A significant improvement in BER is achieved using FEC for a given SNR at the receiver (see Figure 7.2).

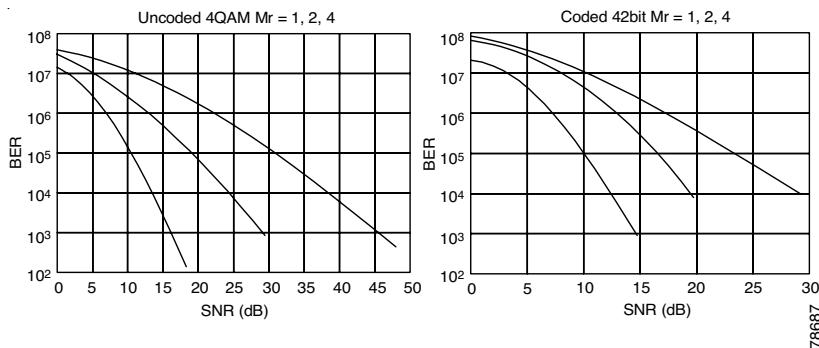


FIGURE 7.2 BER against signal-to-noise for coded and uncoded data streams

7.4 ADVANCED SIGNALING TECHNIQUES USED TO MITIGATE MULTIPATH

Several techniques have been used to make digital modulation schemes more robust: QAM with Decision Feedback Equalization (DFE), Direct Sequence Spread Spectrum (DSSS), Frequency-Division Multiplexing (FDM), and Orthogonal Frequency-Division Multiplexing (OFDM).

7.4.1 QAM with DFE

In wireless QAM systems, DFE is used to mitigate the effects of the Intersymbol Interference (ISI) caused by multipath. When delay spread is present, the echoes of previous symbols corrupt the sampling instant for the current symbol. The DFE filter oversamples the incoming signal and filters out the echoed carriers. The complexity of DFE schemes causes them not to scale with increases in bandwidth. The complexity of the DFE filter (number of taps) is proportional to the size of the delay spread. The number of required taps is proportional to the delay spread (in seconds) multiplied by the symbol rate.

For a QAM-based wireless system transmitting in the MMDS band (6-MHz-wide channel) to survive a 4- μ sec delay spread, the number of taps required would equal 24. To equalize a system with 24 taps, a DFE system would need 72 feedforward and 24 feedback taps. In addition to the number of taps needed, the complexity of the math needed for each tap increases with the number of taps. Therefore, the increase in complexity becomes an exponential function of the bandwidth of the carrier signal. Figure 7.3 compares the complexity rate of QAM/DFE and OFDM. Orthogonal Frequency-Division Multiplexing (OFDM) is discussed later in this paper.

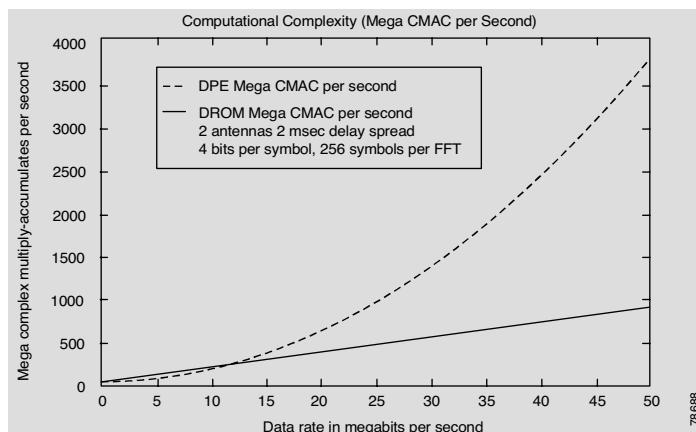


FIGURE 7.3 Computational complexity of QAM versus OFDM

7.4.2 Spread Spectrum

The *spread spectrum* is a method commonly used to modulate the information into manageable bits that are sent over the air wirelessly. Spread spectrum was invented by Hedy Lamar, a film actress who retained the patent and was the recipient of a governmental award for this accomplishment.

Essentially, spread spectrum refers to the concept of splitting information over a series of radio channels or frequencies. Generally, the number of frequencies is in the range of about 70, and the information is sent over all or most of the frequencies before being demodulated, or combined at the receiving end of the radio system.

Two kinds of spread spectrum are available:

- Direct Sequence Spread Spectrum (DSSS)
- Frequency Hopping Spread Spectrum (FHSS)

DSSS typically has better performance, while FHSS is typically more resilient to interference.

A commonly used analogy to understand spread spectrum is that of a series of trains departing a station at the same time. The payload is distributed relatively equally among the trains, which all depart at the same time. Upon arrival at the destination, the payload is taken off each train and is collated. Duplications of the payload are common to the spread spectrum so that when data arrives excessively corrupted, or fails to arrive, the redundancies inherent to this architecture provide a more robust data link.

The *direct sequence spread spectrum* (DSSS) is a signaling method that avoids the complexity and the need for equalization. Generally, a narrowband QPSK signal is used. This narrowband signal is then multiplied (or spread) across a much wider bandwidth. The amount of spectrum needed is expressed as: $10(\text{SNR}/10) \times \text{narrowband symbol rate}$.

Therefore, if a SNR of 20 dB is required to achieve the appropriate BER, the total spread bandwidth needed to transmit a digital signal of 6 Mbps equals 600 MHz.

This is not very bandwidth-efficient. In addition, the sampling rate for the receiver needs to be about 100 times the data rate. Therefore, for this hypothetical system, the sampling rate would also need to be 600 megasamples per second.

With DSSS, all trains leave in an order, beginning with Train 1 and ending with Train N, depending on how many channels the spread spectrum system allocates. In the DSSS architecture, the trains always leave in the same order, although the numbers of railroad tracks can be in the hundreds or even thousands.

Code Division Multiple Access (CDMA) is used to allow several simultaneous transmissions to occur. Each data stream is multiplied with a pseudorandom noise code (PN code). All users in a CDMA system use the same frequency band. Each signal is spread out and layered on top of each other and is overlaid using code spreading in the same time slot. The transmitted signal is recovered by using the PN code. Data transmitted by other users looks like white noise and drops out during the reception phase. Any narrow-band noise is dispersed during the de-spreading of the data signal. The advantage of CMDA is that the amount of bandwidth required is now shared over several users. However, in systems in which there are multiple transmitters and receivers, proper power management is needed to ensure that one user does not overpower other users in the same spectrum. These power management issues are mainly confined to CMDA architectures.

7.4.3 FHSS

With the FHSS architecture, the trains leave in a different order—that is, not sequentially from Train 1 to Train N. In the best of FHSS systems, trains that run into interference are not sent out again until the interference abates. In FHSS systems, certain frequencies (channels) are avoided until the interference abates.

Interference tends to cover more than one channel at a time. Therefore, DSSS systems tend to lose more data from interference as the data sent out is done so over sequential channels. FHSS systems hop between channels in nonsequential order. The best of FHSS systems adjust channel selection so that highly interfered channels are avoided as measured by excessively low bit error rates. Either approach is appropriate and depends on customer requirements, with the selection criteria primarily being that of a severe multipath or interfering RF environment.

7.4.4 FDM

In a *Frequency-Division Multiplexing* (FDM) system, the available bandwidth is divided into multiple data carriers. The data to be transmitted is then divided among these sub-carriers. Because each carrier is treated independently of the others, a frequency guard band must be placed around it. This guard band lowers the bandwidth efficiency. In some FDM systems, up to 50 percent of the available bandwidth is wasted. In most FDM systems, individual users are segmented to a particular sub-carrier; therefore, their burst rate cannot exceed the capacity of that sub-carrier. If some sub-carriers are idle, their bandwidth cannot be shared with other sub-carriers.

7.4.5 OFDM

In orthogonal frequency division multiplexing (OFDM, Figure 7.4), multiple carriers (or tones) are used to divide the data across the available spectrum, similar to FDM. However, in an OFDM system, each tone is considered to be orthogonal (independent or unrelated) to the adjacent tones and, therefore, does not require a guard band. Because OFDM requires guard bands only around a set of tones, it is more efficient spectrally than FDM. Because OFDM is made up of many narrowband tones, narrowband interference will degrade only a small portion of the signal and has no or little effect on the remainder of the frequency components.

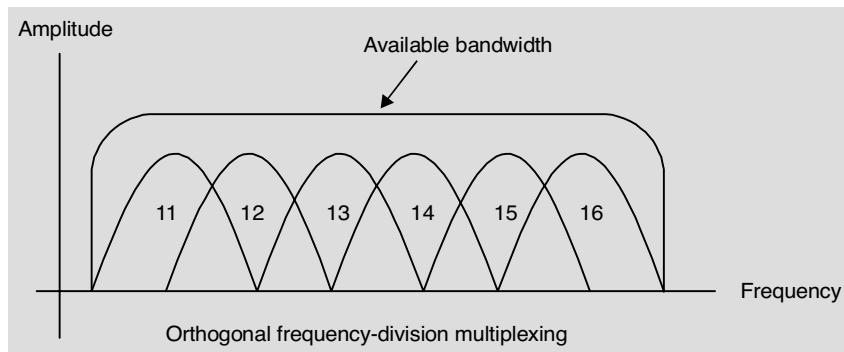


FIGURE 7.4 Example of OFDM tones

OFDM systems use bursts of data to minimize the ISI caused by delay spread. Data is transmitted in bursts, and each burst consists of a cyclic prefix followed by data symbols. An example OFDM signal occupying 6 MHz is made up of 512 individual carriers (or tones), each carrying a single QAM symbol per burst. The cyclic prefix is used to absorb transients from previous bursts caused by multipath signals. An additional 64 symbols are transmitted for the cyclic prefix. For each symbol period, a total of 576 symbols are transmitted by only 512 unique QAM symbols per burst. In general, by the time the cyclic prefix is over, the resulting waveform created by the combining multipath signals is not a function of any samples from the previous burst. Hence, there is no ISI. The cyclic prefix must be greater than the delay spread of the multipath signals. In a 6-MHz system, the individual sample rate is 0.16 μ secs. Therefore, the total time for the cyclic prefix is 10.24 μ secs, greater than the anticipated 4 μ secs delay spread.

7.4.6 VOFDM

In addition to the standard OFDM principles, the use of spatial diversity can increase the system's tolerance to noise, interference, and multipath. This is referred to as *vectored OFDM*, or VOFDM. Spatial diversity is a widely accepted technique for improving performance in multipath environments. Because multipath is a function of the collection of bounced signals, that collection is dependent on the location of the receiver antenna. If two or more antennae are placed in the system, each would have a different set of multipath signals. The effects of each channel would vary from one antenna to the next, so carriers that may be unusable on one antenna may become usable on another. Antenna spacing is at least ten times the wavelength.

Significant gains in the S/N are obtained by using multiple antennae. Typically, a second antenna adds about 3 dB in LOS and up to 10 dB in non-LOS environments.

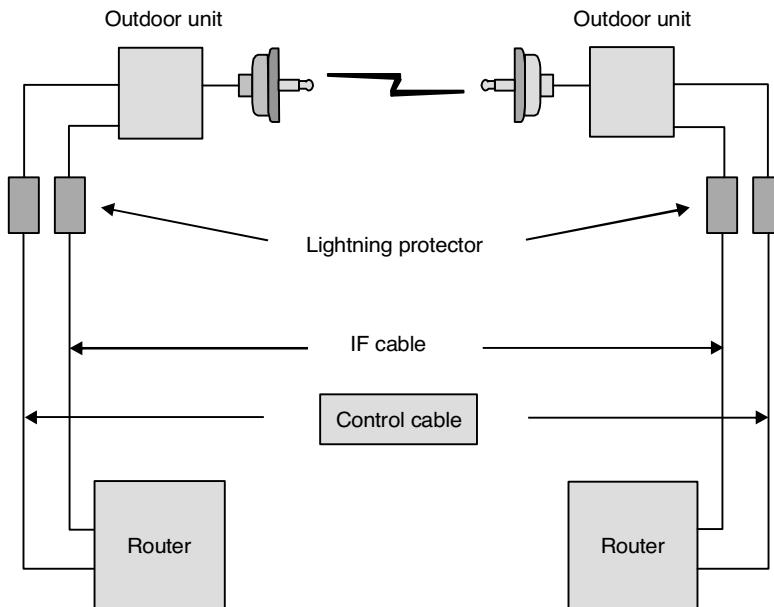


FIGURE 7.5 Spectrum technology

7.5 BENEFITS OF USING WIRELESS SOLUTIONS

The following list summarizes the main benefits of using wireless technologies:

- **Completes the access technology portfolio:** Customers commonly use more than one access technology to service various parts of their network and during the migration phase of their networks, when upgrading occurs on a scheduled basis. Wireless enables a fully comprehensive access technology portfolio to work with existing dial, cable, and DSL technologies.
- **Goes where cable and fiber cannot:** The inherent nature of wireless is that it doesn't require wires or lines to accommodate the data/voice/video pipeline. As such, the system will carry information across geographical areas that are prohibitive in terms of distance, cost, access, or time. It also sidesteps the numerous issues of ILEC colocation.

Although paying fees for access to elevated areas such as masts, towers, and building tops is not unusual, these fees, the associated logistics, and contractual agreements are often minimal compared to the costs of trenching cable.

- **Involves reduced time to revenue:** Companies can generate revenue in less time through the deployment of wireless solutions than with comparable access technologies because a wireless system can be assembled and brought online in as little as two to three hours.

This technology enables service providers to sell access without having to wait for cable-trenching operations to complete or for incumbent providers to provide access or backhaul.

- **Provides broadband access extension:** Wireless commonly both competes with and complements existing broadband access. Wireless technologies play a key role in extending the reach of cable, fiber, and DSL markets, and it does so quickly and reliably. It also commonly provides a competitive alternative to broadband wireline or provides access in geographies that don't qualify for loop access.

7.6 EARTH CURVATURE CALCULATION FOR LINE-OF-SIGHT SYSTEMS

Line-of-sight systems that carry data over distances in excess of 10 miles require additional care and calculations. Because the curvature of the Earth

causes bulges at the approximate rate of 10 feet for every 18 miles, a calculation is required to the maintain line-of-sight status.

The Fresnel (pronounced fren-NEL) zone refers to that which must clear the Earth's bulge and other obstructions. This is the elliptically shaped free space area directly between the antennae. The center area in this zone is of the greatest importance and is called the first Fresnel zone. Although the entire Fresnel zone covers an area of appreciable diameter between the antennae, the first Fresnel zone is considered as a radius about the axis between the antennae. A calculation is required to determine the radius (in feet) that must remain free from obstruction for optimal data transfer rates. The formula for this calculation is: $D2/8$.

Table 7.2 helps to calculate the distance/bulge ratio.

Table 7.2 Wireless Distance Calculations.

Distance (Miles)	Earth bulge
8	8.0
10	12.5
12	18.0
14	24.5
16	32.0

While observing these calculations, it's important to remember that this accounts only for Earth bulge. Vegetation such as trees and other objects such as buildings must have their elevations added into this formula. A reasonable rule of thumb is 75 feet of elevation at both ends of the data link for a distance of 25 miles, but this should be considered an approximation only.

$$Rft = 72.1 \sqrt{d1 \cdot d2 / FD}$$

where

Rft = radius of the first Fresnel zone in feet

F = carrier frequency

d1 = distance from the transmitter to the first path obstacle

d2 = distance from the path obstacle to the receiver

D = $d1 + d2$ (in miles)

The industry standard is to keep 60 percent of the first Fresnel zone clear from obstacles. Therefore, the result of this calculation can be reduced by up

to 60 percent without appreciable interference. This calculation should be considered as a reference only and does not account for the phenomenon of refraction from highly reflective surfaces.

7.7 MICROWAVE COMMUNICATION LINKS

Since the beginning of the development of microwave wireless transmission equipment, manufacturers and operators have tried to mitigate the effects of reflected signals associated with signal propagation. These reflections are called *multipath*. In real-world situations, microwave systems involve careful design to overcome the effects of multipath. Most existing multipath mitigation approaches fall well short of the full reliable information rate potential of many wireless communications systems. This section discusses how to create a digital microwave transmission system that not only can tolerate multipath signals, but that also can actually take advantage of them.

Digital microwave systems fall into two categories: wavelengths less than 10 GHz and wavelengths greater than 10 GHz (referred to as a *millimeter wave*). Several bands exist below 10 GHz for high-speed transmissions. These may be licensed bands, such as MMDS (2.5 GHz), or unlicensed bands, such as U-NII (5.7 GHz). Bands that are below 10 GHz have long propagation distances (up to 30 miles). They are only mildly affected by climatic changes, such as rain. These frequencies are generally not absorbed by objects in the environment. They tend to bound and thus result in a high amount of multipath.

Bands over 10 GHz, such as 24 GHz, LMDS (28 GHz), and 38 GHz, are very limited in terms of distance (less than 5 miles). They are also quite susceptible to the signal fades attributed to rain. Multipath tends not to be an issue because the transmission distances are less and because most of the multipath energy is absorbed by the physical environment. However, when these frequencies are used in highly dense urban areas, the signals tend to bounce off objects, such as metal buildings or metalized windows. The use of repeaters can add to the multipath propagation by delaying the received signal.

7.7.1 What is Multipath?

Multipath is the composition of a primary signal plus duplicate or echoed images caused by reflections of signals off objects between the transmitter and the receiver. In Figure 7.6, the receiver hears the primary signal sent

directly from the transmission facility, but it also sees secondary signals that are bounced off nearby objects.

These bounced signals arrive at the receiver later than the incident signal. Because of this misalignment, the out-of-phase signals causes intersymbol interference or distortion of the received signal. Although most of the multipath is caused by bounces off tall objects, multipath can also occur from bounces off low objects, such as lakes and pavement.

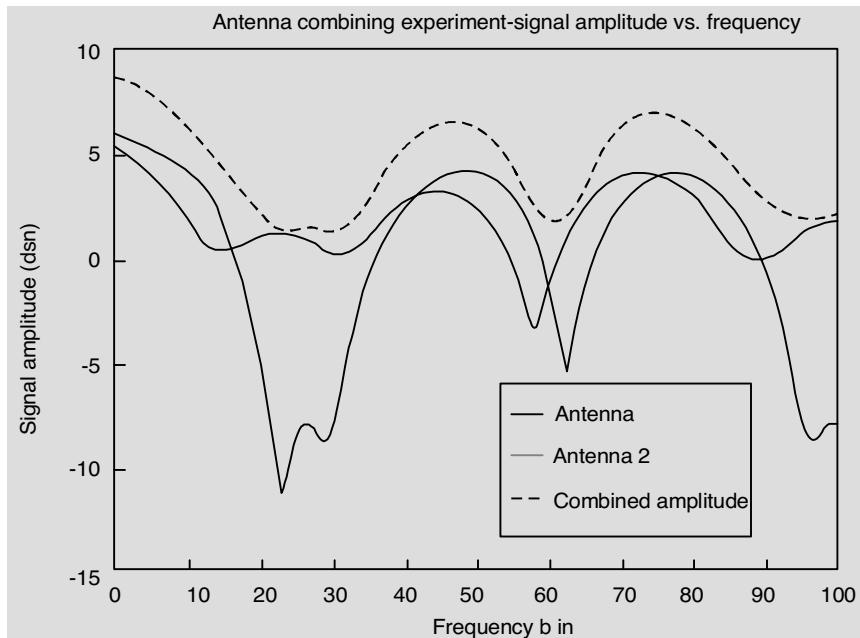


FIGURE 7.6 Multipath reception

The actual received signal is a combination of a primary signal and several echoed signals. Because the distance traveled by the original signal is shorter than the bounced signal, the time differential causes two signals to be received. These signals are overlapped and combined into a single one. In real life, the time between the first received signal and the last echoed signal is called the delay spread, which can be as high as 4 μ sec.

In the example shown in Figure 7.7, the echoed signal is delayed in time and reduced in power. Both are caused by the additional distance that the bounced signal traveled over the primary signal. The greater the distance, the longer the delay and the lower the power of the echoed signal. You might think that the longer the delay, the better off the reception would be. However, if

the delay is too long, the reception of an echoed symbol S1 and the primary symbol S2 can also interact. Because there may be no direct path for the incident signal in non-line-of-sight (LOS) environments, the primary signal may be small in comparison to other secondary signals.

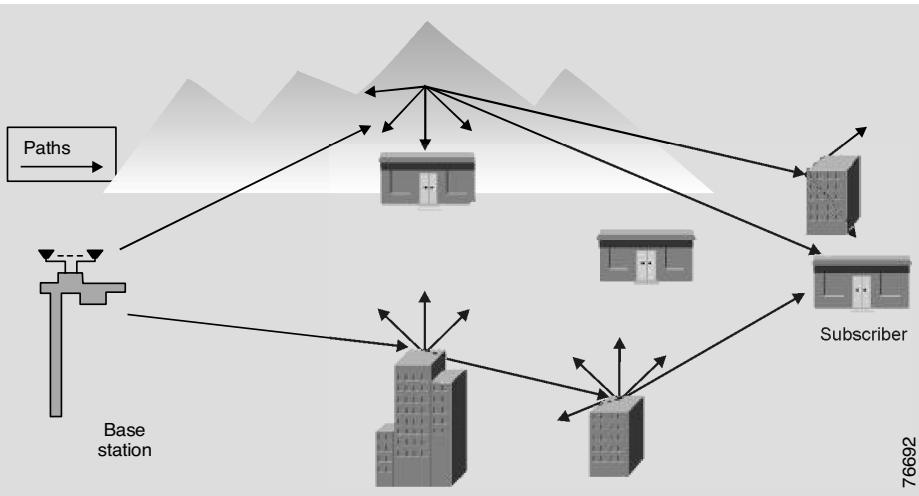


FIGURE 7.7 Typical multipath example

In analog systems such as television, this multipath situation can actually be seen by the human eye. Sometimes, there is a ghost image on your television, and no matter how much you adjust the set, the image does not go away. In these analog systems, this is an annoyance. In digital systems, it usually corrupts the data stream and causes the loss of data or a lower performance. Correction algorithms must be put in place to compensate for the multipath, resulting in a lower available data rate.

In digital systems, the input signal is sampled at the symbol rate. The echoed signal actually interferes with the reception of the second symbol, thus causing intersymbol interference (ISI). This ISI is the main result of multipath, and digital systems must be designed to deal with it.

7.7.2 Multipath in Non-LOS Environments

In LOS environments, multipath is usually minor and can be overcome easily. The amplitudes of the echoed signals are much smaller than the primary one and can be effectively filtered out using standard equalization techniques. However, in non-LOS environments, the echoed signals may have higher power levels because the primary signal may be partially or totally obstructed,

and generally because more multipath is present. This makes the equalization design more difficult.

In all the previous discussions, the multipath has been a semifixed event. However, other factors such as moving objects enter into play. The particular multipath condition changes from one sample period to the next. This is called *time variation*. Digital systems must be capable of withstanding fast changes in the multipath conditions, referred to as *fast fading*. To deal with this condition, digital systems need fast AGC circuits. Adaptive equalizers, discussed next, need fast training times.

7.8 ELEMENTS OF A TOTAL NETWORK SOLUTION

The issue of what comprises a solution is the subject of considerable discussion and conjecture. Commonly, the term solution includes the following primary elements:

- Premises networks
- Access networks
- Core networks
- Network management
- Deployment

A fully comprehensive wireless solution must also include the issues of deployment, maintenance, legacy, migration, and value propositions. The scope of what comprises a fully comprehensive solution can readily exceed these items.

7.8.1 Premises Networks

Premises networks are the voice, data, or video distribution networks that exist or will exist within the subscriber premises. Typical points of demarcation between the access and premises networks for purposes of this discussion include channel banks, PBXs, routers, or multiservice access devices.

Customer premises equipment receives signals from the hub, translates them into customer-usable data, and transmits returning data back to the hub. The transmitter, the receiver, and the antenna are generally housed in a Compact Rooftop Unit (RTU) that is smaller than a satellite TV minidish. It is mounted on the subscriber's roof in a location where it will have a clear line

of sight to the nearest LMDS hub site. Installation includes semi-precision pointing to ensure maximum performance of the RF link.

The indoor unit, the Network Interface Unit (NIU), does the modulation, demodulation, in-building wire-line interface functions, and provides an intermediate frequency to the RTU. Many interfaces required by end customer equipment require the NIU to have a breadth of physical and logical interfaces. The NIUs are designed to address a range of targeted subscribers whose connectivity requirements may range from T1/E1, POTS, Ethernet, or any other standard network interface. These interfaces are provided by the NIU with Interworking Function (IWF) cards. Different types of IWF cards are required in the NIU to convert the inputs into ATM cells and provide the appropriate signaling. Common IWFs include 10BaseT, T1/E1 circuit emulation, and others. The NIU also has an IF that is translated by the CPE RTU.

7.8.2 Access Networks

The access networks are the transport and distribution networks that bridge the premises network and the core network demarcation points. For the purpose of this discussion, the primary means of providing the transport from an access network Point-of-Presence (POP) to the premises is radio and the distribution between access network POPs is either fiber or radio.

7.8.3 Core Networks

The core networks are the public or private backbone networks that, in a general sense, will be utilized by the access network operators to connect their multitude of regionally dispersed POPs and to interconnect to public service provider network elements. For the purpose of this discussion, the point of demarcation between the access network and the core network is a core switch that serves as an upstream destination point for a multitude of access network branches or elements.

7.8.4 Network Management

The glue that ties all the network elements together and supports the key information processing tasks that make a business run effectively is performed by the Network Management System (NMS) inclusive of Operational Support System (OSS) functionality. In its full implementation, the NMS is an exceptionally complex set of moderately- to highly-integrated software platforms. For the purpose of this section, the element managers necessary within each system-level piece of the access network are assumed, but the overarching NMS is beyond the intended scope.

Ideally, the NMS should provide end-to-end functionality throughout both the wireless and wireline elements of the network, including the backbone and the customer premises. A network management system performs the service, network, and element management across multivendor and multitechnology networks, including

- Topology management
- Connectivity management
- Event management

The functions of the network management system can be further outlined as follows:

- Integrated topology map that displays an entire set of nodes and links in the network, shown with mapped alarms
- Store of network-wide physical (nodes/links) and logical topology (circuits/PVCs) for inventory
- Customer care interface to provide network and end-user status
- Performance statistics on PCR, SCR, MBS, CDVT, and network/link status
- SLA reporting with customer partitioning and alerting of customer violations
- Alarm correlation and root-cause analysis
- Network simulations to test whether a problem was completely corrected
- Trouble ticketing/workforce management
- Performance reports based on the statistics collected, with customer and network views
- Usage-based billing for ATM connections
- Read-only CNM for viewing the network and connection

7.8.5 Deployment

As stated previously, tier 1 customers utilize Cisco's ecosystem of deployment partners. Deployment for systems covering BTA, MTA, or nationwide footprints requires the following areas of expertise and resources:

- Construction (towers, masts)
- Licensing (FCC and local compliance for RF, construction, and access)
- Site survey (RF environment evaluation)
- Integration (selection and acquisition of various RF components)
- Prime (customer engagement through contract)
- Finance (securing or provisioning of project financing)
- Installation (assembly of components)
- Provisioning (spare components)

7.9 BILLING AND MANAGEMENT OF WIRELESS SYSTEMS

The issue of billing and network management is a considerable one. In the most general terms, you should consider the wireless links as a network section managed by the standard Cisco IOS and SNMP tools. Accordingly, key customer items such as billing, dynamic host control, testing, and configuration are managed remotely with standard router tools, as indicated in Figure 7.8.

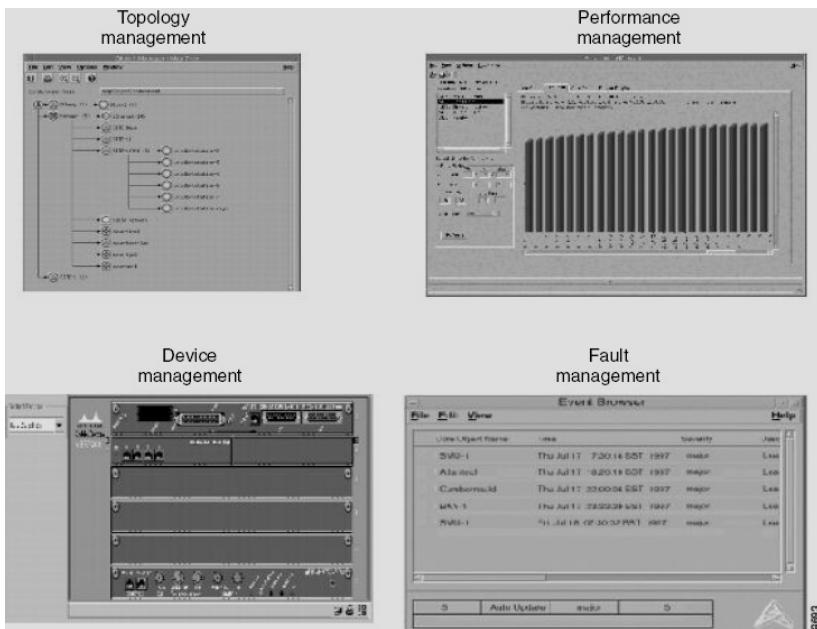


FIGURE 7.8 Management interfaces

7.9.1 Example Implementation

Cisco's MMDS/U-NII system is designed with the following objectives:

- For a service provider to offer differentiated services via wireless access, the wireless access system should offer a higher capacity than alternative access technologies.
- The capacity of the system is increased by three items:
 - A highly efficient physical layer that is robust to interference, resulting in high bandwidth efficiency per sector
 - A statistically efficient industry standard Medium Access Control (MAC) protocol that delivers Quality of Service (QoS)
 - A multicellular system
- A large bandwidth enables differentiated services such as Voice over IP (VoIP) now, and interactive video in the future, both with QoS
- A multitiered CPE approach, satisfying the needs of Small and Medium Businesses (SMB), small Office/Home Office (SOHO) applications, and residential customers
- Ease of base installation and back-haul
- Ease of provisioning and network management

7.10 IP WIRELESS SYSTEM ADVANTAGES

Table 7.3 summarizes the advantages of a wireless system.

Table 7.3 Features and Benefits of Wireless Communication.

Feature	Benefit
Shared-bandwidth system	Point-to-multipoint wireless architecture Shared bandwidth among many small and medium businesses Burst data rate up to 22 Mbps
Dedicated high-bandwidth System	Point-to-point wireless architecture High data rate (22 to 44 Mbps) Shared head end with point-to-multipoint equipment

(continued)

(continued)

Feature	Benefit
Small-cell and single-cell deployment	Variety of available cellular deployment plans Capability to scale with successful service penetration from tens of customers to thousands of customers Single cells of up to 45 km radius Small cells of up to 10 km radius for maximum revenue
Third-generation microwave technology	Higher-percentage coverage of customers in business district Capability of non-line-of-sight technology to service customers that older technology cannot service Capability to configure marginal RF links to improve performance Tolerant of narrow-band interference Receivers capable of adapting to changing environment for every packet Licensed frequency bands Options that include unlicensed 5.7 to 5.8 GHz bands
Open interfaces	Part of Cisco's dedication to open architectures Partners capable of supplying outdoor unit (ODU), antenna, cable, and all other components outside the router Availability of in-country manufacturers as partners Capability of MAC protocol to enhance DOCSIS, a proven industry multipoint standard
Integrated into Cisco routers	IOS system software and Cisco management software features that treat the radio link as simply another WAN interface Two systems created in one unit: a radio and a multiservice router Wireless integrated into management, provisioning, and billing systems Minimized cost of spare hardware
Native IP	Voice over IP Video over IP Virtual private networks Quality of service Queuing features Traffic policies

(continued)

Feature	Benefit
Cost-effective solution	Competitively priced Large pool of personnel already trained on Cisco routers and protocols Reduced training time Addition of broadband wireless access to Cisco's total end-to-end network solution and support
Link encryption	Privacy ensured through use of 40/56-bit DES encryption on every user's wireless link

7.11 IP WIRELESS SERVICES FOR SMALL AND MEDIUM BUSINESSES

The Small to Medium Business (SMB) customer requires services that range from typical Internet data access to business voice services. Most small businesses today have separate voice and data access lines. Almost all SMB customers use native IP in their networks. Voice access lines are typically analog POTS lines or Key Telephone System (KTS) trunks. As businesses grow, they may require a digital T1 trunk for their Private Branch Exchange (PBX). Data access is typically anything from dial to ISDN, fractional T1 Frame Relay, and potentially up to a dedicated leased line T1 service.

- SMB access technologies include
 - Plain Old Telephone Service (POTS)
 - KTS trunks
 - Digital T1 PBX trunks
 - Internet data access (Fast Ethernet)
- SMB service technologies include
 - Internet access (IP service)
 - Intranet access (VPN)
 - Voice services (VoIP)
 - Video conferencing
 - Service-level agreements for guaranteed data rates

- Residential access offerings include
 - POTS
 - Internet data access
- Residential service offerings include
 - Internet access (IP service)
 - Intranet access
 - Voice services (VoIP)
 - Video conferencing

7.12 IP POINT-TO-MULTIPOINT ARCHITECTURE

The Point-to-Multipoint (P2MP) system consists of a hub, or Head End (HE), or a Base Station (BS) which serves several sectors in the cell. Each sector consists of one radio communicating with many customers.

The head end is an outdoor unit, or transverter, connected to a wireless modem card inside a Cisco UBR7246 or 7223 router.

At the customers' premises is another transverter, which is connected to a wireless network module in a router.

The Cisco P2MP objectives are

- Provide an integrated end-to-end solution (one box, one management and provisioning platform)
- Complete multiservice offerings (Voice over IP, data, Video over IP)
- Scalability and flexibility (scalable head end and CPE offerings)
- Enabled for non-line-of-sight (substantially better coverage)
- Native IP packet transport
- Part of an overall standards-based strategy to provide many Cisco hosts and many frequency bands on a global basis

The shared-bandwidth, or multipoint, product delivers 1 to 22 Mbps aggregate full-duplex, shared-bandwidth, P2MP fixed-site data in the MMDS band for both residential and small business applications, as shown in Figure 7.9.

The P2MP wireless router is an integrated solution. At the base station (or head end or hub), it consists of a base universal router (UBR 7246 or

UBR7223), a wireless modem card, an Outdoor Unit (ODU) for the appropriate frequency band, cables, and antenna subsystems, as shown in Figure 7.10.

At the small business customer premises, the system consists of a network module in a 3600-family router, with an Outdoor Unit (ODU) and antenna. This CPE equipment is simpler and, therefore, less expensive than the Head End (HE) equipment. The 3600 family has a wide variety of interfaces to match all types of customer equipment.

At the SOHO or telecommuter customer premises, the system consists of a network module in a 2600- or 900-family router, with an Outdoor Unit (ODU) and an antenna. This CPE equipment is simpler and, therefore, less expensive than the Head End (HE) equipment. The 2600 and 900 families have a wide variety of interfaces to match all types of customer equipment. A consumer unit by one or more of Cisco's ecosystem partners is expected by the first quarter of 2001.

These Wireless Broadband Routers (WBBR) are then used to blanket an urban area by dividing the business district into small cells.

The product is also be capable of working as a single cell—that is, one hub serving an entire business district in a cell of radius less than 45-miles, because the low frequencies in the MMDS band are not impacted by rain. However, a single cell does not have the revenue potential of small cells.

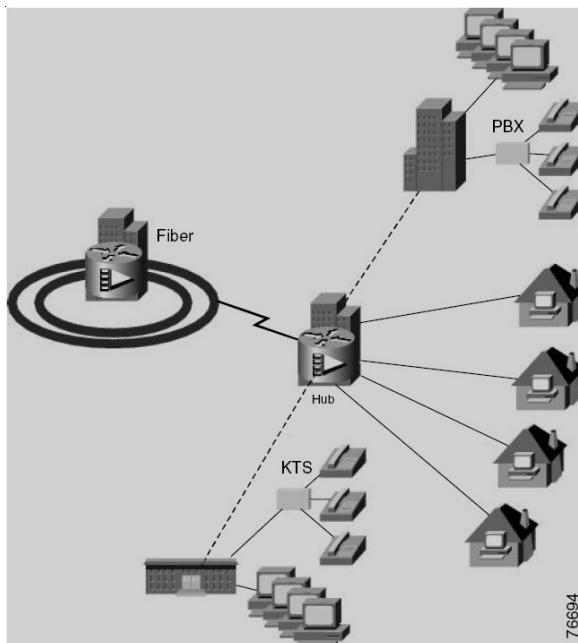


FIGURE 7.9 Basic components of the P2MP base station (head end, hub)

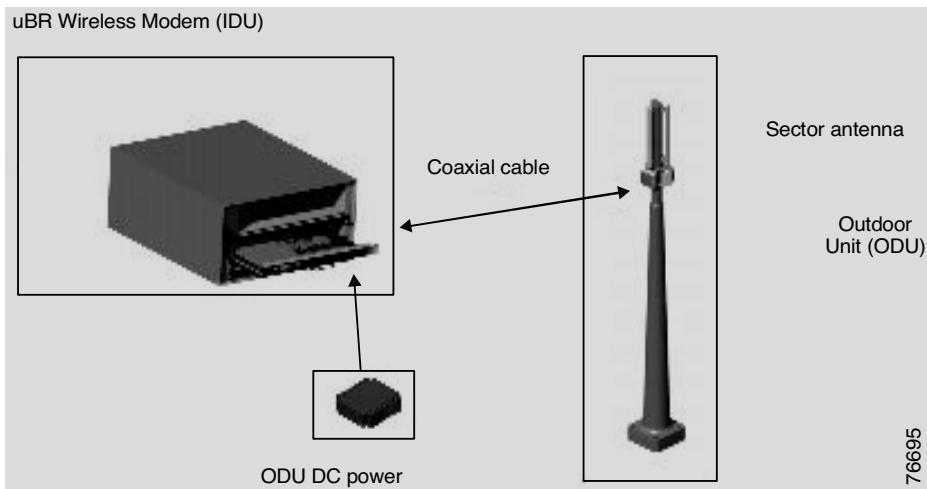


FIGURE 7.10 The Cisco MMDS broadband wireless access system

Technology has always been a Cisco Systems differentiator, and the proposed system fits that market position. The proposed P2MP system uses patented third-generation microwave technology to overcome the classical microwave constraint that the transmitter and receiver must have a clear line of sight. The proposed technology takes advantage of the waves that bounce off buildings, water, and other objects to create multiple paths from the transmitter to the receiver. The receivers are capable of making all these multipath signals combine into one strong signal, rather than having them appear to be interference.

The capability to operate with high levels of multipath permits obstructed links to be deployed. This, in turn, enables multicellular RF deployments and virtually limitless frequency reuse. The antennae in the system can be mounted on short towers or rooftops.

Although this is primarily a savings in the cost of installing a system, it has the added benefit of making the installation less visible to the user's neighbors, which is very important in some regions.

The typical point-to-multipoint system for an SMB is shown in Figure 7.11.

The point-to-point system is similar, except that the customer premises equipment is another UBR7223 or 7246 router. The point-to-point system is shown in Figure 7.12.

Some SMB customers will require a data rate that is higher than the service provider can supply within the traffic capacity of the multipoint system. The service provider may satisfy those customers by installing Cisco's Point-

to-Point (P2P) links from the same hub as the P2MP system. Thus, the hub can be a mixture of P2MP and P2P systems.

In both cases, integrating the wireless card directly into the router brings with it all the Cisco IOS features and network management.

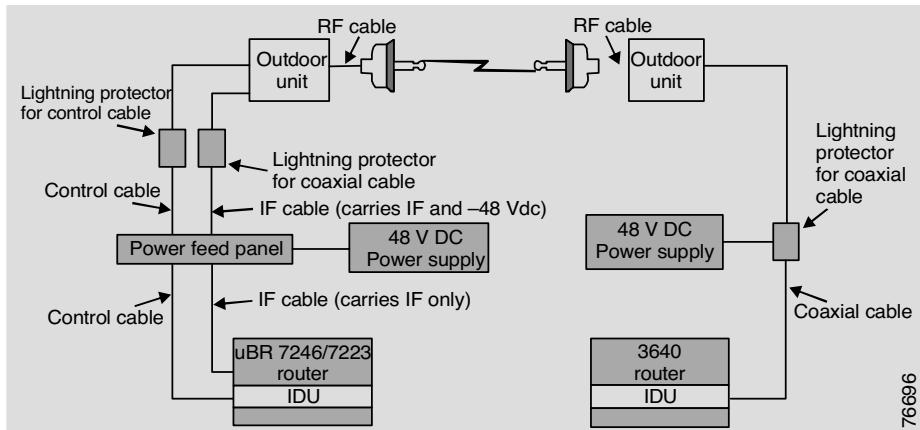


FIGURE 7.11 Hardware components of the point-to-point system (only one of four possible sectors shown on HE equipment)

7.13 IP WIRELESS OPEN STANDARDS

This open architecture permits many different vendors to participate, creating new products, features, and services. Figure 7.13 shows why many vendors are interested in this approach. By using a common IF, many different frequency bands can be utilized for many different services.

Other router and switch vendors are capable of entering the market because the IF-to-WAN conversion is something that they can work into their product line. It also shows how Cisco can migrate the IF into WAN interfaces in a range of existing and future router products.

Equally important, by making the wireless interface just one more WAN interface on a router, Cisco has integrated all the network management for the wireless system into the normal router and switch network management, such as CiscoView and Cisco Works 2000.

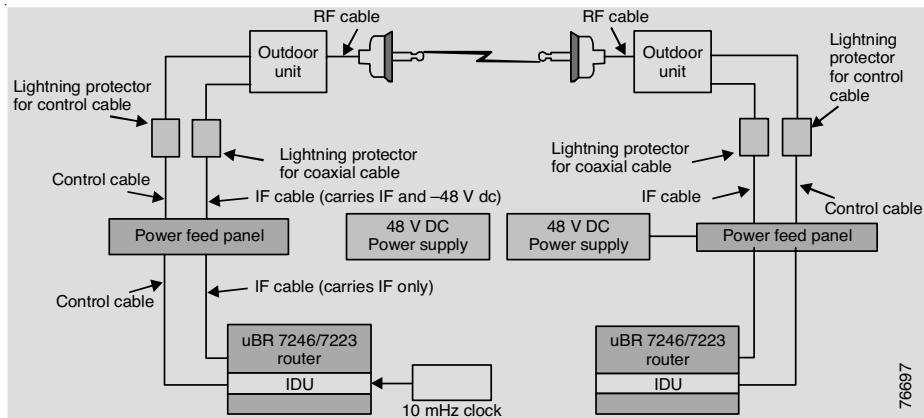


FIGURE 7.12 Hardware components of the point-to-point system

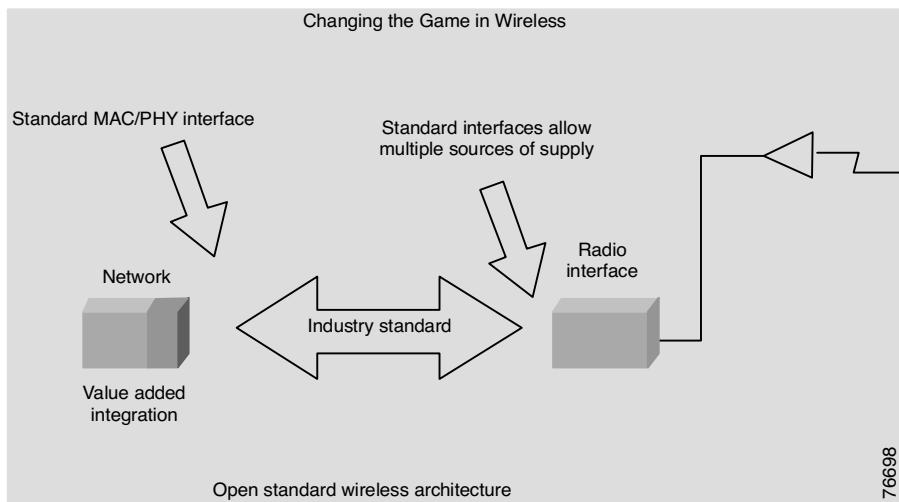


FIGURE 7.13 Open standard wireless architecture

7.14 IP VECTOR ORTHOGONAL FREQUENCY-DIVISION MULTIPLEXING

The system uses the next generation microwave technology invented by Cisco Systems, called Vector Orthogonal Frequency-Division Multiplexing (VOFDM), to resolve the issue of multipath signals in a radiating environment.

A transmitted signal gets reflected off buildings, vegetation, bodies of water, and large, solid surfaces, causing ghosts of the carrier (main or intentional) frequency to arrive at the receiver later than the carrier frequency.

Multipath signal issues are a liability for all radio systems except those without VOFDM or a feature to cancel or filter the late-arriving signals.

Embedded in the OFDM-modulated carrier frequencies are training tones. These allow multipath-channel compensation on a burst-by-burst basis. This is especially important on the uplink because each OFDM burst may be transmitted by a different Subscriber Unit (SU) over a different multipath channel. The overall effect of the VOFDM scheme is an RF system that is extremely resilient to multipath signals.

Multiple Access and Error Control Schemes

This section describes various multiple access techniques and error control schemes.

7.14.1 Channel Data Rate

Raw channel over-the-air data rates are 36, 24, 18, 12, 9, and 6 Mbps. Excluding physical layer overhead, the user rates are different for the downstream and upstream links. Downstream, these rates are 22.4, 17, 12.8, 10.1, 7.6, and 5.1 Mbps. Upstream, the rates are 19.3, 15.2, 11.4, 8.1, 6.2, 4.4, 4.2, 3.2, and 1.4 Mbps. Various combinations of these rates are supported, depending on the cell type.

These are configuration parameters that can be set and changed. Thus, if a customer requests an increased data rate service, the change can be made from the Network Operations Center (NOC) without personnel having to visit the customer site.

The service provider can make this as simple or as complicated as desired. Thus, one service city can have all subscribers on a single data rate plan, while another city can offer time-of-day and day-of-week data rate premium services.

7.14.2 Downstream and Upstream User Bandwidth Allocation

Downstream and upstream user bandwidth can be assigned dynamically for session-based traffic or, during initial registration for best-effort Internet data access, based on service flows for each user.

7.14.3 Duplexing Techniques

Time-division duplexing and frequency-division duplexing are two common techniques used for duplexing. There are challenges with each of the two approaches, related to implementation, flexibility, sensitivity, network synchronization, latency, repeaters, asymmetrical traffic, AGC, and the number of SAW filters. The TDD vs. FDD selection is not a simple decision to make—indeed, the advantages are relatively close to one another. However, having the benefit of the most recent information regarding the procurement of low-cost duplexors, development priority was given to FDD scheme.

TDD is a duplexing technique that utilizes time sharing to transmit and receive data in both directions. Each side is allotted a certain amount of time to transmit, generally in symmetric amounts. The TDD algorithms are embedded into each of the RF processor boards and are synched by protocol instruction when the units are first powered up. Commonly, these synching protocols are updated on a routine basis.

In FDD, the total allocated spectrum of frequency is divided so that each end of the radio link can transmit in parallel with the other side. FDD is commonly divided equally, but it is not symmetric on many links.

7.15 MULTIPLE ACCESS TECHNIQUE

User bandwidth allocation is carried out by means of a Medium Access Control (MAC) protocol. This protocol is based on the MAC portion of DOCSIS protocol developed by the Cable Labs consortium. The MAC protocol assigns service flows (SID) to each user; depending on the Quality of Service (QoS) requirements, the upstream MAC scheduler provides grants to fulfill the bandwidth needs. Similarly, the downstream bandwidth is divided between active users of unicast and multicast services.

Each upstream channel is divided into intervals. Intervals are made up of one or more minislots. A *minislot* is the smallest unit of granularity for upstream transmit opportunities.

Upstream transmit opportunities are defined by the MAC MAP message. This message is sent from the base station to all registered CPEs. In the MAP message, each interval is assigned a usage code that defines the type of traffic that can be transmitted during that interval, as well as whether the interval is open for contention by multiple CPEs or for the sole use of one CPE. The interval types are request, request/data, initial maintenance, station maintenance, short data grant, long data grant, and acknowledgment.

For example, the request interval is used for CPE bandwidth requests and is typically multicast to all CPEs; therefore, it is an interval open for contention. Multiple CPEs can attempt to send a bandwidth request; if the request is granted, the base station will assign a series of minislots to the CPE in the next MAP message. Contention is resolved via a truncated binary exponential back-off algorithm.

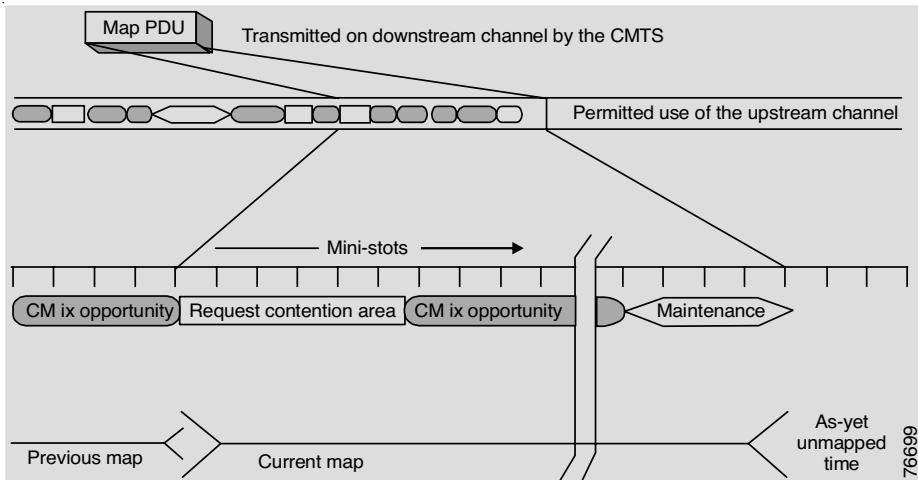


FIGURE 7.14 Frame Allocation MAP

To support customer QoS requirements, six types of service flows are specified: Unsolicited Grant Service (UGS), real-time Polling Service (rtPS), Unsolicited Grant Service with Activity Detection (UGS-AD), non-real-time Polling Service (nrtPS), Best Effort (BE) Service, and a Committed Information Rate (CIR) Service.

7.15.1 Unsolicited Grant Service

The intent of UGS is to reserve fixed-size data grants at periodic transmission opportunities for specific real-time traffic flows. The MAC scheduler provides fixed-size data grants at periodic intervals to the service flow. The QoS parameter for the given service flow sets the grant size, the nominal grant interval, and the tolerant grant jitter.

7.15.2 Real-Time Polling Service

The intent of the rtPS grants is to reserve upstream transmission for real-time traffic flows such as VoIP. Such service flows receive periodic transmission opportunities regardless of network congestion, but they release their

transmission reservation when they are inactive. As such, the base station MAC scheduler sends periodic polls to rtPS service flows using unicast request opportunities enabling subscriber wireless port to request for the upstream bandwidth that it needs. The QoS parameter for such a service flow is the nominal polling interval and the jitter tolerance for the request/grant policy.

7.15.3 Unsolicited Grant Service with Activation Detection

The intent of UGS-AD is to reserve upstream transmission opportunities for real-time traffic flows such as VoIP with silence suppression. USG-AD is designed to emulate the capabilities of UGS service when active, and rtPS service when inactive. The QoS parameter for such a service flow is the nominal polling interval, tolerated polling jitter, nominal grant interval, tolerated grant jitter, unsolicited grant size, and the request/transmission policy.

7.15.4 Non-Real-Time Polling Service

The intent of nrtPS is to set aside upstream transmission opportunities for non-real-time traffic flows, such as FTP transfer. These service flows receive a portion of transmission opportunities during traffic congestion. The base station MAC scheduler typically polls such service flows either in a periodic or a nonperiodic fashion. The subscriber wireless port can use either the unicast request opportunities or contention request opportunities to request upstream grants. The QoS parameter for such a service flow is the nominal polling interval, reserved minimum traffic rate, maximum allowed traffic rate, traffic priority, and request/transmission policy.

7.15.5 Best Effort Service

The intent of the BE service is to provide an efficient way of transmission for the best-effort traffic. As such, the subscriber wireless port can use either contention or unicast request opportunities to request upstream grants. The QoS parameter for such a service flow is the reserved minimum traffic rate, maximum sustained traffic rate, and traffic priority.

7.15.6 Committed Information Rate

CIR can be implemented in several different ways. As an example, it could be a BE service with a reserved minimum traffic rate or nrtPS with a reserved minimum traffic rate.

7.15.7 Frame and Slot Format

The *frame and slot format* is based on the MAC protocol. The downstream transmission is broadcast, similar to Ethernet, with no association with framing or a minislot. The recipients of the downstream packets perform packet filtering based on the Layer 2 address—the Ethernet MAC or SID value.

The upstream transmission is based on a Time-Division Multiple Access (TDMA) scheme, and the unit of time is a minislot. The minislot size varies based on upstream configuration settings and can carry data between 8 bytes and approximately 230 bytes. TDMA synchronization is done by time-stamp messages, and the time of transmission is communicated to each subscriber by the MAP messages (a MAP message carries the schedule information (map) for each minislot of the next data Protocol Data Unit (PDU)). MAP messages are initiated by the MAC scheduler in the base station and thus convey how each minislot is used (reserved for user traffic, for initial invitation, or as contention slots). The contention slots are used for best-effort traffic to request bandwidth from the scheduler. The frame time is programmable and can be optimized for a given network.

7.16 SYNCHRONIZATION TECHNIQUE (FRAME AND SLOT)

Accurately receiving an Orthogonal Frequency-Division Multiplexing (OFDM) burst requires burst timing and frequency offset estimation. *Burst timing* means determining where in time the OFDM burst begins and ends. Determining the difference between the local demodulating oscillator and the modulating oscillator at the transmitter is called *frequency synchronization* or *frequency offset estimation*.

Burst timing and frequency offset are determined simultaneously through the use of the extra-cyclic-prefix samples in every downstream OFDM burst. Regardless of the channel, the samples in the extra-cyclic-prefix are equal to a set of time-domain samples in the OFDM burst. This structure is optimally exploited to simultaneously identify the correct OFDM burst timing and frequency offset. These estimates are then filtered over successive OFDM bursts.

Burst timing and frequency synchronization are required for the downstream link. The upstream link will be frequency locked once the subscriber unit frequency locks the downstream signal. That is, the local oscillator at the subscriber unit is synchronized to the base station oscillator by use of a

frequency-locked loop. In this way, the upstream transmission that arrives at the base station receiver has nearly zero frequency offset.

As in any TDMA upstream, each subscriber unit must lock to the TDMA slot when transmitting on the upstream. This is referred to as *ranging* each subscriber. As prescribed by the MAC layer, a subscriber fills periodic ranging slots with a known sequence. This known sequence is used at the base station to determine proper slot timing.

The upstream TDMA synchronization is done via time-stamp messages that carry the exact timing of the base station clock. Each subscriber phase locks its local clock to the base station clock and then synchronizes its minislot (unit of TDMA) counter to match that of the base station. Furthermore, ranging is performed as part of initial acquisition to advance the transmission time of the subscriber so that the precise arrival of its transmission is synchronized with the expected minislot time of the base station.

7.17 AVERAGE OVERALL DELAY OVER LINK

The average overall delay in the link depends on the particular bandwidth and spectral efficiency setting. In all cases, the physical layer delay is limited to 5 ms in the downstream. The upstream physical layer delay is less than 2 ms.

7.18 POWER CONTROL

Power control is done in real time to track the rapidly changing environment. It is one facet of the system capability to adjust on a packet-by-packet basis.

The power control is capable of adjusting for fades as deep as 20 dB.

The receiver signal power at the subscriber is controlled through the use of an Automatic Gain Control (AGC) system. The AGC system measures received power at the analog-to-digital converter (ADC) and adaptively adjusts the analog attenuation. With AGC, channel gain variations on the order of 200 Hz can be accommodated without impacting the OFDM signal processing.

Receive power at the base station is regulated through the use of an Automatic Level Control (ALC) system. This system measures power levels of each subscriber and creates transmit attenuator adjustments at the subscriber so that future communication is done with the correct transmit power level. This power control loop is implemented in the physical layer hardware so as not to impact the MAC-layer performance.

Power measurements for the purposes of ALC occur on upstream traffic. Specific power bursts may be requested by the base station MAC to power control subscribers.

7.19 ADMISSION CONTROL

A new user is admitted to the system by means of a software suite. Components of this suite are the User Registrar, Network Registrar, Modem Registrar, and Access Registrar.

The User Registrar enables wireless network subscribers to self-provision via a web interface. Subscriber self-provisioning includes account registration and activation of the subscriber's CPE and personal computers over the wireless access network. The User Registrar activates subscriber devices with account-appropriate privileges through updates to an LDAPv3 directory.

The Network Registrar supplies DHCP and DNS services for the CPEs and personal computers. The Network Registrar DHCP allocates IP addresses and configuration parameters to clients based on per-device policies, which are obtained from an LDAP directory. The Network Registrar allocates limited IP addresses and default configuration parameters to inactivated devices, to steer users to the User Registrar's activation page.

The Modem Registrar adds TFTP and time services to the Network Registrar for the CPEs. The Modem Registrar TFTP builds DOCSIS configuration files for clients based on per-CPE policies, which are obtained from an LDAP directory. The Modem Registrar builds limited-privilege configuration files to inactivated CPEs.

The Access Registrar supplies RADIUS services for the CPEs and the clients that are connected to the CPEs. The Access Registrar RADIUS returns the configuration parameters to NAS clients based on per-subscriber policies, which are obtained from an LDAP directory. The Access Registrar returns limited-privilege NAS and PPP parameters to unregistered subscribers and to inactivated CPEs.

7.20 REQUIREMENTS FOR THE CELL RADIUS

As stated previously, the system employs Frequency-Division Duplexing (FDD). Unlike Time-Division Multiplexing (TDD) systems, the equipment is not subject to cell radius limits. The only limitation to the cell radius is that

imposed by Free Space Attenuation (FSA) of radio signals. Timing limitations imposed by the MAC protocol on upstream and downstream channels permit the cell radius to be beyond the radio horizon (FSA notwithstanding).

Unless the subscriber density is very low, as in rural areas, very large cells are not recommended because they may quickly become capacity-limited and are difficult to scale to meet the higher capacity demand.

7.20.1 Requirement for Frequency Reuse

Two architectures are described in this section. The first is a multicellular architecture, called the *small-cell pattern*. The second architecture is known as a *single-cell architecture*. The baseline small-cell architecture employs a 4×3 frequency reuse pattern. This means that a four-cell reuse pattern and three sectors are employed within each cell or Base Station (BS). Within a cellularized network, a hexagonal cell is tiled out to completely cover the service area. In our system, a four-cell tiling pattern is repeated over the service area. This tiling pattern puts a lower limit on the reuse distance, thereby controlling interference levels. Employing three sectors within each BS further reduces interference compared to omnidirectional antennae. A 4×3 reuse pattern works well for operations using obstructed (OBS) links. It uses a total of 24 channels: 12 channels downstream, plus 12 channels upstream.

Note that the number of MMDS channels used may be different than $12 + 12$ because the system supports channels narrower than 6 MHz. At any given BS, signals are received from Subscriber Units (SUs) over both OBS links and Line-of-Sight (LOS) links. In general, OBS links are attenuated at 40 dB/decade ; that is, R^4 path loss. LOS links, however, propagate based on R^{-2} path loss. Thus, if an SU has a LOS link to its BS, it likely also has a LOS link to a reuse BS. To suppress cochannel interference from LOS links, horizontal and vertical polarization are alternatively used on reuse cells. This results in an overall reuse scheme of $4 \times 3 \times 2$ and (still) using a total of 24 channels. This reuse pattern can be continued to cover as large a geographical area as desired. Improved frequency reuse can occur when networks are of more limited extent. This is because there are fewer tiers of interfering cells. In these cases, greater C/I ratios are obtained, along with a greater network capacity (given a fixed amount of spectrum).

Our approach to frequency reuse for OBS links is conservative in that it does not rely on polarization discrimination. Signals transmitted over multipath channels often experience depolarization. If a dual polarization reuse scheme were employed, it may be very difficult to achieve 99.99 percent link availability because as a reuse (undesired) channel is depolarized, it results in higher levels of interference to the cochannel (desired) signal.

7.20.2 Radio Resource Management

Radio resource management falls under the following broad categories:

- Spectrum management in a cell
- Load balancing of CPEs within an upstream channel
- Time-slotted upstream architecture
- Contention resolution
- Traffic policing
- Traffic shaping
- Quality-of-service controls

7.20.3 Spectrum Management in a Cell

Within a cell, the upstream frequency band can be split into as many as four channels. The channel bandwidths that can be configured are 1.5 MHz, 3 MHz, and 6 MHz. The downstream frequency consists of one channel that again can be configured as 1.5, 3, or 6 MHz. The capability of the Cisco wireless products to operate in these wide ranges of channel bandwidths allows the operator great flexibility in designing the network for efficient usage.

7.20.4 Load Balancing of CPEs Within an Upstream Channel

A cell within the Cisco wireless access architecture consists of a downstream channel with up to four upstream 1.5 MHz channels. By default, load balancing is performed on the upstream channels as CPEs enter the network. Special algorithms run on CPE and the head end to ensure a uniformity of CPE loading on the upstream channels. This allocation of CPEs across the upstream channels can also be done under user control.

7.20.5 Time-Slotted Upstream

Cisco's wireless access solution uses a MAC protocol based on DOCSIS. This protocol is based on a broadcast downstream architecture and a time-slotted upstream architecture. The time slots for the upstream govern the access rights of each CPE on to the upstream channel. There is a very sophisticated scheduler that runs on the head end and allocates these time slots to all the CPEs. Resource sharing of the upstream bandwidth is realized by each CPE making its request to send in a contention time slot, the head end responding

to it in a subsequent message downstream (called a MAP message), and the CPE using the information in the MAP message to send the data in an appropriate time slot upstream.

7.21 CONTENTION RESOLUTION

The DOCSIS protocol uses the notion of contention time slots in the upstream. These time slots are used by the CPEs to send a request to the head end for a time slot grant to send data. In a loaded upstream network, it is possible for these contention slots to become very congested themselves and nonproductive. Cisco's wireless head end uses an intelligent algorithm to ensure that the contention slots are evenly spaced, especially in times of high upstream load. The CPEs also implement an intelligent algorithm that ensures that the request grants from the CPEs are spaced evenly over all the contention slots in a given MAP message.

7.21.1 Traffic Policing

One of the most important features of the wireless access channel is its shared nature. At any given time, several hundreds or thousands of CPE subscribers may be sharing an upstream or downstream channel. Although this shared nature is useful for reducing the per-subscriber investment for the operator, two important aspects must be considered while providing this access service:

- The need to allocate this bandwidth fairly among all the users.
- The need to prevent misbehaving users from completely monopolizing the access.

The Cisco solution uses sophisticated algorithms at the head end to police the traffic from each subscriber CPE.

The traffic from each CPE can also be shaped by algorithms running at the head end. This allows the operator to provision services to the CPE based on the critical nature of the traffic, customer needs, and so on. The peak rates of traffic from each subscriber are measured on a continuous basis and are policed at every request from the CPE. If the request from the CPE exceeds its allotted rate, the grant is delayed, thereby effectively controlling the rate of data transfer from the CPE. Differentiated services can be offered to subscribers. The operator can configure different maximum data rates for different wireless subscribers and can charge accordingly. Subscribers requiring

higher peak rates (and who are willing to pay for them) can be configured with a higher peak rate limit dynamically or statically.

7.22 INTERFACE SPECIFICATIONS BASED ON THE GENERIC REFERENCE MODEL

We now discuss the generic reference model in Figure 7.15. In this figure, reference points I–VIII refer to specific interfaces and/or functions.

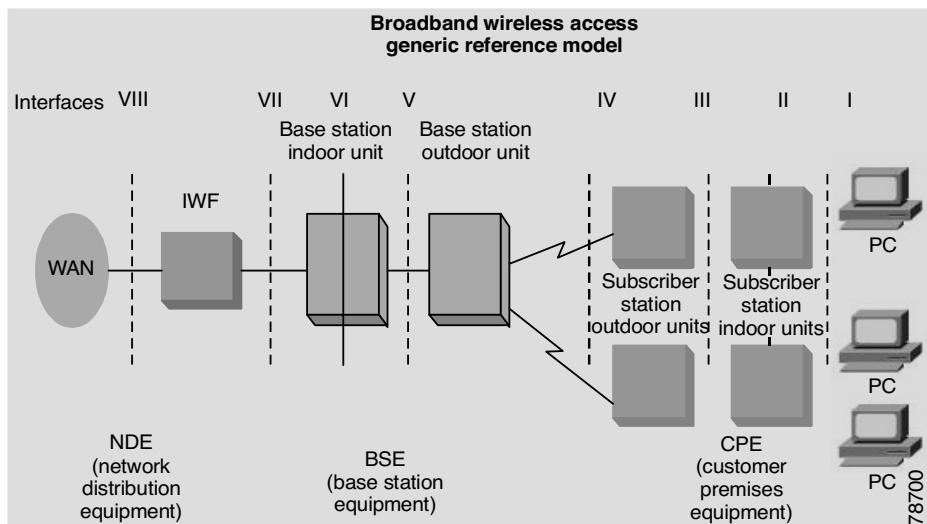


FIGURE 7.15 Generic reference model for broadband wireless access

- Interface I: Wireless Subscriber Interface
- Interface II: Subscriber Indoor Unit PHY/MAC Interface
- Interface III: Subscriber Radio IF/RF Interface—This interface is a physical coaxial cable carrying IF, digital control information, and DC powering for the ODU.
- Interface IV: RF Air Interface—This interface is the over-the-air RF interface. It is in the MMDS frequency band (2.500 to 2.690 GHz, and 2.150 to 2.162 GHz). The energy radiated from the antenna is governed by FCC Rules and Regulations, Part 21.

The above descriptions of interfaces I to III represent Cisco Systems products by way of example. Such products interwork with the base station described next, via the interface IV compatibility, but will have many new and different CPE interfaces, features, and services.

- *Interface V: Base Station RF/IF Interface*—This interface at the base station is a physical coaxial cable carrying the Intermediate Frequency (IF), digital control information, and DC powering for the Outdoor Unit (ODU).
- *Interface VI: Base Station Indoor PHY/MAC Interface*—This interface is internal to the Cisco router.
- *Interface VII: Network Connection Interface*

7.23 WIRELESS PROTOCOL STACK

The access wireless architecture consists of a base station system that serves a community of subscriber systems. It is a point-to-multipoint architecture in the sense that the entire bandwidth on the upstream and downstream is shared among all the subscribers. The protocol stack implemented to make all this work is based on the DOCSIS standards developed by the Cable Labs consortium.

The current state of the art is the version by Cisco that includes a base station end (a UBR7200 router); the subscriber end is in the 3600 or a 900 router series. The base station end and the subscriber end operate as forwarding agents and also as end systems (hosts). As forwarding agents, these systems can also operate in bridging or routing mode. The principal function of the wireless system is to transmit Internet Protocol (IP) packets transparently between the base station and the subscriber location. Certain management functions also ride on IP that include, for example, spectrum management functions and the software downloading.

Both the subscriber end and the base station end of the wireless link are IP hosts on a network, and they fully support standard IP and Logical Link Control (LLC) protocols, as defined by the IEEE 802 LAN/MAN Standards Committee standards. The IP and Address Resolution Protocol (ARP) protocols are supported over DIX and SNAP link layer framing. The minimum link

layer Minimum Transmission Unit (MTU) on transmit from the base station is 64 bytes; there is no such limit for the subscriber end. IEEE 802.2 support for TEST and XID messages is provided.

The primary function of the wireless system is to forward packets. As such, data forwarding through the base station consists of transparent bridging or network layer forwarding such as routing and IP switching. Data forwarding through the subscriber system is link layer transparent bridging as with Layer 3 routing based on IP. Forwarding rules are similar to [ISO/IEC10038], with modifications as described in DOCSIS specifications Section 3.1.2.2 and Section 3.1.2.3. Both the base station end and the subscriber end support DOCSIS-modified spanning-tree protocols and include the capability to filter 802.1d bridge PDUs (BPDUs). The DOCSIS specification also assumes that the subscriber units will not be connected in a configuration that would create network loops.

Both the base station end and the subscriber end provide full support for Internet Group Management Protocol (IGMP) multicasting.

Above the network layer, the subscribers or end users can use the transparent IP capability as a bearer for higher-layer services. Use of these services is transparent to the subscriber end and the base station end.

In addition to the transport of user data, several network management and operation capabilities are supported by the base station end and the subscriber end:

- Simple Network Management Protocol (SNMP), [RFC-1157], for network management
- Trivial File Transfer Protocol (TFTP), [RFC-1350], a file transfer protocol, for downloading software and configuration information, as modified by RFC 2349, “TFTP Timeout Interval and Transfer Size Options” [RFC-2349]
- Dynamic Host Configuration Protocol (DHCP), [RFC-2131], a framework for passing configuration information to hosts on a TCP/IP network
- Time of Day Protocol [RFC-868], to obtain the time of day

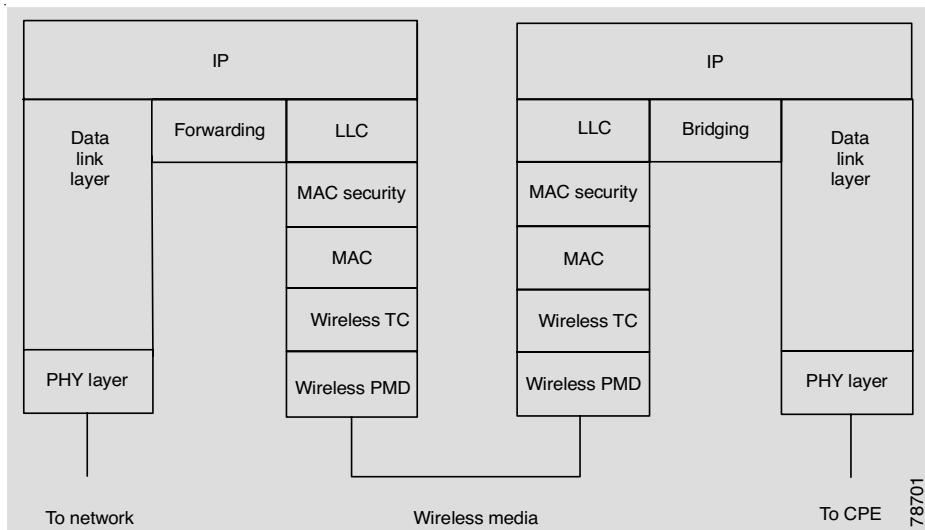


FIGURE 7.16 Wireless protocol stack for network management and operation

Link layer security is provided in accordance with DOCSIS baseline privacy specifications.

7.24 SYSTEM PERFORMANCE METRICS

Table 7.4 shows the network capacity for a high capacity suburban/urban small-cell network.

Table 7.4 Network Capacity for High-capacity Suburban/Urban Small-cell Network.

Cell radius (Mile)	No. of cells	Small businesses served	Small business penetration	Households served	Household penetration
2.2	83	5,229	10.5%	126,990	15.9%

These capacities are based on 6 MHz downstream channels and 3 MHz upstream channels:

- **Downstream**—6 MHz; at the medium VOFDM-throughput setting, 17.0 Mbps.

- **Upstream**—3.0 MHz; at the low VOFDM-throughput setting, 4.4 Mbps. The low setting is so that both SB and residential SUs may be serviced by the same upstream channels.

This high-capacity network has 83 cells in the network, each with 3 sectors. A total of 249 sectors are in the network, each serving 21 SB and 510 HH. Overall, a total of 249 ¥ 21, or 5,229, SBs are served by the network; similarly, a total of 126,990 HHs are served.

This network design graphically illustrates the power and scalability of Cisco's technology.

7.25 SUPERCELL NETWORK DESIGN

The supercell (very large cell) network design is one that provides low coverage and low overall network capacity. However, it may be attractive for initial network rollout because of the availability of existing (tall) towers. In our supercell design, assuming that a sufficient number of MMDS channels are available, up to 18 sectors may be used. No frequency reuse is performed within the supercell, again because of sector-to-sector isolation requirements that are greater than sector antennae can provide. Each sector operates independently. Also, at least four MMDS-channels must be set aside as guard bands.

The number of sectors deployed on the supercell may be scaled as the demand for capacity grows. Because there is no frequency reuse, no special requirements are placed on the design of the sector antennae. For example, the same panel antenna used for a 3-sector supercell could also be used all the way up to 18 sectors. However, to increase RF coverage, narrower-beam antennae may be employed to increase EIRP. This will be effective as long as the supercell isn't capacity-limited (which is often the case).

The capacity of the supercell is given in Table 7.5. In this deployment model, we have not differentiated between suburban and urban deployments. The assumption is that the desire is to provide service primarily to subscribers in which LOS operation is possible. Because macro-diversity is not possible in a supercell design, coverage becomes difficult. For example, the COST-231 Hata model predicts 80 percent coverage at a radius of only 15 miles—much smaller than the desired cell radius. Moreover, this coverage is computed at the limits of the model's antenna heights—200 m for the HE, and 10 m for the SU over suburban terrain.

Table 7.5 Network Capacity for a Supercell.

Number of sectors	Cell radius (Mile)	Small businesses served	Small business penetration	Households served	Household penetration
3	16	1,116	2%	26,856	3%
18	23	2,232	4%	53,712	7%

These capacities are based on 6 MHz downstream channels and 3 MHz upstream channels, both at the medium VOFDM-throughput setting.

If a lower availability objective were desired, the fade margin could be greatly reduced, thereby extending the cell radius. More importantly, the sector-to-sector isolation would be greatly reduced, perhaps admitting frequency reuse within the supercell. Because the cell is capacity-limited (there are many more subscribers in the cell's radio footprint than there is capacity to service), this would be a tremendous benefit.

The multipath channel from both the front (desired) antenna and the rear (undesired) antenna must be the same so that the fading from the desired and undesired antennae must be highly correlated.

The time rate-of-change of the multipath channel must be slow enough such that power control errors are very small.

The following sections present both the general functions performed by the various configuration items or building blocks segmented into transport and services products.

7.26 TRANSPORT LAYER PRODUCTS

The transport layer is composed of the equipment that provides the transmission and reception of IF signals between the rooftop and router equipment and the RF signals over the air. The transport equipment is designed to work in an outdoor environment mounted on buildings or telecommunications towers. The P2MP transport layer is physically segmented into the hub and terminal equipment categories, as depicted in Figure 7.17.

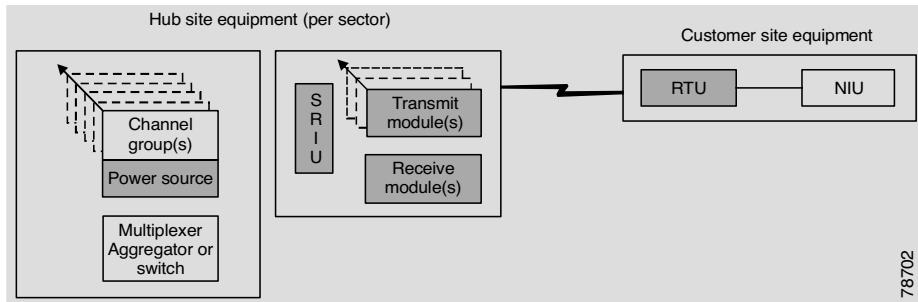


FIGURE 7.17 Hub site equipment (per sector)

7.26.1 P2MP Transport Equipment Element—Customer Premises

The terminal equipment consists of an integrated RF transceiver/antenna, commonly referred to as the *rooftop unit* (RTU). This equipment is easily installed on any customer rooftop using a standard mounting device. The RTU requires two RG-11 coaxial cables to the indoor equipment for transmit, receive, and power. The RTU operates on 12.5 VDC (nominal) at the input and in standard configuration must be installed within 60 m of the Network Interface Unit (NIU), although longer spans can be engineered and supported.

7.26.2 Rooftop Unit

The sole element of the P2MP transport layer at the terminal site is the RTU. The RTU is an integrated antenna and RF transceiver unit that provides wireless transmission and reception capabilities in the 5.7 GHz frequency region. Received and transmitted signals are frequency translated between the 5.7 GHz region and an intermediate frequency (IF) in the 400 MHz range to the Network Interface Unit (NIU).

The RTU consists of an antenna(s), a down-converter/IF strip, and an up-converter/transmitter. It receives/transmits using orthogonal polarization. Selection of polarization (horizontal/vertical) occurs at installation and is dictated by the hub-sector transmitter/receiver. This selection remains fixed for the duration that service is provided to that site.

The RTU mounts on the exterior of a subscriber's building. Some alignment is required to gain Line of Sight (LOS) to the hub serving the RTU. Multiple RTUs can be deployed to provide path redundancy to alternate hub sites. The RTU requires dual coax cable (RG-11) runs to the NIU for signal

and power. The maximum standard separation between the RTU and the NIU is 60 m. This separation can be extended via application-specific designs.

7.26.3 Basic Receiver

A single basic receiver is required per 90° sector, if no return-path redundancy is required. The receive module is an integrated 5.7 GHz receiver/down-converter/antenna. A collection of signals is received from customer units operating in the 5.7 GHz band and is block down-converted to an intermediate frequency signal. This signal is provided to any of the channel group types. Vertical or horizontal polarization is selectable, and a redundant receiver per sector can be deployed as an option.

7.26.4 High-Gain Receiver

A high-gain receiver is used in lieu of the basic receiver when higher link margin is required because of the specific geographic conditions of deployment. The high-gain receive module is intended to be matched only with the high-gain transmit module. The specifications are identical to those of the basic receive module, except for physical package and antenna gain.

Because of the modularity of the SP2200 products, there is no one standard rack or set of racks. All SP2200 elements are designed to mount in a standard 19-inch (48.3 cm) open relay rack with a standard EIA hole pattern or an equipment enclosure with 19 inches (48.3 cm) of horizontal equipment mounting space. The final assembly of the equipment into racks is accomplished on site at initial install or over time, as capacity demands.

7.27 LMDS ENVIRONMENTAL CONSIDERATIONS

Environmental conditions such as rain and smog must be considered when deploying RF systems that transmit at frequencies above 10 GHz because these conditions degrade the signal path and shorten the maximum range for a given data link.

LMDS data links are generally about one-fourth that of MMDS or U-NIII links and require fairly strict adherence to a line-of-sight implementation. One of the more favorable aspects of the LMDS frequency, however, is that it has exceptional frequency reuse capabilities.

The data link availability is expressed in terms of the number of nines that follow the decimal point. For example, 99.999 percent link availability

means that a data link will be up and online (available) for all but .001 percent of the year. Link availability is dependent on a wide range of items, but these generally begin with fundamental RF system design issues such as antenna size, range between antennae, and atmospheric conditions (for LMDS band).

7.28 WLAN STANDARDS COMPARISON

Table 7.6 provides a brief comparison of WLAN standards.

Table 7.6 A Brief Comparison of HomeRF, Bluetooth, and 802.11 WLAN Standards.

	HomeRF	Bluetooth	802.11
Physical Layer	FHSS ¹	FHSS	FHSS, DSSS ² , IR ³
Hop Frequency	50 hops per second	1600 hops per second	2.5 hops per second
Transmitting Power	100 mW	100 mW	1 W
Data Rates	1 or 2 Mbps	1 Mbps	11 Mbps
Max # Devices	Up to 127	Up to 26	Up to 26
Security	Blowfish format	0-, 40-, and 64-bit	40- to 128-bit RC4
Range	150 feet	30 to 300 feet	400 feet indoors, 1000 feet LOS
Current Version	V1.0	V1.0	V1.0

¹FHSS—frequency hopping spread spectrum

²DSSS—direct sequence spread spectrum

³IR—infrared

The following should be noted in Table 7.6:

- 40 to 128-bit RC4 refers to very robust data security algorithms.
- An 802.11 range of 1,000 feet refers to outdoor conditions. Indoor conditions are more difficult for these types of RF systems.
- 802.11 power output of 1 W is substantial.

- The maximum number of devices supported depends on data rate per device.
- The Aironet acquisition uses 802.11.

Although there are three standards in use in the United States, and an additional two are in use in Europe (HyperLAN and HyperLAN2), the FCC thinks highly of the 802.11b standard, and a close relationship exists between the FCC and the IEEE, which backs the standard.

EXERCISES

1. What is the primary difference between narrowband and broadband wireless?
2. What is the primary difference between WLAN and fixed wireless?
3. Identify three different applications for wireless.
4. What are the fundamental hardware elements of a wireless solution?
5. What are the primary benefits of wireless?

CHAPTER 8

VIRTUAL PRIVATE NETWORKS

Chapter Goals

- Comparison of IPSec to Cisco Encryption Technology
- IPSec protocols
- IKE protocol
- Internet Engineering Task Force (IETF)
- NAT-Traversal
- Gateway-to-gateway architecture
- Virtual Private Networking (VPN)
- Encryption process
- End-to-end encryption
- ESP version 3

8.1 SECURITY POLICY

IPSec is a framework of open standards developed by the Internet Engineering Task Force (IETF). IPSec provides security for the transmission of sensitive information over unprotected networks (such as the Internet). IPSec acts at the network layer, protecting and authenticating IP packets between participating IPSec devices (“peers”), such as Cisco routers. IPSec provides a range of network security services. These services are optional. In general, local security policy will dictate the use of one or more of these services:

- **Data Confidentiality:** The IPSec sender can encrypt packets before transmitting them across a network.

- **Data Integrity:** The IPSec receiver can authenticate packets sent by the IPSec sender to ensure that the data has not been altered during transmission.
- **Data Origin Authentication:** The IPSec receiver can authenticate the source of the IPSec packets sent. This service is dependent upon the data integrity service.
- **Anti-Replay:** The IPSec receiver can detect and reject replayed packets.

With IPSec, data can be transmitted across a public network without fear of observation, modification, or spoofing. This enables applications such as virtual Private Networks (VPNs), including intranets, extranets, and remote user access. IPSec services are similar to those provided by Cisco Encryption Technology (CET), a proprietary security solution introduced in Cisco IOS. However, IPSec provides a more robust security solution and is standards-based. IPSec also provides data authentication and anti-replay services in addition to data confidentiality services, while CET provides only data confidentiality services.

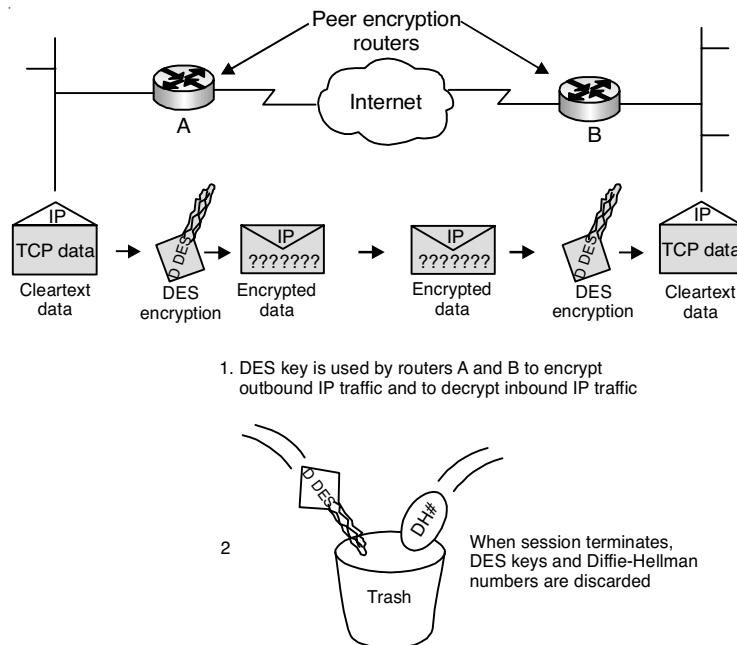


FIGURE 8.1 Cisco encryption technology

IPSec shares the same benefits as Cisco Encryption Technology: both technologies protect sensitive data that travels across unprotected networks,

and, like Cisco Encryption Technology, IPSec security services are provided at the network layer, so you do not have to configure individual workstations, PCs, or applications. This benefit can provide a great cost savings. Instead of providing the security services you do not need to deploy and coordinate security on a per-application, per-computer basis, you can simply change the network infrastructure to provide the needed security services. IPSec also provides additional benefits not present in Cisco Encryption Technology. Because IPSec is standards-based, Cisco devices are able to interoperate with other IPSec-compliant networking devices to provide the IPSec security services. IPSec-compliant devices could include both Cisco devices and non-Cisco devices, such as PCs, servers, and other computing systems. Cisco and its partners, including Microsoft, are planning to offer IPSec across a wide range of platforms, including Cisco IOS software, the Cisco PIX Firewall, Windows 95, and Windows NT. Cisco is working closely with the IETF to ensure that IPSec is quickly standardized.

- A mobile user will be able to establish a secure connection back to his office. For example, the user can establish an IPSec “tunnel” with a corporate firewall—requesting authentication services—in order to gain access to the corporate network; all of the traffic between the user and the firewall will then be authenticated. The user can then establish an additional IPSec tunnel—requesting data privacy services—with an internal router or end system.
- IPSec provides support for the Internet Key Exchange (IKE) protocol and for digital certificates. IKE provides negotiation services and key derivation services for IPSec. Digital certificates allow devices to be automatically authenticated to each other without the manual key exchanges required by Cisco Encryption Technology. (For more information, see the “Internet Key Exchange Security Protocol” feature documentation.) This support allows IPSec solutions to scale better than Cisco Encryption Technology solutions, making IPSec preferable in many cases for use with medium-sized, large-sized, and growing networks, where secure connections between many devices is required. These and other differences between IPSec and Cisco Encryption Technology are described in the following sections.

8.2 IPSEC NETWORK SECURITY

Should you implement Cisco Encryption Technology (CET) or IPSec network security in your network? The answer depends on your requirements. If

you require only Cisco-router-to-Cisco-router encryption, then you could run Cisco Encryption Technology, which is a more mature, higher-speed solution. If you require a standards-based solution that provides multivendor interoperability or remote client connections, then you should implement IPSec. If you want to implement data authentication with or without privacy (encryption), then IPSec is the right choice. If you want, you can configure both Cisco Encryption Technology and IPSec simultaneously in your network, even simultaneously, on the same device. A Cisco device can simultaneously have Cisco Encryption Technology secure sessions and IPSec secure sessions, with multiple peers. IPSec is an extension of the IP protocol, which provides security to the IP and the upper-layer protocols. It was first developed for the new IPv6 standard and then backported to IPv4. The IPSec architecture is described in the RFC2401. The following few paragraphs provide a short introduction into IPSec.

IPSec uses two different protocols—AH and ESP—to ensure the authentication, integrity and confidentiality of the communication. It can protect either the entire IP datagram or only the upper-layer protocols. The appropriate modes are called tunnel mode and transport mode. In *tunnel mode*, the IP datagram is fully encapsulated by a new IP datagram using the IPSec protocol. In *transport mode*, only the payload of the IP datagram is handled by the IPSec protocol inserting the IPSec header between the IP header and the upper-layer protocol header.

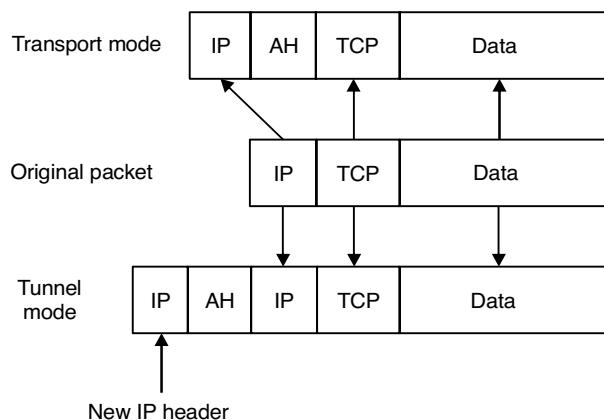


FIGURE 8.2 IPSec tunnel and transport mode

To protect the integrity of the IP datagram, the IPSec protocols use Hash Message Authentication Codes (HMAC). 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. To protect the confidentiality of the IP datagram, the IPSec protocols use standard symmetric encryption algorithms. The IPSec standard requires the implementation of NULL and DES. Today, stronger algorithms are used, like 3DES, AES, and Blowfish. To protect against denial-of-service attacks the IPSec protocols use a sliding window. Each packet gets assigned a sequence number and is only accepted if the packet's number is within the window or newer. Older packets are immediately discarded. This protects against replay attacks, where the attacker records the original packets and replays them later. For the peers to be able to encapsulate and encapsulate the IPsec packets they need a way to store the secret keys, algorithms and IP addresses involved in the communication. All these parameters needed for the protection of the IP diagram are stored in a Security Association (SA). The security associations are in turn stored in a Security Association Database (SAD). Each security association defines the following parameters:

- Source and destination IP address of the resulting IPSec header: These are the IP addresses of the IPsec peers protecting the packets.
- IPSec protocol (AH or ESP): sometimes compression (IPCOMP) is supported, too.
- The algorithm and secret key used by the IPsec protocol.
- Security Parameter Index (SPI): This is a 32-bit number that identifies the security association. Some implementations of the security association database allow further parameters to be stored:
 - IPsec mode (tunnel or transport)
 - Size of the sliding window to protect against replay attacks
- Lifetime of the security association. Because the security association defines the source and destination IP addresses, it can only protect one direction of the traffic in a full duplex IPSec communication. To protect both directions, IPSec requires two unidirectional security associations. The security associations only specify how IPsec is supposed to protect the traffic. Additional information is needed to define which traffic to protect when. This information is stored in the Security Policy (SP), which

in turn is stored in the Security Policy Database (SPD). A security policy usually specifies the following parameters:

- Source and destination address of the packets to be protected: In transport mode these are the same addresses as in the SA. In tunnel mode, they may differ.
- The protocol (and port) to protect: Some IPsec implementations do not allow the definition of specific protocols to protect. In this case, all traffic between the mentioned IP addresses is protected.
- The security association to use for the protection of the packets: The manual set-up of the security association is error prone and not very secure. The secret keys and encryption algorithms must be shared between all peers in the virtual private network. The exchange of the keys poses critical problems for the system administrator: How can secret symmetric keys be exchanged when no encryption is yet in place? To solve this problem, the Internet key exchange protocol (IKE) was developed. This protocol authenticates the peers in the first phase. In the second phase the security associations are negotiated and the secret symmetric keys are chosen using a Diffie-Hellmann key exchange. The IKE protocol then even takes care of periodically rekeying the secret keys to ensure their confidentiality.

8.3 IPSEC PROTOCOLS

The IPsec protocol family consists of two protocols: Authentication Header (AH) and Encapsulated Security Payload (ESP). Both are independent IP protocols. AH is the IP protocol 51 and ESP is the IP protocol 50. The following two sections briefly discuss their properties.

8.3.1 Authentication Header (AH)

The AH protocol protects the integrity of the IP datagram. To achieve this, the AH protocol calculates a HMAC to protect the integrity. When calculating the HMAC, the AH protocol bases it on the secret key, the payload of the packet, and the immutable parts of the IP header like the IP addresses. It then adds the AH header to the packet. The AH header is shown in Figure 8.3.

Next header	Payload length	Reserved
Security Parameter Index (SPI)		
Sequence number (Replay defense)		
Hash message authentication code		

FIGURE 8.3 The AH header protects the integrity of the packet

The AH header is 24 bytes. The first byte is the *Next Header* field. This field specifies the protocol of the following header. In tunnel mode, a complete IP datagram is encapsulated; therefore, the value of this field is 4. When encapsulating a TCP datagram in transport mode, the corresponding value is 6. The next byte specifies the length of the payload. This field is followed by two reserved bytes. The next double word specifies the 32-bit *Security Parameter Index* (SPI). The SPI specifies the security association to use for the encapsulation of the packet. The 32-bit *Sequence Number* protects against replay attacks. Finally, 96 bits hold the *Hash Message Authentication Code* (HMAC). This HMAC protects the integrity of the packets since only the peers knowing the secret key can create and check the HMAC. Since the AH protocol protects the IP datagram, including the immutable parts of the IP header like the IP addresses, the AH protocol does not allow NAT. Network Address Translation (NAT) replaces an IP address in the IP header (usually the source IP) with a different IP address. After the exchange, the HMAC is not valid anymore. The NAT Traversal extension of the IPsec protocol implements ways around this restriction.

8.3.2 Encapsulated Security Payload (ESP)

The ESP protocol can both ensure the integrity of the packet using a HMAC and the confidentiality using encryption. After encrypting the packet and calculating the HMAC, the ESP header is generated and added to the packet. The ESP header consists of two parts and is shown in Figure 8.4.

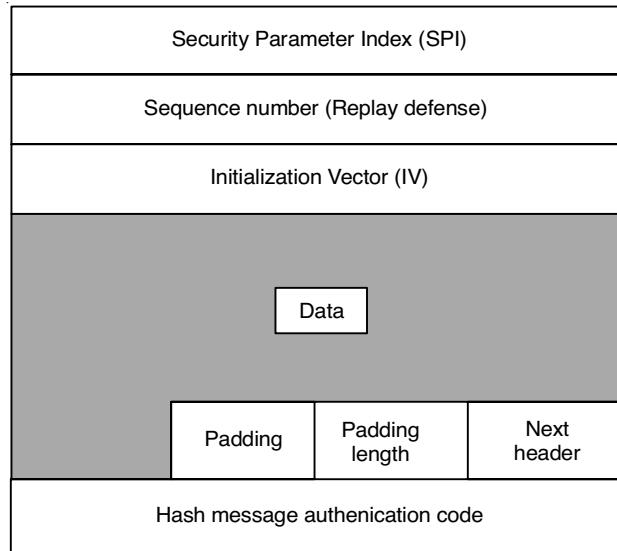


FIGURE 8.4 The ESP header

The first double word in the ESP header specifies the *Security Parameter Index* (SPI). This SPI specifies the SA to use for the encapsulation of the ESP packet. The second double word holds the *Sequence Number*. This sequence number is used to protect against replay attacks. The third double word specifies the *Initialization Vector* (IV), which is used in the encryption process. Symmetric encryption algorithms are susceptible to a frequency attack if no IV is used. The IV ensures that two identical payloads lead to different encrypted payloads. IPsec uses block ciphers for the encryption process. Therefore, the payload may need to be padded if the length of the payload is not a multiple of the block length. The length of the pad is then added. Following the pad length, the 2-byte long *Next Header* field specifies the next header. Lastly, the 96-bit long HMAC is added to the ESP header, ensuring the integrity of the packet. This HMAC only takes the payload of the packet into account. The IP header is not included in the calculation process. The usage of NAT therefore does not break the ESP protocol. In most cases, NAT is not possible in combination with IPsec. The NAT-Traversal offers a solution in this case by encapsulating the ESP packets within UDP packets.

8.3.3 IKE Protocol

The IKE protocol solves the most prominent problem in the set-up of secure communication: the authentication of the peers and the exchange of the

symmetric keys. It then creates the security associations and populates the SAD. The IKE protocol usually requires a user space daemon and is not implemented in the operating system. The IKE protocol uses 500/UDP for its communication. The IKE protocol functions in two phases. The first phase establishes the *Internet Security Association Key Management Security Association* (ISAKMP SA). In the second phase, the ISAKMP SA is used to negotiate and set up the IPsec SAs. The authentication of the peers in the first phase can usually be based on Pre-Shared Keys (PSK), RSA keys and X.509 certificates (racoon even supports Kerberos). The first phase usually supports two different modes: main mode and aggressive mode. Both modes authenticate the peer and set up an ISAKMP SA, but the aggressive mode uses only half the number of messages to achieve this goal. This does have its drawbacks though, because the aggressive mode does not support identity protection and is therefore susceptible to a man-in-the-middle attack if used in conjunction with pre-shared keys.

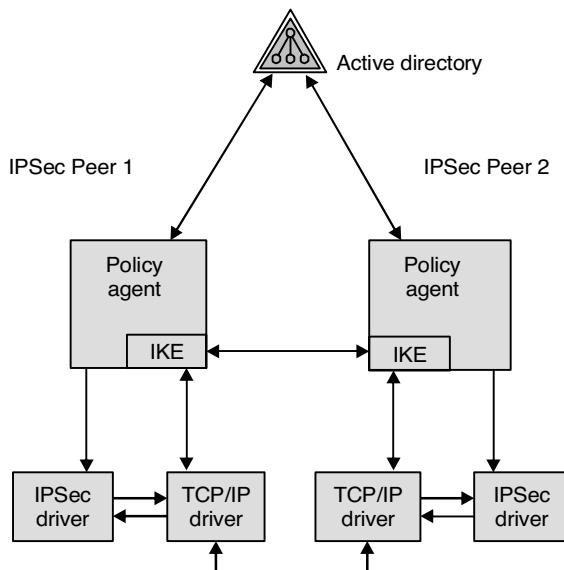


FIGURE 8.5 IKE protocol exchanges security

This is the only purpose of the aggressive mode. Because of the internal workings of the main mode, it does not support the usage of different pre-shared keys with unknown peers. The aggressive mode does not support identity protection and transfers the identity of the client in the clear. The peers therefore know each other before the authentication takes place and

different pre-shared keys can be used for different peers. In the second phase, the IKE protocol exchanges security association proposals and negotiates the security associations based on the ISAKMP SA. The ISAKMP SA provides the authentication to protect against a man-in-the-middle attack. This second phase uses the quick mode. Usually two peers negotiate only one ISAKMP SA, which is then used to negotiate several (at least two) unidirectional IPSec SAs.

8.4 NAT-TRAVERSAL

What is NAT-Traversal and why is it needed? Often, one peer in the VPN is behind a NAT-device. (We just assume Source-NAT devices are used in this example.) Here, NAT refers to Source-NAT or Masquerading. What does this mean concerning the VPN? The original IP address of the peer is hidden by the NAT-device. The NAT-device conceals the original source IP address and replaces it with its own IP address. This makes the IPSec AH protocol immediately unusable. But ESP can still be used if both sides are configured correctly. So why do you need NAT-Traversal? Because as soon as two machines behind the same NAT device try to build a tunnel to the outside, both will fail. Why is this happening?

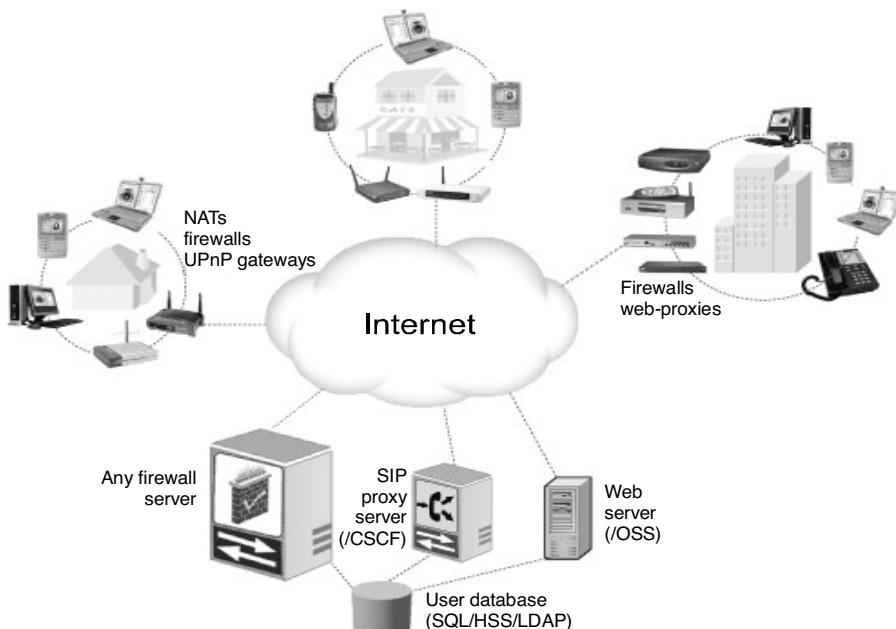


FIGURE 8.6 NAT-traversal again encapsulates the ESP packets in UDP packets

The NAT device needs to keep track of the connections to be able to the replay packets back to the original client. Therefore, the NAT device maintains an internal table where all “natted” connections are stored. Let's assume one client connects to a web server on the Internet. The NAT device conceals the original address by replacing it with its own address. It then makes a note in its internal table that all packets coming back on the chosen client port have to be sent to the original client. As soon as the second client starts a connection, it handles that connection in the same way. If the second client chose the same client port by coincidence, the NAT device will also modify the client port for clarity. This works very well using TCP and UDP because those protocols provide ports. ESP does not use ports. Therefore, the NAT device can only use the protocol to distinguish the packets. When the first client connects, it stores the information in the table that all ESP packets have to be “de-natted” to the first client. When the second client connects, it will overwrite this entry with the appropriate entry for the second one, thus breaking at least the first connection. What does NAT traversal do to help? NAT-traversal again encapsulates the ESP packets in UDP packets. These can easily be handled by a NAT device since they provide ports. By default, port 4500/UDP is used. NAT traversal is specified in several drafts. There are no RFCs at the moment. A nice feature of NAT traversal is the fact that, once activated, the peers automatically use it when needed.

8.5 VIRTUAL PRIVATE NETWORK (VPN)

The most common use of IPsec implementations is providing Virtual Private Network (VPN) services. A VPN is a virtual network, built on top of existing physical networks, that can provide a secure communications mechanism for data and IP information transmitted between networks. Because a VPN can be used over existing networks, such as the Internet, it can facilitate the secure transfer of sensitive data across public networks. This is often less expensive than alternatives, such as dedicated private telecommunications lines between organizations or branch offices. VPNs can also provide flexible solutions, such as securing communications between remote telecommuters and the organization servers, regardless of where the telecommuters are located. A VPN can even be established within a single network to protect particularly sensitive communications from other parties on the same network. Discuss these three models: gateway-to-gateway, host-to-gateway, and host-to-host.

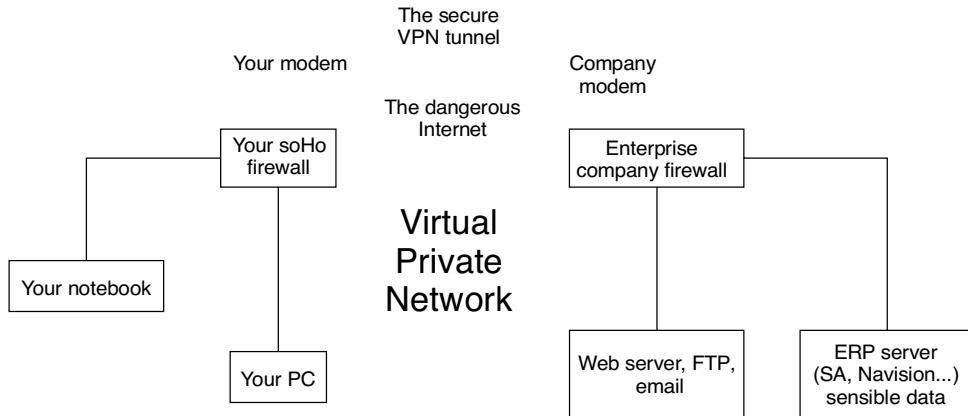


FIGURE 8.7 VPN Communications mechanism

VPNs can use both symmetric and asymmetric forms of cryptography. Symmetric cryptography uses the same key for both encryption and decryption, while asymmetric cryptography uses separate keys for encryption and decryption, or to digitally sign and verify a signature. Symmetric cryptography is generally more efficient and requires less processing power than asymmetric cryptography, which is why it is typically used to encrypt the bulk of the data being sent over a VPN. One problem with symmetric cryptography is with the key exchange process; keys must be exchanged out-of-band to ensure confidentiality. Common algorithms that implement symmetric cryptography include the Digital Encryption Standard (DES), Triple DES (3DES), Advanced Encryption Standard (AES), Blowfish, RC4, International Data Encryption Algorithm (IDEA), the Hash Message Authentication Code (HMAC) versions of Message Digest 5 (MD5), and Secure Hash Algorithm (SHA-1).

Asymmetric cryptography (also known as *public key cryptography*) uses two separate keys to exchange data. One key is used to encrypt or digitally sign the data, and the other key is used to decrypt the data or verify the digital signature. These keys are often referred to as public/private key combinations. If an individual public key (which can be shared with others) is used to encrypt data, then only that same individual private key (which is known only to the individual) can be used to decrypt the data. If an individual private key is used to digitally sign data, then only that same individual public key can be used to verify the digital signature. Common algorithms that implement asymmetric cryptography include RSA, Digital Signature Algorithm (DSA), and Elliptic Curve DSA (ECDSA). Although there are numerous

ways in which IPsec can be implemented, most implementations use both symmetric and asymmetric cryptography. Asymmetric cryptography is used to authenticate the identities of both parties, while symmetric encryption is used for protecting the actual data because of its relative efficiency. It is important to understand that VPNs do not remove all risk from networking. While VPNs can greatly reduce risk, particularly for communications that occur over public networks, they cannot eliminate all risk for such communications. One potential problem is the strength of the implementation. For example, flaws in an encryption algorithm or the software implementing the algorithm could allow attackers to decrypt intercepted traffic; random number generators that do not produce sufficiently random values could provide additional attack possibilities. Another issue is encryption key disclosure; an attacker who discovers a key can not only decrypt traffic, but potentially also pose as a legitimate user. Another area of risk involves availability. A common model for information assurance is based on the concepts of confidentiality, integrity, and availability. Although VPNs are designed to support confidentiality and integrity, they generally do not improve availability, the ability for authorized users to access systems as needed. In fact, many VPN implementations actually tend to decrease availability somewhat because they add more components and services to the existing network infrastructure. This is highly dependent upon the chosen VPN architecture model and the details of the implementation. The following sections describe each of the three primary VPN architectures: host-to-host, host-to-gateway, and gateway-to-gateway.

8.6 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 8.8 shows an example of an IPSec network architecture that uses the gateway-to-gateway model to provide a protected connection between the two networks.

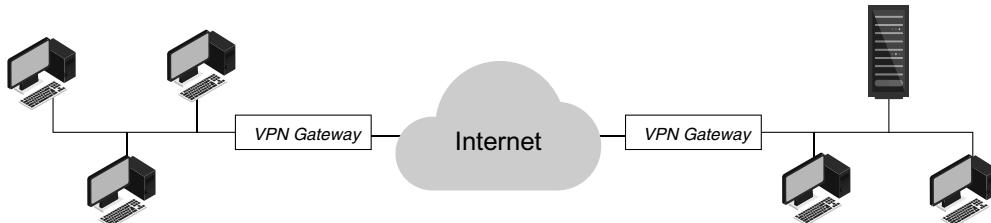


FIGURE 8.8 Gateway-to-gateway architecture example

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, and 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 8.8 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 the 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. 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.

8.7 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 8.9 shows an example of an IPSec host-to-gateway architecture that provides a protected connection for the remote user.

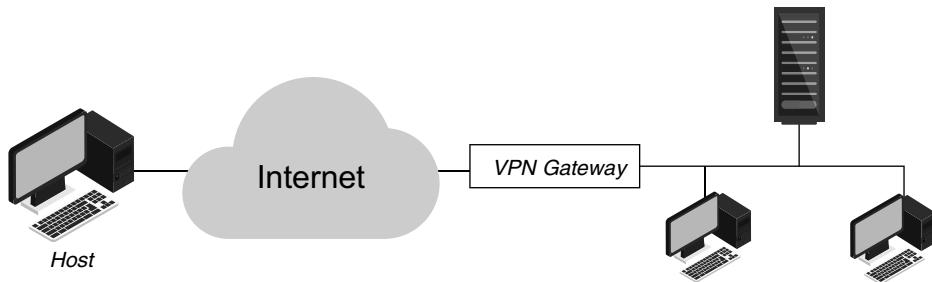
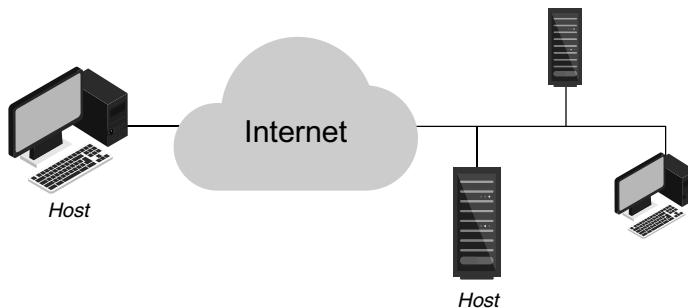


FIGURE 8.9 Host-to-gateway architecture

In this model, IPSec connections are created as needed for each individual VPN user. 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, host, and the VPN gateway is 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. 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 traveling 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 and hosts need to have VPN client software configured.

**FIGURE 8.10** Host-to-host architecture

In this model, IPSec connections are created as needed for each individual VPN user. The users and hosts are configured to act as IPSec clients with the IPSec server. When a user wishes to use resources on the IPSec server, the user 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 users, host, and the server is protected by the IPsec connection. As shown in Figure 8.10 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 utilized 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. All user systems and servers that participate in VPNs need to have VPN software installed and/or configured.

8.8 MODEL COMPARISON

Table 8.1 Comparison of 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)

(continued)

Feature	Gateway-to-gateway	Host-to-gateway	Host-to-host
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

8.9 TCP/IP NETWORK SECURITY PROTOCOL

We now move on to discuss the TCP/IP model and its layers: application, transport, network, and data link. We also explain how the security controls at each layer provide different types of protection for TCP/IP communications. IPSec, a network layer security control, can provide several types of protection for data, depending on its configuration. Most IPSec implementations provide VPN services to protect communications between networks. This section describes VPNs and highlights the three primary VPN architecture models.

TCP/IP is widely used throughout the world to provide network communications. The TCP/IP model is composed of the following four layers, each having its own security controls that provide different types of protection.

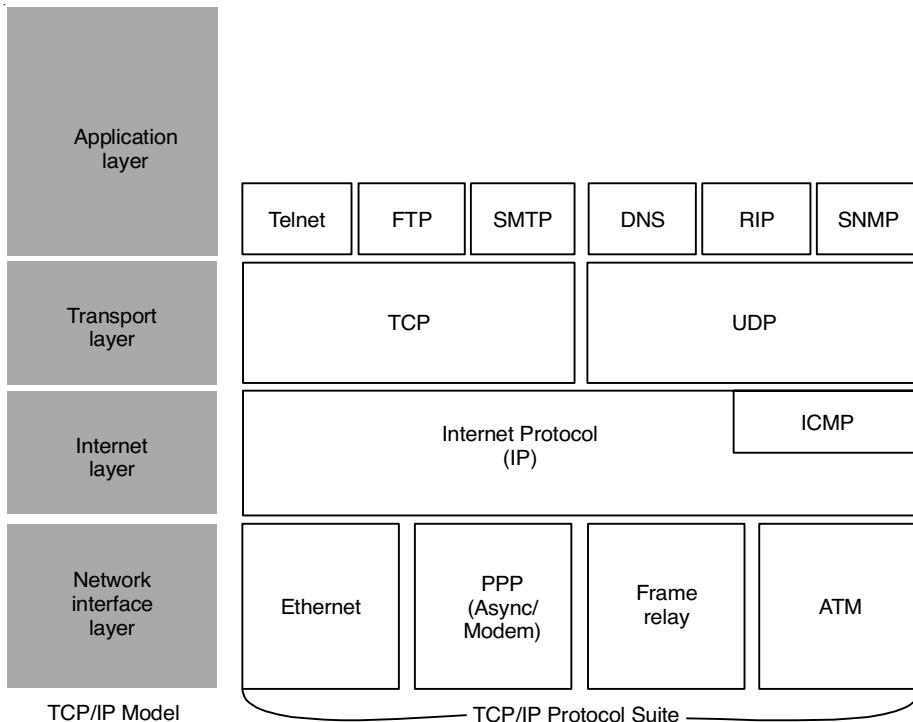


FIGURE 8.11 The four layers of the TCP/IP model

- The **application layer**, which sends and receives data for particular applications. Separate controls must be established for each application; this provides a very high degree of control and flexibility over each application's security, but it may be very resource-intensive. Creating new application layer security controls is also more likely to create vulnerabilities. Another potential issue is that some applications may not be capable of providing such protection or being modified to do so.
- The **transport layer**, which provides connection-oriented or connectionless services for transporting application layer services across networks. Controls at this layer can protect the data in a single communications session between two hosts. The most frequently used transport layer control is TLS/SSL, which most often secures HTTP traffic. To be used, the transport layer controls must be supported by both the clients and servers.
- The **network layer**, which routes packets across networks. Controls at this layer apply to all applications and are not application-specific, so applications do not have to be modified to use the controls. However,

this provides less control and flexibility for protecting specific applications than transport and application layer controls. Network layer controls can protect both the data within packets and the IP information for each packet.

- The **data link layer**, which handles communications on the physical network components. The data link layer's controls are suitable for protecting a specific physical link, such as a dedicated circuit between two buildings or a dial-up modem connection to an ISP. Because each physical link must be secured separately, data link layer controls generally are not feasible for protecting connections that involve several links, such as connections across the Internet.
- IPsec is a framework of open standards for ensuring private communications over IP networks which has become the most commonly used network layer security control. It can provide several types of protection, including maintaining confidentiality and integrity, authenticating the origin of data, preventing packet replay and traffic analysis, and providing access protection.
- A VPN is a virtual network built on top of existing networks that can provide a secure communications mechanism for data and IP information transmitted between networks. VPNs generally rely on both symmetric and asymmetric cryptography algorithms. Asymmetric cryptography is used to provide peer authentication; symmetric encryption is used to protect the actual data transfers because of its relative efficiency.
- Although VPNs can reduce the risks of networking, they cannot eliminate it. For example, a VPN implementation may have flaws in algorithms or software that attackers can exploit. Also, VPN implementations often have at least a slightly negative impact on availability, because they add components and services to existing network infrastructures.

There are three primary models for VPN architectures:

- **Gateway-to-gateway:** This connects two networks by deploying a gateway to each network and establishing a VPN connection between the two gateways. Communications between hosts on the two networks are then passed through the VPN connection, which provides protection for them. No protection is provided between each host and its local gateway. The gateway-to-gateway is most often used when connecting two secured networks, such as a branch office and headquarters, over the Internet. This often replaces more costly private WAN circuits. Gateway-to-gateway

VPNs are typically transparent to users and do not involve installing or configuring any software on clients or servers.

- **Host-to-gateway:** This connects hosts on various networks with hosts on the organization's network by deploying a gateway to the organization's network and permitting external hosts to establish individual VPN connections to that gateway. Communications are protected between the hosts and the gateway, but not between the gateway and the destination hosts within the organization. The host-to-gateway model is most often used when connecting hosts on unsecured networks to resources on secured networks, such as linking traveling employees to headquarters over the Internet. Host-to-gateway VPNs are typically not transparent to users because each user must authenticate before using the VPN and each host must have VPN client software installed and configured.
- **Host-to-host:** This connects hosts to a single target host by deploying VPN software to each host and configuring the target host to receive VPN connections from the other hosts. This is the only VPN model that provides protection for data throughout its transit. It is most often used when a small number of users need to use or administer a remote system that requires the use of insecure protocols and can be updated to provide VPN services. The host-to-host model is resource-intensive to implement and maintain because it requires configuration on each host involved, including the target.

8.10 NODE-TO-NODE ENCRYPTION

Node-to-node encryption is also referred to as link-to-link encryption. In the OSI model, the data link layer is concerned with node-to-node or link-to-link connections. As a result, if you encrypt the packet at the data link layer, it must be decrypted by the data link layer recipient before passing it up to the network layer to determine how to forward the packet. When encrypting at the data link layer, a packet has to be decrypted and re-encrypted for each node-to-node hop along the route. Node-to-node encryption operating at the data link layer requires compatible devices, sharing a protocol, and a key management process for every device on the network. If the devices on the network are not compatible, they will not be able to relay the packets they receive. This is an issue that must be considered, because if the network is large, the management requirements will be significant.

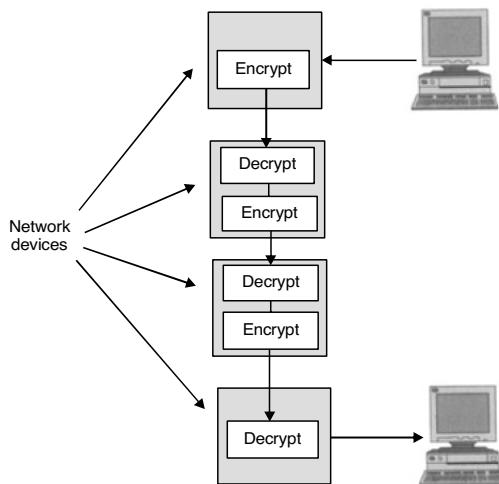


FIGURE 8.12 Node-to-node encryption

8.11 SITE-TO-SITE ENCRYPTION

As an alternative, end-to-end encryption operates at the upper layers of the OSI models and can encapsulate data into standard network protocols. As a result, no special considerations are necessary for the intermediate hops along the network. The encryption and decryption of the encapsulated data is done at either end of the connection.

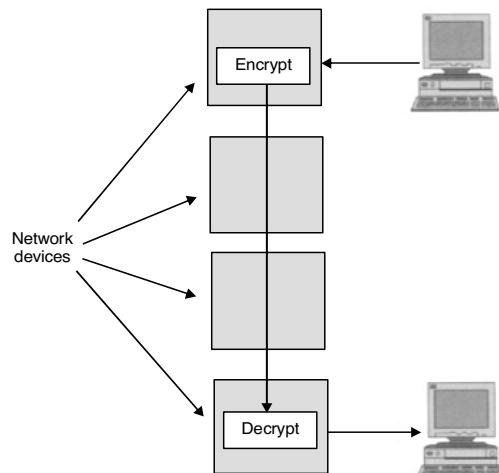


FIGURE 8.13 End-to-end encryption

However, a consideration with end-to-end encryption is that the further up the protocol stack you move the encryption, the more information you may be providing a potential eavesdropper. As you will see, as you move the encryption higher up the protocol stack, more information is revealed about the sender, the recipient, and the nature of the data.

8.12 WHERE TO ENCRYPT

The level of security achieved differs depending on where the encryption takes place. The level of security required should dictate where your encryption is performed. In the OSI model, if you encrypt at the network layer (layer 3), information identifying the devices or machines can be intercepted. For instance, information on the IP addresses of the source and destination can be monitored. This information can be used for network traffic analysis. Network traffic analysis in itself can provide a wealth of information that can be utilized by individuals or entities sniffing the network. If the encryption takes place further up the protocol stack at the transport layer (layer 4), then someone eavesdropping on the communications can tell which port you are communicating with on the recipient system. From that information, eavesdroppers can surmise what protocol you are using.

For example, if you are communicating with port 161, then you are most likely using SNMP for network management. If you are communicating with port 25, then you are probably using SMTP for e-mail. Knowing the protocols that are running on a device or system can be used to plan an attack. A TCP/IP port is a logical connection to a server that usually handles a specific service or protocol. TCP/IP network servers often provide a variety of services or protocols such as SNMP, HTTP, or SMTP. Each of the available services “listens” for an outside connection on a particular port number or uses a specified port number. Table 8.2 lists examples of well-known port numbers and their associated services.

Table 8.2 Well-known Port Numbers and their Associated Services.

Port	Service
25	Simple Mail Transport Protocol (SMTP)
53	DNS
80.81	HTTP
161	SNMP

The port numbers range from 1 to 65,535, with the privileged ports ending at 1,024. Non-privileged ports range from 1,025 to 65,535. Sometimes the port numbers are displayed at the end of a URL. Take, for example, <http://www.someurl.com:81>. In this example, the server is using port 81 for the particular URL address. It indicates the port number that the TCP/IP connection is using on the Web server. At the application layer (layer 7), even more information is available. If email is encrypted and transmitted at this level, it may be secure from disclosure and modification, but anyone monitoring the transmission will know you sent email, to whom you sent it, and where. As a result, when implementing encryption on a network, you have to determine where you need the encryption to take place and what is an adequate level of security based upon the sensitivity of the data.

8.13 ENCRYPTION PROCESS

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).

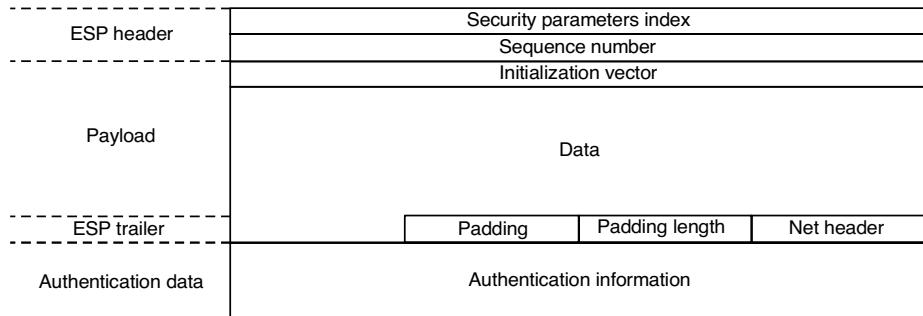
8.14 ESP PACKET FIELDS

ESP adds a header and a trailer around each packet's payload. Each ESP header is composed of two fields.

- 1. 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, ESP), to determine which SA is being used.

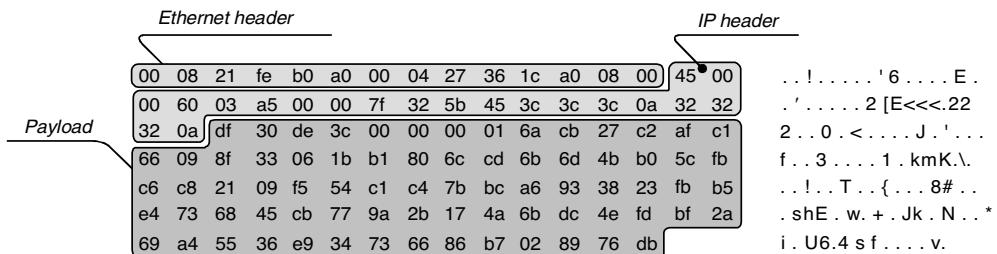
2. **Sequence Number.** Each packet is assigned a sequential sequence number, and only packets within a sliding window of sequence numbers are accepted. This provides protection against replay attacks because duplicate packets use the same sequence number. This also helps to thwart denial-of-service attacks because old packets that are replayed will have sequence numbers outside the window, and will be dropped immediately without performing any more processing.
3. 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.
4. The third part of the packet is the ESP trailer, which contains at least two fields and may optionally include one more.
5. **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.
6. **Padding Length.** This number indicates how many bytes long the padding is. The Padding Length field is mandatory.
7. **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. Like AH, the field contains the MAC output. Unlike AH, the MAC in ESP does not include the outermost IP header in its calculations. The recipient of the packet can recalculate the MAC to confirm that the portions of the packet other than the outermost IP header have not been altered in transit.

**FIGURE 8.14** ESP packet fields

8.15 HOW ESP WORKS

Reviewing and analyzing actual ESP packets can provide a better understanding of how ESP works, particularly when compared with AH packets. Figure 8.15 shows the bytes that compose an actual ESP packet and their ASCII representations in the same format. The alphabetic sequence that was visible in the AH-protected payload cannot be seen in the ESP-protected payload because it has been encrypted. The ESP packet only contains five sections: the Ethernet header, IP header, ESP header, encrypted data (payload and ESP trailer), and (optionally) authentication information. From the encrypted data, it is not possible to determine if this packet was generated in transport mode or tunnel mode. However, because the IP header is unencrypted, the IP protocol field in the header does reveal which protocol the payload uses (in this case, ESP) the unencrypted fields in both modes (tunnel and transport) are the same.

**FIGURE 8.15** ESP packet capture

Although it is difficult to tell from Figure 8.15, the ESP header fields are not encrypted. Figure 8.16 shows the ESP header fields from the first four packets in an ESP session between hosts A and B. The SPI and Sequence

Number fields work the same way in ESP that they do in AH. Each host uses a different static SPI value for its packets, which corresponds to an ESP connection being composed of two one-way connections, each with its own SPI. Both hosts initially set the sequence number to 1, and both incremented the number to 2 for their second packets.

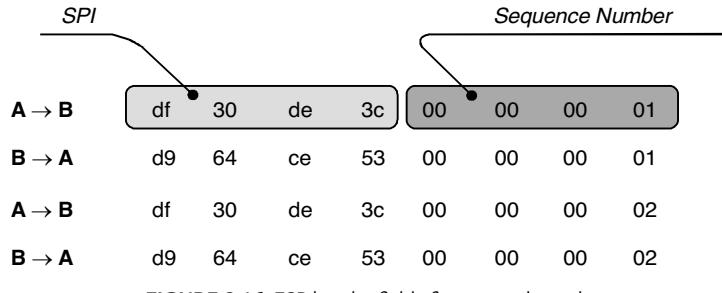


FIGURE 8.16 ESP header fields from sample packets

8.16 ESP VERSION 3

A new standard for ESP, version 3, is currently in development. Based on the current standard draft, there should be several major functional differences between version 2 and version 3, including the following:

1. The standard for ESP version 2 required ESP implementations to support using ESP encryption only (without integrity protection). The proposed ESP version 3 standard makes support for this optional.
2. ESP can use an optional longer sequence number, just like the proposed AH version 3 standard.
3. ESP version 3 supports the use of combined mode algorithms (*e.g.*, AES Counter with CBC-MAC [AES-CCM]). Rather than using separate algorithms for encryption and integrity protection, a combined mode algorithm provides both encryption and integrity protection.

The version 3 standard draft also points to another standard draft that lists encryption and integrity protection cryptographic algorithm requirements for ESP. For encryption algorithms, the draft mandates support for the null encryption algorithm and 3DES-CBC, strongly recommends support for AES-CBC (with 128-bit keys), recommends support for AES-CTR, and discourages support for DES-CBC. For integrity protection algorithms, the draft mandates support for HMAC-SHA1-96 and the null authentication

algorithm, strongly recommends support for AES-XCBC-MAC-96, and also recommends support for HMAC-MD5-96. The standard draft does not recommend any combined mode algorithms.

8.17 INTERNET KEY EXCHANGE (IKE)

The purpose of the Internet Key Exchange (IKE) protocol is to negotiate, create, and manage security associations. *Security Association* (SA) is a generic term for a set of values that define the IPsec features and protections applied to a connection. SA can also be manually created, using values agreed upon in advance by both parties, but these SA cannot be updated; this method does not scale for real-life large-scale VPNs. IKE uses five different types of exchanges to create security associations, transfer status, and error information, and define new Diffie-Hellman groups. In IPsec, IKE is used to provide a secure mechanism for establishing IPsec-protected connections. The following sections describe the five types of IKE exchanges (main mode, aggressive mode, quick mode, informational, and group) and explain how they work together for IPsec. This section also briefly discusses IKE version 2 and how it differs from the original IKE protocol.

8.18 PHASE ONE EXCHANGE

The purpose of the IKE phase one exchange is for the two IPsec endpoints to successfully negotiate a secure channel through which an IPsec SA can be negotiated. The secure channel created during phase one is commonly known as an IKE SA. The purpose of the IKE SA is to provide bidirectional encryption and authentication for other IKE exchanges: the negotiations that comprise phase two, the transfer of status and error information, and the creation of additional Diffie-Hellman groups. In fact, a phase one IKE exchange must be successfully completed before any of the other IKE exchange types can be performed. An IKE SA can be established through one of two modes: main mode and aggressive mode.

8.19 MAIN MODE

Main mode negotiates the establishment of the IKE SA through three pairs of messages. In the first pair of messages, each endpoint proposes parameters to

be used for the SA. Four of the parameters are mandatory and are collectively referred to as the protection suite:

- **Encryption Algorithm:** This specifies the algorithm to be used to encrypt data. Examples of encryption algorithms are DES, 3DES, CAST, RC5, IDEA, Blowfish, and AES. It provides guidance on selecting an encryption algorithm.
- **Integrity Protection Algorithm:** This indicates which keyed hash algorithm should be used for integrity protection. HMAC-MD5 and HMAC-SHA-1 are commonly used keyed hash algorithms.
- **Authentication Method:** There are several possible methods for authenticating the two endpoints to each other, including the following:
 - **Pre-shared Keys:** Each endpoint has been given the same secret key in advance. The endpoints use the key to generate a value that is then used to create the secret keys that are used to protect the phase 1 secure channel, as well as the eventual IPSec SA. Successful completion of the phase 1 IKE negotiation constitutes proof that each peer possesses the pre-shared secret key, which serves to authenticate the peers to each other.
 - **Digital Signatures:** Each endpoint has its own digital certificate that contains a public key. The endpoint uses the corresponding private key to digitally sign data before sending it to the other endpoint, which verifies the signature using the peers public key. The digital signature algorithm choices are RSA and the Digital Signature Standard (DSS).
 - **Public Key Encryption:** Instead of using the public/private key pair for signing data, each peer encrypts data with its own private key and decrypts data with the peer's public key. The algorithm typically used for public key encryption is RSA. Public key encryption-based authentication typically relies upon the establishment of a Public Key Infrastructure (PKI) implementation and the issuance of digital certificates. This authentication method is defined in the IKE standard, but it is not commonly implemented or used.
 - **External Authentication:** Although not specified by the current IKE standard, some IPSec implementations support the use of external authentication servers and services such as Kerberos v5. In the Kerberos method, a Kerberos server maintains all of the keys for all devices within its domain. Kerberos may also be used to authenticate

the hosts; however, the identity of the endpoints will not be concealed until the third set of messages, as described later in this section. (When some authentication methods are used, such as pre-shared keys or digital signatures, the identity of the endpoints is protected during all three sets of messages.)

8.20 DIFFIE-HELLMAN (DH) GROUP

A Diffie-Hellman group is used to generate a shared secret for the endpoints in a secure manner, so that an observer of the IKE phase one exchange cannot determine the shared secret. This shared secret is then used to generate a value that is used as input to the calculations for the phase 1 and 2 secret keys. Each DH group number corresponds to a key length and an encryption generator type (exponentiation over a prime modulus [MODP] or elliptic curve over G also known as EC2N). Although groups using elliptic curves may be more efficient than prime modulus groups, elliptic curve groups are not generally used because of intellectual property concerns involving the licensing of elliptic curve cryptography algorithms.

Table 8.3 Diffie-Hellman Group Definitions.

Group number	Generator	Modulus of field size
1	MODP	768-bit modulus
2	MODP	1024-bit modulus
3	EC2N	155-bit field size
4	EC2N	185-bit field size
5	MODP	1536-bit modulus
14	MODP	2048-bit modulus
15	MODP	3072-bit modulus
16	MODP	4096-bit modulus
17	MODP	6144-bit modulus
18	MODP	8192-bit modulus

Besides negotiating the parameters of the IKE protection suite, the first pair of main mode messages also includes the exchange of cookies. The cookies are based partially on the other host's IP address and a time-based counter. This provides some protection against denial of service attacks because it occurs before the cryptographically-intensive operations in subsequent steps. Figure 8.17 shows the Ethereal interpretation of the initial message in the first pair of main mode messages. (Both messages in the pair contain the same fields, so the second message in the pair is omitted for brevity.) Besides the initial cookie value, Figure 8.17 shows many other items of interest, including which mode is being used (in this case, main mode); which encryption and hash algorithms should be used; and which authentication method should be used (in this case, pre-shared keys).

```
User Datagram Protocol, Src Port: isakmp (500), Dst Port: isakmp (500)
Internet Security Association and Key Management Protocol
    Initiator cookie: 0x04874D4D109ECCF4
    Responder cookie: 0x0000000000000000
    Next payload: Security Association (1)
    Version: 1.0
    Exchange type: Identity Protection (Main Mode) (2)
    Flags
        .... .0 = No encryption
        .... 0. = No commit
        .... 0.. = No authentication
    Message ID: 0x00000000
    Length: 84
    Security Association payload
        Next payload: NONE (0)
        Length: 56
        Domain of interpretation: IPSEC (1)
        Situation: IDENTITY (1)
        Proposal payload # 1
            Next payload: NONE (0)
            Length: 44
            Proposal number: 1
            Protocol ID: ISAKMP (1)
            SPI size: 0
            Number of transforms: 1
            Transform payload # 1
                Next payload: NONE (0)
                Length: 36
                Transform number: 1
                Transform ID: KEY_IKE (1)
                Encryption-Algorithm (1): DES-CBC (1)
                Hash-Algorithm (2): MD5 (1)
                Group-Description (4): Default 768-bit MODP group (1)
                Authentication-Method (3): PSK (1)
                Life-Type (11): Seconds (1)
                Life-Duration (12): Duration-Value (28800)
```

FIGURE 8.17 Ethereal interpretation of a first pair main mode message

The second pair of main mode messages performs a key exchange through Diffie-Hellman, using the parameters negotiated during the first step. Figure 8.17 shows Ethereal interpretation of the initial message in the second pair of main mode messages. (Both messages in the pair contain the same fields, so the second message in the pair is omitted for brevity.) Most of the packet is composed of the key exchange data, as well as a nonce. The contents of the second pair of messages vary somewhat based on authentication method. Messages involving pre-shared key or digital signature-based authentication have the same fields header, key, and nonce. Messages involving public key encryption-based authentication encrypt the nonce with the other endpoints public key and exchange IDs (also protected by public keys). When pre-shared key or digital signature-based authentication is used, IDs are not exchanged until the third pair of messages so that the keys established through the Diffie-Hellman exchange can protect the IDs.

```
User Datagram Protocol, Src Port: isakmp (500), Dst Port: isakmp (500)
Internet Security Association and Key Management Protocol
    Initiator cookie: 0x04874D4D109ECCF4
    Responder cookie: 0x38945FD052E53D60
    Next payload: Key Exchange (4)
    Version: 1.0
    Exchange type: Identity Protection (Main Mode) (2)
    Flags
        .... .0 = No encryption
        .... .0 = No commit
        .... .0.. = No authentication
    Message ID: 0x00000000
    Length: 152
    Key Exchange payload
        Next payload:Nonce (10)
        Length: 100
        Key Exchange Data
    Nonce payload
        Next payload: NONE (0)
        Length: 24
        Nonce Data
```

FIGURE 8.18 Ethereal interpretation of a second pair main mode message

In the third pair of messages, each endpoint is authenticated to the other. Again, this depends on the negotiated authentication method. If pre-shared keys are specified, authenticating hash digests are exchanged; if digital signatures are specified, they are used. Regardless of the method in use, these messages are encrypted based on the information exchanged in the second pair of messages. Figure 8.18 shows ethereal interpretation of the initial message in the third pair of main mode messages. Other than the IKE header fields, the rest of the data is shown as encrypted.

```
User Datagram Protocol, Src Port: isakmp (500), Dst Port: isakmp (500)
Internet Security Association and Key Management Protocol
    Initiator cookie: 0x04874D4D109ECCF4
    Responder cookie: 0x38945FD052E53D60
    Next payload: Identification (5)
    Version: 1.0
    Exchange type: Identity Protection (Main Mode) (2)
    Flags
        .... .1. = Encryption
        .... .0. = No commit
        .... .0.. = No authentication
    Message ID: 0x00000000
    Length: 60
    Encrypted payload (32 bytes)
```

FIGURE 8.19 Ethereal interpretation of a third pair main mode message

Any of the pairs of main mode messages might also contain a vendor ID, which is a value that indicates the vendor of the sender's IPSec software. The vendor ID can be used to identify some of the sender's characteristics and preferences.

To summarize, main mode uses three pairs of messages. Each of the three pairs of messages has a different purpose. The first pair of messages negotiates the IKE SA parameters, the second pair performs a key exchange, and the third pair authenticates the endpoints to each other.

8.21 AGGRESSIVE MODE

1. Aggressive mode offers a faster alternative to main mode. It negotiates the establishment of the IKE SA through three messages instead of three pairs of messages. The first two messages negotiate the IKE SA parameters and perform a key exchange; the second and third messages authenticate the endpoints to each other. The following provides more detail on each message.
2. In the first message, endpoint A sends all the protection suite parameters, as well as its portion of the Diffie-Hellman key exchange, a nonce, and its identity.
3. In the second message, endpoint B sends the protection suite parameters, its portion of the Diffie-Hellman key exchange, a nonce, its identity, and its authentication payload (through digital signature or hash).
4. In the third message, endpoint A sends its authentication payload.

5. Aggressive mode negotiates all the same parameters as main mode through fewer messages. Also, unlike main mode, aggressive mode can be used with pre-shared secret key authentication for hosts without fixed IP addresses. However, with the increased speed of aggressive mode comes decreased security. Since the Diffie-Hellman key exchange begins in the first packet, the two parties do not have an opportunity to negotiate the Diffie-Hellman parameters. Also, the identity information is not always hidden in aggressive mode, so an observer could determine which parties were performing the negotiation. (Aggressive mode can conceal identity information in some cases when public keys have already been exchanged.) Aggressive mode negotiations are also susceptible to pre-shared key cracking, which can allow user impersonation and man-in-the-middle attacks. Another potential issue is that while all IPSec devices must support main mode, aggressive mode support is optional. Unless there are performance issues, it is generally recommended to use main mode for the phase one exchange.

8.22 PHASE TWO EXCHANGE

The purpose of phase two is to establish an SA for the actual IPSec connection. This SA is referred to as the *IPSec SA*. Unlike IKE SAs, which are bidirectional, IPSec SAs are unidirectional. This means that an IPSec connection between two systems requires two security associations. The pair of IPSec SAs is created through a single mode, quick mode. Quick mode uses three messages to establish the SA. Remember that quick mode communications are encrypted by the method specified in the IKE SA created by phase one and the Ethereal interpretation of a quick mode message. Although certain fields are visible (*e.g.*, cookies, message ID, and flags), most of the content of the message is encrypted. The following items are the most significant contents of the encrypted portion of the quick mode messages.

1. In the first message, endpoint A sends keys and IPSec SA parameter suggestions (such as the anti-replay measure).
2. In the second message, endpoint B sends keys and IPSec SA parameter selections, plus a hash for authentication.
3. In the third message, endpoint A sends a hash for authentication.

```
User Datagram Protocol, Src Port: isakmp (500), Dst Port: isakmp (500)
Internet Security Association and Key Management Protocol
    Initiator cookie: 0x04874D4D109ECCF4
    Responder cookie: 0x38945FD052E53D60
    Next payload: Hash (8)
    Version: 1.0
    Exchange type: Quick Mode (32)
    Flags
        .... .1 = Encryption
        .... .0. = No commit
        .... .0.. = No authentication
    Message ID: 0x7DEA6802
    Length: 164
    Encrypted payload (136 bytes)
```

FIGURE 8.20 Ethereal interpretation of a quick mode message

After endpoint B validates the third message, the IPSec SA are established. All active SA are stored in a Security Association Database (SAD). The SAD includes the following information for each protected connection:

- Source IP address
- Destination IP address
- SPI
- IPSec security protocol (AH or ESP)
- Mode (transport or tunnel)
- Encryption algorithm for ESP (*e.g.*, AES-CBC)
- Integrity protection algorithm (*e.g.*, HMAC-MD5 and HMAC-SHA-1)
- Secret keys used by the selected algorithms
- Key length, if any of the selected algorithms can use multiple key sizes
- SA lifetime (described later in this section)
- Sequence number information
- Anti-replay information
- Types of traffic to which this SA should be applied (*e.g.*, specific ports and/or protocols)

An SA can be uniquely identified by the combination of three parameters: the destination IP address, the SPI, and the IPSec security protocol. When an endpoint needs to know which SA applies to a particular packet, it looks it up in the SAD using these parameters. The SA describes the security measures that IPsec should use to protect communications; however, it does not fully

describe what types of traffic should be protected, and under what circumstances. That information is stored in the Security Policy Database (SPD), which classifies traffic as requiring IPsec protection (protect), not requiring IPsec protection (bypass), or being prohibited (discard). The SPD typically contains the following information for each type of traffic that needs to be protected:

- Source and destination IP address
- IP protocol (*e.g.*, TCP or UDP)
- TCP or UDP port number (optional)
- IPsec protections to be applied
- Pointer to the SA within the SAD, if an SA has already been negotiated for a particular type of traffic

Most implementations have a GUI that enables the user to configure the SPD; the SAD entries are created as a result of IKE negotiations. In some implementations, it is not obvious how the terms used in the configuration tool match up with the SAD and SPD database entries. Both databases should be protected, and only an administrator or super user should be able to configure the SPD.

Both IKE and IPsec SA typically have a limited lifetime, which cannot be increased after the SA is created. If an SA is nearing the end of its lifetime, the endpoints must create a new SA and use it instead, through a process known as *rekeying*. The SA lifetime specifies how often each SA should be rekeyed, either based on elapsed time or the amount of network traffic. The lifetime is most often based on an elapsed time of no more than a day.

8.23 INFORMATIONAL EXCHANGE

The purpose of the IKE informational exchange is to provide the endpoints a way to send each other status and error messages. The IKE SA provides protection for the status and error information, ensuring that unauthorized messages do not disrupt an IPsec negotiation or prematurely end an existing IPsec SA. For example, one endpoint can tell another endpoint that a particular SA should no longer be used. However, the messages sent through the informational exchange are UDP-based, and the recipient does not acknowledge them, so there is no guarantee that the other endpoint will receive them.

8.24 GROUP EXCHANGE

In the pre-defined Diffie-Hellman groups, each group number specifies a modulus size and an encryption generator type. The IKE group exchange can be used to negotiate the creation of additional Diffie-Hellman groups. Once two endpoints agree on the characteristics of a new Diffie-Hellman group, they can specify its number in future phase one negotiations. Defining a new Diffie-Hellman group is not a trivial matter, so in practice, the group exchange is not commonly used.

8.25 IKE VERSION 2

A standard for version 2 of IKE has been in development for some time. According to Appendix A of the current draft of the IKE version 2 standards, motivations for developing the version 2 standard include the following:

1. Creating a single RFC that defines IKE (version 1 was defined through multiple RFCs)
2. Simplifying IKE, including the elimination of extraneous features such as the aggressive and group exchanges, and authentication through public key encryption
3. Establishing reliable message delivery, including acknowledged informational messages
4. Providing additional protection against denial of service attacks
5. Resolving issues with using IKE through NAT gateways
6. Fixing bugs
7. Defining how error conditions and ambiguous situations should be handled.

IKE version 2 supports the Extensible Authentication Protocol (EAP), which permits IPSec to use external authentication services such as Kerberos and RADIUS. IKE version 2 also includes the Peer Authorization Database (PAD), which includes the valid identities (*e.g.*, IP addresses) for peers and the valid authentication methods for each peer. Another significant functional difference between version 1 and version 2 is that version 2 can establish both the IKE SA and the IPSec SA in a total of 4 messages, as follows:

- In the first pair of messages, the endpoints negotiate various security parameters, as well as sending each other Diffie-Hellman values.
- In the second pair of messages, the endpoints authenticate each other and establish an IPSec SA.

Numerous IKEv2 implementations are expected to be available in 2006; it is likely that IKEv2 will be widely deployed a few years after that.

8.26 IP PAYLOAD COMPRESSION PROTOCOL (IPCOMP)

In communications, it is often desirable to perform lossless compression on data to repackage information in a smaller format without losing any of its meaning. For example, if host A wants to send host B a string of a thousand X, it would be more efficient to send host B a single X and tell it to use a thousand of them. Similarly, using a compression protocol for IPSec communications should improve the efficiency of IPSec in terms of network bandwidth because fewer bytes of data will need to be transmitted. However, there is a problem with this. Ideally, the process of encryption makes data appear random to an observer; for example, the letters, digits, and punctuation of an email message may be converted into many different printable and non-printable characters. Random data is very difficult to compress, because compression works by communicating the same information in a smaller format. Therefore, it is much more effective to first compress data and then encrypt it.

0	1	2	3
Version	Hrd len	Service type	Total length
Identification		Flags	Fragment offset
Time to live	Protocol	Header checksum	
Source IP address			
Destination IP address			
IP options (if needed)		Padding	
Data ...			

FIGURE 8.21 IP payload compression protocol

The IP Payload Compression Protocol (IPComp) is often used with IPSec. By applying IPComp to a payload first, then encrypting the packet through ESP, effective compression can be achieved. However, this is somewhat dependent on the data in each packet. For example, compression

provides little savings on very small payloads. Also, some data may already be compressed by an application or other means. In these cases, it is a waste of resources to compress the payload, as the overhead of compressing and decompressing the data outweighs the benefit of a trivial reduction in payload size (or no reduction at all). Accordingly, IPComp only uses compression if it actually makes the packet smaller. If it attempts to compress a packet and discovers that no benefit is gained, it will send the original, non-compressed packet so that the receiver does not waste resources performing decompression.

Each packet that has had compression applied will contain an IPComp header. Each header has three fields:

- **Next Header:** This field contains the IP protocol number for the packet payload, such as 6 for TCP or 17 for UDP.
- **Reserved:** This field is reserved for future use, so it should be set to 0.
- **Compression Parameter Index (CPI):** This is similar to the SPI. The CPI and the destination IP address form a compression security association.

IPComp can be configured to provide compression for IPsec traffic going in one direction only (*e.g.*, compress packets from endpoint A to endpoint B, but not from endpoint B to endpoint A) or in both directions. IPComp allows administrators to choose from multiple compression algorithms,

8.27 ESP IN A GATEWAY-TO-GATEWAY ARCHITECTURE

In this scenario, the goal is to establish an IPsec connection that provides encryption and integrity protection services between endpoints A and B. The IPsec architecture is gateway-to-gateway; endpoint A uses gateway A on network A, and endpoint B uses gateway B on network B. The first step in establishing the connection is to create an IKE SA (if one does not already exist), as follows:

1. Endpoint A creates and sends a regular (non-IPsec) packet that has a destination address of endpoint B.
2. Network A routes the packet to gateway A.
3. Gateway A receives the packet and performs NAT, altering the packets source IP address.

4. Gateway A matches the packets' characteristics against those in its Security Policy Database. It determines that the packet should be protected by encryption and integrity protection through ESP, and also determines the address of the destination gateway. Because the SPD entry does not have a pointer to an IKE SA, it knows that no IKE SA currently exists to protect this particular traffic.
5. Gateway A initiates an IKE SA negotiation with Gateway B using either main mode or aggressive mode. At the end of the negotiation, the IKE SA is created. The next step in establishing the ESP connection is to create IPSec SA as follows:
6. Gateway A uses the parameters set in the IKE SA to initiate an IPSec SA negotiation with gateway B. The IKE SA provides protection for the negotiation, which is performed using quick mode. The parameters specify that ESP tunnel mode will be used and that it will provide encryption and integrity protection. At the end of the negotiation, a pair of unidirectional IPSec SA is created for the ESP tunnel. Each SA provides protection only for traffic going in one direction.
7. Once the two IPSec SA have been created, gateway A can finish processing the packet sent by endpoint A in step 1. The packet will first be encrypted, then processed for integrity protection. The following steps outline how the data actually reaches its destination:
 - Gateway A modifies the packet so it is protected in accordance with the SA parameters. This includes adding a new IP packet header that uses gateway A's IP address as the source IP address, and gateway B's IP address as the destination address, encrypting the data, then adding the authentication information. Gateway A then sends the packet to Gateway B.
 - Gateway B receives the packet and uses the value in the unencrypted SPI field from the ESP header to determine which SA should be applied to the packet. After looking up the SA parameters (including the secret keys needed for integrity protection and decryption), gateway B processes and validates the packet. This includes removing the additional IP packet header, checking the integrity of the encrypted data, optionally performing a replay check, and decrypting the original payload. Gateway B checks the SPD to ensure that the required protections were applied to the packet then sends the packet to its actual destination, endpoint B.

If endpoint B wishes to reply to the packet, the last step of this process is repeated, except the parties are switched. Endpoint B would send a packet to endpoint A; routing would direct it to gateway B. Gateway B would modify the packet appropriately and send it to gateway A. Gateway A would process and validate the packet, apply NAT to restore the original IP address, then send the packet to endpoint A.

Assuming that the IPsec connection between the gateways is sustained, eventually the IKE or IPsec SA will approach one of the SA lifetime thresholds (maximum time, maximum bytes transmitted). The first gateway that determines the SA lifetime is approaching initiates its keying process. This causes some of the steps listed previously to be performed again, depending on which type of SA (IKE or IPsec) needs to be rekeyed. Once the new SAs have been created, the gateways send all new traffic over the new SAs, and eventually the old SAs are deleted. (The precise details of the rekeying process can vary significantly among IPsec implementations.)

8.28 ESP AND IPCOMP IN A HOST-TO-GATEWAY ARCHITECTURE

In this scenario, the goal is to establish an IPsec connection that provides encryption, integrity protection, and compression services between endpoints A and B. The IPsec architecture is host-to-gateway; endpoint A is located on network A, and endpoint B uses gateway B on network B.

The first step in establishing the connection is to create an IKE SA, as follows:

1. Endpoint A creates a regular (non-IPsec) packet that has a destination address of endpoint B. When endpoint A attempts to send this packet, its IPsec client software matches its characteristics against those in its Security Policy Database and determines that ESP and IPComp should be applied to the packet. It also determines the IP address of the destination gateway, Gateway B.
2. Endpoint A initiates an IKE SA negotiation with Gateway B using either main mode or aggressive mode. At the end of the negotiation, the IKE SA is created.

The next step in establishing the ESP and IPComp connection is to create IPsec SA as follows.

3. Endpoint A uses the parameters set in the IKE SA to initiate an IPsec SA negotiation with gateway B. The IKE SA provides protection for the negotiation, which is performed using quick mode. The parameters specify that ESP tunnel mode will be used and that it will provide both encryption and integrity protection, and that IPComp will also be applied. At the end of the negotiation, a pair of unidirectional IPsec SA is created for the tunnel, as well as a pair of IPComp SA.
 - Once the two IPsec SA have been created, endpoint A can finish processing the initial packet. The following steps outline how the data actually reaches its destination:
4. Endpoint A modifies the packet so it is protected in accordance with the SA parameters. IPComp is applied first, then ESP. This includes adding a new IP packet header that uses gateway B's IP address as the destination address. Endpoint A then sends the packet to Gateway B.
5. Gateway B receives the packet and uses the value in the unencrypted SPI field from the ESP header to determine which SA should be applied to the packet. After looking up the SA parameters, gateway B processes and validates the packet. This includes removing the additional IP packet header, decompressing the data (if necessary), performing the integrity verification, optionally performing a replay check, and decrypting the original payload. Gateway B checks the SPD to ensure that the policy was followed properly, then sends the packet to its actual destination, endpoint B.

If endpoint B wishes to reply to the packet, the last two steps of this process are repeated, except the parties are switched. Endpoint B would send a packet to endpoint A; routing would direct it to gateway B. Gateway B would modify the packet appropriately and send it to endpoint A.

8.29 ESP AND AH IN A HOST-TO-HOST ARCHITECTURE

In this scenario, the goal is to establish a transport mode IPsec connection that provides encryption and authentication between endpoints A and B. Because of security concerns, AH authentication has been selected instead of ESP authentication because AH can check the integrity of the IP header. The IPsec architecture is host-to-host, with both endpoints on the same network. The first step in establishing the connection is to create an IKE SA.

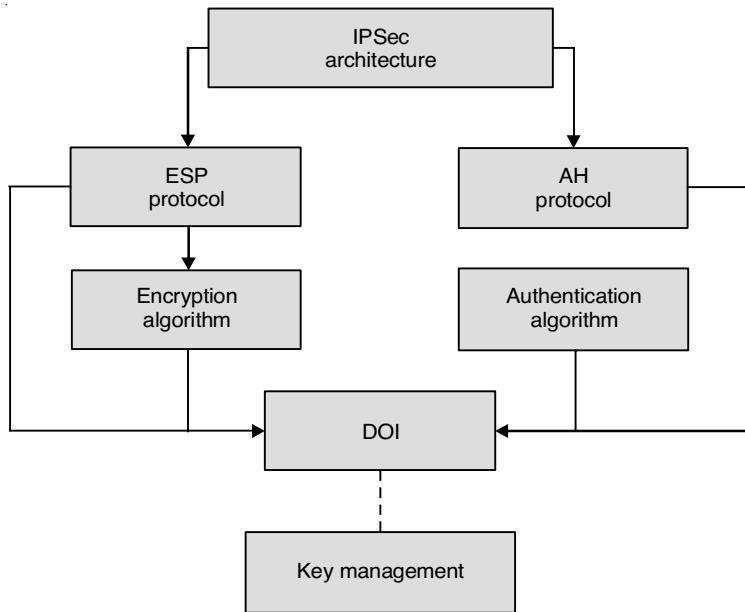


FIGURE 8.22 An illustration of the IPsec architecture that is host-to-host, with both endpoints on the same network

- Endpoint A creates a regular (non-IPSec) packet that has a destination address of endpoint B. When endpoint A attempts to send this packet, its IPSec client software matches its characteristics against those in its Security Policy Database and determines that ESP and AH should be applied to the packet. It also determines that the packet should be sent to endpoint B (*e.g.*, no need to change the address to point to an IPSec gateway).
- Endpoint A initiates an IKE SA negotiation with endpoint B using either main mode or aggressive mode. At the end of the negotiation, the IKE SA is created.
 1. The next step in establishing the ESP and AH connection is to create IPsec SA as follows:
 2. Endpoint A uses the parameters set in the IKE SA to initiate an IPsec SA negotiation with endpoint B for the AH service. The IKE SA provides protection for the negotiation, which is performed using quick mode. The parameters specify that AH transport mode will be used. At the end of the negotiation, a pair of unidirectional SA is created for the tunnel.
 3. Step 3 is repeated to negotiate the SA for the ESP service.

4. Once the four IPSec SA have been created, endpoint A can finish processing the initial packet. The following steps outline how the data actually reaches its destination 5. Endpoint A modifies the packet so it is protected in accordance with the SA parameters. ESP is applied first, then AH. (This allows AH to provide integrity for the ESP portions of the packet.) Endpoint A then sends the packet to endpoint B.
5. Endpoint B receives the packet and uses the SPI value from the AH header to determine which SA should be applied to the packet. After looking up the SA parameters, endpoint B processes and validates the packet, in terms of AH. Next, Endpoint B uses the value in the unencrypted SPI field from the ESP header to determine which SA should be applied to the packet next. After looking up the SA parameters, endpoint B processes and validates the packet, in terms of ESP. The IPSec client on endpoint B then releases the packet so the host can process it.
6. If endpoint B wishes to reply to the packet, the last two steps of this process are repeated, except the parties are switched. Endpoint B would create a packet and send it directly to endpoint A.

EXERCISES

1. Compare IPSec to CISCO encryption technology.
2. Discuss IPSec protocols.
3. What is the IKE protocol?
4. What is the Internet Engineering Task Force (IETF)?
5. What is NAT-Traversal?
6. Explain the gateway-to-gateway architecture.
7. Discuss Virtual Private Networks (VPNs).
8. Discuss the encryption process.
9. What is end-to-encryption?
10. What is ESP version 3?
11. Explain ESP and AH in a host-to-host architecture.
12. Explain the IP payload compression protocol (IPCOMP).
13. What is the site-to-site encryption?

CHAPTER 9

THE SECURITY OF EMERGING TECHNOLOGIES

Chapter Goals

- **Introduction**
- **Security of Big Data Analytics**
- **Security of Cloud Computing**
- **Security of the Internet of Things (IoT)**
- **Security of the Smart Grid**
- **Security of Supervisory Control and Data Acquisition (SCADA) Control Systems**
- **Security of Wireless Sensor Network (WSN)**

This chapter¹ discusses the security of recent emerging technologies, including big data analytics, cloud computing, Internet of Things (IoT), smart grids, Supervisory Control and Data Acquisition (SCADA) Control Systems, and Wireless Sensor Networks (WSNs).

¹ This chapter was published as Chapter 14 in *Network Security and Cryptography*. Sarhan Musa.

Copyright ©2018 by Mercury Learning and Information LLC. All rights reserved.

9.1 SECURITY OF BIG DATA ANALYTICS

Big data is a large-scale information management and analysis technology that exceeds the capability of traditional data processing technologies. Big data can be defined by two or more characteristics:

- **Volume:** A system that gathers large amounts of data
- **Variety:** The data being gathered and analyzed varies in structure and format
- **Velocity:** Data is gathered at a high speed
- **Value:** Significant value is derived from the analysis of the data
- **Visibility:** Data is accessed or visible from disparate or multiple geographic regions
- **Variability:** Data flows can be highly inconsistent with periodic peaks
- **Complexity:** Management of complex data that is coming from multiple sources. The data must be linked, matched, cleansed, and transformed into the required formats before processing.

Big data involves two types of processing:

- *Batch processing*: the analytics of the data at rest (for example, using Hadoop)
- *Stream processing*: the analytics of the data in motion (for example, using Storm)

Big data analytics is the process of analyzing and mining big data. It can produce operational and product knowledge at an unprecedented scale and specificity. The technological advances in the storage, processing, and analysis of big data include the following:

- a. Rapidly decreasing the cost of storage and CPU power in recent years
- b. Increasing the flexibility and cost-effectiveness of datacenters and cloud computing for elastic computation and storage
- c. Development of new frameworks such as Hadoop, which allow users to take advantage of these benefits through the implementation of distributed computing systems storing large quantities of data through flexible parallel processing.

9.1.1 Big Data Analysis Can Transform Security Analytics

- a. Accumulate data from various internal organizational sources and external sources to create a consolidated view of the required data into a vulnerability database.
- b. Perform in-depth analytics on the data using security intelligence, hence uncovering unique patterns that could be the source of many security issues.
- c. Provide a one-dimensional view of all the related information.
- d. Provide real time analysis of streaming data and use the previous results as feedback to the system.

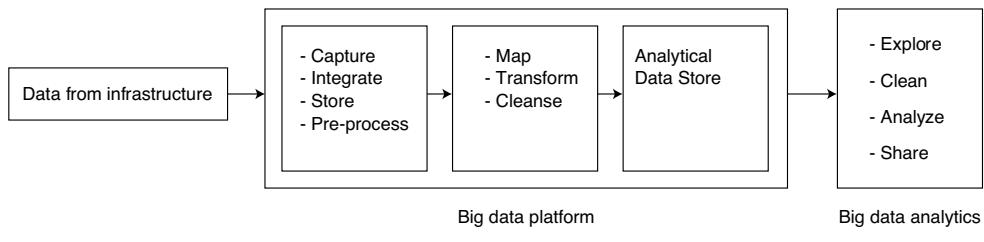


FIGURE 9.1 Big data flow

9.1.2 Big Data Analytics for Security Issues and Privacy Challenges

- **Protected database storage and transaction log file:** The availability and scalability of the system require auto-tiering for big data management. Auto-tiering solutions do not keep track of where the database is actually stored, but serve to protect database storage.
- **Secure computations in distributed frameworks:** Parallelism is used in computations and physical storage to process very large amounts of data. The MapReduce framework is an example. Protecting the mappers and protecting the data in the presence of untrusted mappers are two major attack prevention measures.
- **Privacy issues for non-relational data stores:** No SQL database embedded protection in the middleware. It does not provide any type of support for enforcing it explicitly in the database. The gathering aspect of No SQL databases imposes additional demands on the strength of privacy practices.

- **End-point input validation/filtering:** This method is used to identify trusted data and verify that the source of data input is not harmful.
- **Extensible privacy that preserves data mining and analysis:** Data mining and analytics raise privacy issues, and affect the appropriation of security, reduce civil freedoms, and increase state and corporate control. Anonymizing data for analysis is not sufficient to ensure user security.
- **Real-time security and compliance monitoring:** This method gives the number of alerts generated by privacy devices. These alerts lead to many false positives, which are mostly ignored. This approach is used to provide real-time problem detection based on scalable privacy analysis.
- **Granular audits:** This method is used for compliance, regulation, and forensics reasons. It is used to deal with data objects, which probably are allocated.
- **Cryptographically enforced access control and secure communication:** This method is used to ensure fairness, authentication, and agreement among the distributed entities.
- **Granulated access control:** This helps to secure access to data by limiting the people who have access. Granulated access control gives data managers the ability to share as much data as possible without compromising privacy.
- **Information security:** Tackling big data from a security point of view is a hard task. Thus, information security is a big data issue.
- **Metadata provenance:** Data increases in complexity due to large provenance graphs generated from provenance-enabled programming environments in big data applications. The analysis of such large provenance graphs to detect metadata dependencies for security and confidentiality purpose is computationally intensive.

9.2 SECURITY OF CLOUD COMPUTING

- *Cloud computing* is a type of computing in which services are delivered through the Internet. It depends on the sharing of computing resources, rather than having local servers or personal devices handle the applications.

- Cloud computing makes use of the increasing computing power to execute millions of instructions per second.
- Cloud computing uses networks of a large group of servers with specialized connections to distribute data processing among the servers. Instead of installing a software suite for each computer, this technology requires the installation of a single kind of software in each computer that allows users to log into a Web-based service; it also hosts all the programs required by the user.
- Local computers no longer have to take the entire burden when it comes to running applications.
- Cloud computing technology is being used to minimize the usage cost of computing resources.
- The cloud network, consisting of a network of computers, handles the load. The cost of software and hardware on the user end decreases. The only thing that must be done at the user's end is to run the cloud interface software to connect to the cloud.

Cloud computing consists of two ends:

- **Front end:** This includes the user's computer and software required to access the cloud network.
- **Back end:** This consists of various computers, servers, and database systems that create the cloud.

The user can access applications in the cloud network from anywhere by connecting to the cloud using the Internet. Some of the real time applications that use cloud computing are Gmail, Google Calendar, and Dropbox.

9.2.1 Cloud Deployment Models

- **Public cloud:** This cloud infrastructure is typically owned by an organization selling cloud services, known as a cloud service provider (CSP), and it delivers it to the general public on a subscription basis.
- **Private cloud:** This cloud infrastructure is operated only for an individual organization. The infrastructure may be managed by the organization itself or by a third party, and it may be located either on premises or off.

- **Community cloud:** This cloud infrastructure is shared by several organizations and supports a specific community that has shared concerns. The infrastructure may be managed by the organizations or by a third party and may be located on-premise or off-premise.
- **Hybrid cloud:** This cloud infrastructure is a combination of both private and public cloud instances that remain unique entities, but are bound together by standardized or proprietary technology that enables data and application portability.

9.2.2 The Three Layers of the Cloud Computing Services Model (Software, Platform, or Infrastructure (SPI) Model)

- *Software as a Service (SaaS)* provides a way of delivering centrally hosted applications over the Internet as a service. The user consumes a software application across the Internet. The user has no infrastructure or applications to manage and update, no setup or hardware costs, and application accessibility from any Internet connection. It is a user interface and a network-hosted application (such as IBM Lotus Live).
- *Platform as a Service (PaaS)* provides the service and management that is similar to the operating system. The CSP provides an additional layer on top of the infrastructure. Services include the operating system, network access, storage, database management systems, hosting, server-side scripting, and support. The user can use this environment and the tools provided to create software applications. It is a network-hosted software development platform (such as Google App Engine (GAE) and Microsoft Azure).
- *Infrastructure as a Service (IaaS)* provides the cloud services of the basic hardware, such as the CPU, network, and storage. The CSP provides the virtualized computing infrastructure. This generally includes virtual compute instances, network connectivity, IP infrastructure, bandwidth, load balancers, and firewalls. The user is responsible for installing and maintaining everything above the hypervisor (from the operating system upward). The provider hosts the customer's Virtual Machine (VMs) or provides network storage (for example, Amazon Web Service (AWS)).

9.2.3 Security Concerns and Challenges of Cloud Computing

- **Authentication:** Applications and data are accessed over the Internet, which increases the complexity of the authentication procedures.
- **Authorization:** This computing environment requires the use of the cloud service and providers' services for identifying the access policies.
- **Data integrity:** Appropriate mechanics are required for detecting accidental and intentional changes in the data.
- **Security of data while at rest:** Appropriate separation procedures are required to ensure the isolation between applications and data from different organizations.
- **Security of data while in motion:** Appropriate security procedures are required to ensure the security of data while in motion.
- **Auditing:** Appropriate auditing procedures are required to get visibility into the application, data accesses, and actions performed by the application's users.

9.2.4 Cloud Security as a Consumer Service

- Identity Services and Access Management Services
- Data Loss Prevention (DLP)
- Web Security
- Email Security
- Security Assessments
- Intrusion Management, Detection, and Prevention (IDS/IPS)
- Security Information and Event Management (SIEM)
- Encryption
- Business Continuity and Disaster Recovery
- Network Security

9.3 SECURITY OF THE INTERNET OF THINGS (IOT)

- The Internet of Things (IoT) encompasses connected devices, users, and cloud services via the Internet to enable and provide intelligent services for users. Figure 9.2 shows the IoT concept topology.

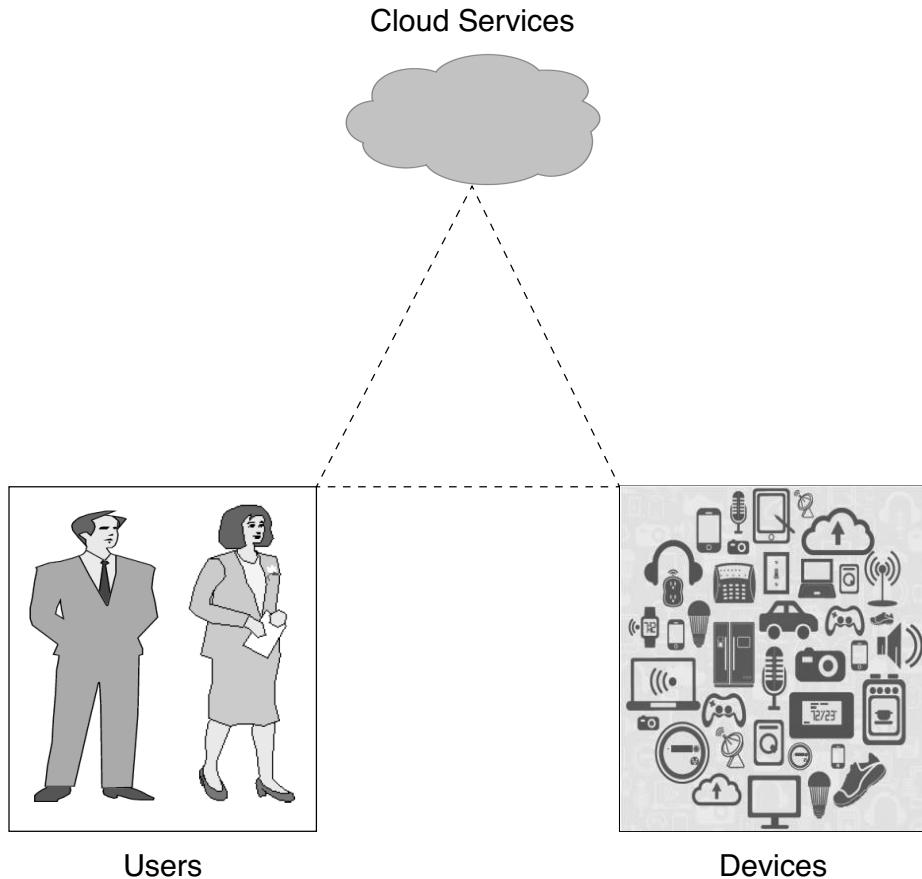


FIGURE 9.2 IoT concept topology

9.3.1 Evolution of the IoT

- Internet Service Providers (ISPs)
- Radio Frequency Identification (RFID)

- Application Service Providers (ASPs)
- Software as a Service (SaaS)

9.3.2 Building Blocks of the Internet of Things (IoT)

- **Sensors/Actuators:** Sensors and actuators are the tools that allow us to monitor and collect data, and control the devices in the IoT.
- **Devices:** Using sensors and actuators, these devices are more intuitive and efficient than we ever thought possible.
- **Gateways:** IoT gateways help devices intelligently communicate for greater efficiency, intuitive data management/classification, and increased security.
- **Master of Devices and Service Providers:** For every device or service in the IoT, there must be a master. This could be the device manufacturer, a cloud service provider, or an IoT solution provider. The master's role is to issue and manage devices, as well as facilitate data analysis.

Table 9.1 Differences between the IoT and Machine-to-Machine (M2M).

M2M	IoT
Focuses on connecting machines (or devices) for use in remote monitoring and control and data exchange— mainly proprietary closed systems	Addresses the way humans and machines connect using common public services
Uses either proprietary or non-IP based communication protocols for communication within its area network, such as ZigBee, Bluetooth, Z-Wave, ModBus, Power Line Communication (PLC), 6LoWPAN, and IEEE 802.15.4. It focuses on the protocols below the network layer.	It focuses on the protocols above the network layer, such as HTTP, CoAP, DDS, and XMPP.
Data is collected in point solutions and can be accessed by on-site applications and storage infrastructure.	Data is collected in the cloud and can be accessed by cloud applications.
Focuses on hardware with embedded modules	Focuses on software

9.3.4 IoT Layer Models

9.3.4.1 Three-Layer Model

- The *application layer* encompasses the information availability, user authentication, information privacy, data integrity, IoT platform stability, middleware security, and management platform (for example, remote medical services, cloud computing, smart grid, smart traffic, smart home, and environment monitoring).
- The *transport layer* is where DOS/DDOS attacks, forgery/middle attack, heterogeneous network attacks, WLAN application conflicts, capacity, and connectivity issues occur (it encompasses LAN, Ad hoc, GPRS, WiFi, and 3G/4G).
- The *sensing layer* is where data is collected and processed. This layer is affected by interruptions, interception, modification, fabrication, non-uniform coding for RFID, and conflict collision for RFID (this layer also includes WSN, RFID, RSN, MEMS, and GPS)

9.3.4.2 Four-Layer Model

- The *service layer* provides the interface and communicates with users. It uses lightweight security.
- The *platform layer* supports the IoT applications and services (for example, it encompasses the interface, context awareness, operating system, and cloud services). It uses privacy preservation for security.
- The *network layer* serves to transmit the data among devices, contents, services, and users. It processes, controls, and manages enormous amounts of network traffic (for example, context connection, context-based network, group mobility, and gateway PnP). It uses Authentication for security.
- The *device layer* perceives the environment with various sensing devices, processes it to send to the sink node or gateway, and responds to it if necessary (for example, actuator, sensor, resource management, and automatic control). It uses sensors' data integrity for security.

9.3.4.3 Seven-Layer Model

- *Physical Devices and Controllers (Layer 1):* This layer controls multiple devices. These are the “things” in the IoT, and they include a wide range of endpoint devices (computing nodes) that send and receive information, e.g., smart controllers, sensors, RFID readers, and different versions of RFID tags. Data confidentiality and integrity must be taken into account from this level upwards. (Secure content: silicon)
- *Connectivity (Layer 2):* This layer includes the communication and processing unit, and ensure the reliable and timely transmission of information. (Secure network access: hardware & protocols)
- *Edge (Fog) Computing (Layer 3):* Data element analysis and transformation. Secure communications (protocols and encryption)
- *Data Accumulation (Layer 4):* Storage and making network data usable by applications. Tamper resistant (software)
- *Data Abstraction (Layer 5):* Aggregation and access, including abstracting the data interface for applications. (Secure storage: hardware & software)
- *Application (Layer 6):* Reporting, analytics, and control. Authentication/Authorization (software)
- *Collaboration and Processes (Layer 7):* Involving people and business processes. Identity management (software)

9.3.5 Applications of the IoT

1. Wearables

- Entertainment
- Fitness
- Smart watches
- Location and tracking

2. Health Care

- Remote monitoring
- Ambulance telemetry
- Drug tracking

- Hospital asset tracking
- Access control
- Predictive maintenance

3. Building & Home Automation

- Access control
- Light and temperature control
- Energy optimization
- Predictive maintenance
- Connected appliances

4. Smart Cities

- Residential emeters
- Smart street lights
- Pipeline leak detection
- Traffic control
- Surveillance cameras
- Centralized and integrated system control

5. Automotive

- Infotainment
- Wire replacement
- Telemetry
- Predictive maintenance
- Car-to-Car (C2C) and Car-to-Infrastructure (C2I)

6. Smart Manufacturing

- Flow optimization
- Real time inventory
- Asset tracking
- Employee safety

- Predictive maintenance
- Firmware updates

Every connected device creates opportunities for attackers. These vulnerabilities are broad, even for a single small device. The risks posed include data transfer, device access, malfunctioning devices, and always-on/always-connected devices. The main challenges in security remain the security limitations associated with producing low cost devices, and the growing number of devices, which creates more opportunities for attacks.

9.3.6 New Challenges Created by the IoT

- **Security:** Prevent attackers from accessing the IoT
- **Privacy:** Protect identity and privacy data from attackers
- **Interoperability and standards:** Ensure that devices communicate securely
- **Legal and regulatory compliance:** Contribute to the legal, tax, and regulatory requirements regarding IoT-related business transactions that involve payment for goods and services at the international, federal, and state levels.
- **Ecommerce and economic development issues:** Connectivity and information sharing to be deployed globally by IoT and the economic rules of engagement for conducting business on the World Wide Web.

9.3.7 Security Requirements of the IoT

- *Confidentiality:* Ensure that only authorized users access the data and information
- *Integrity:* Ensure the completeness, accuracy, and absence of unauthorized data manipulation
- *Availability:* Ensure that all system services are available when requested by an authorized user
- *Accountability:* An ability of a system to hold users responsible for their actions and operations

- *Auditability*: An ability of a system to conduct persistent monitoring of all actions and operations
- *Trustworthiness*: An ability of a system to verify an accurate identity and establish trust in a third party
- *Non-repudiation*: An ability of a system to confirm occurrence/non-occurrence of an action and an operation
- *Privacy*: Ensure that the system obeys privacy policies and enables individuals to control their personal data and information

9.3.8 IoT Attacks

- **Attacks against a device**

An attacker takes advantages of IoT devices because many of the devices have an inherent value by the simple nature of their function. As devices are trusted with the ability to control and manage things, they also present a value for their ability to impact outcomes. The devices have a value based on what is entrusted to those devices.

- **Attacks against the communication between the devices and masters**

An attack involves monitoring and altering messages as they are communicated. The volume and sensitivity of data traversing the IoT environment makes these types of attacks especially dangerous, as messages and data could be intercepted, captured, or manipulated while in transit. All of these threats endanger the trust in the information and data being transmitted, and the ultimate confidence in the overall infrastructure.

- **Attacks against the masters**

Attacks against manufacturers, cloud service providers, and IoT solution providers have the potential to inflict the most amount of harm. These masters are entrusted with large amounts of data, some of it highly sensitive in nature. This data also has value to the IoT providers because of the analytics, which represent a core, strategic business asset—and a significant competitive vulnerability if exposed. Disrupting services to devices also poses a threat as many of the devices depend on the ability to communicate with the masters in order to function. Attacking a master also presents the opportunity to manipulate many devices at once, some of which may already be deployed in the field.

9.3.9 Hybrid Encryption Technique

A hybrid encryption technique provides for information integrity, confidentiality, and being non-repudiation in data exchange for the IoT.

A. *Creating a Key*

- Key production process in AES is used to create a key
- Two four-by-four matrixes (state and key) are used to produce a key for encryption
- Choose a place from the state matrix and a key from the key matrix randomly and produce public key of h by sender in the XOR operation
- The key of h is produced based on a hexadecimal. Then, the public key h is produced.

B. *Encryption*

- A message sent from the sender to the receiver is in a multinomial called a *message*. After making a multinomial message, the sender randomly chooses a multinomial like r from the collection like L_r .
- We can have a message by multinomial r . Therefore, it should not be revealed by the sender.

$$\text{Encryption} = p_r \times h + \text{message}$$

This message will be transmitted to the receiver as an encryption message with a security capability.

C. *Decryption*

- When the message is encrypted, the receiver tries to open the message by its private key or encrypt the message.
- The receiver has both private keys: f and f_p . In fact, f_p is conversed with multinomial of f , so it can be concluded that it will be $f \times f_p = 1$ message receiver multiplies a message on the part of private key that is displayed below with the parameter a :

$$\begin{aligned} a &= f \times \text{encryption} \\ a &= f \times (p_r \times h + \text{message}) \\ a &= f \times P_r \times h + f \times \text{message} \end{aligned}$$

- To choose the correct parameter, the coefficients of the polynomial formula between $-q/2$ and $p/2$ are selected.

- As $p = 3$, then it is drastically reduced and does not have any effect on the process:

$$p_r \times h = 0$$

$$a = f \times \text{message}$$

- Parameter b is calculated. Just multiply private key f in initial message that has been sent by the sender:

$$a = b = f \times \text{message}$$

$$\text{Decryption} = (f_p \times b) / x^2$$

- Whenever Decryption = Message, the message will reach security to the recipient without any problems.

D. *Digital signature*

- A digital signature is used for message validity and proof of identity and security.
- The message must go from the sender to the receiver, so the receiver of the former step acts as the sender now and the sender of the former step acts as the receiver.

$$\text{Encryption sign} = (\text{message} * f) / x^2$$

$$\text{Decryption} = (h / 2 \times f_p \times \text{Encryption sign}) / 2 \times h$$

9.3.10 Hybrid Encryption Algorithm Based on DES and DSA

- Regroup 64-bit data according to blocks and put the output into L_0 , R_0 (two parts, each of 32 bits). Its replacement rule is that it must exchange the 58th bit with the first, the 50th bit with the second, and so on. The last one is the original number.
- L_0 and R_0 are the two parts after the transposition output. L_0 is the left 32 bits of the output, and R_0 is the right 32 bits.
- Sub-key generation algorithm: The 8th, 16th 64th bit is a parity bit according to the DES algorithm rule, and it is not involved in DES operation. Keys actually use 56 bits, and these 56 bits are divided into two parts, C_0 and D_0 , each of 28 bits. The cycle starts on the left for the first time, to obtain C_1 and D_1 , then the C_1 (28 bits) and D_1 (28 bits) obtained are combined to form a 56-bit date. The selection transposition 2 through the narrow, so as to get a key K_0 (48 bits). (And $K_1, K_2 \dots K_{15}$ can be obtained in this way.)

- Calculate the date hash value of the encrypted cipher text using the secure hash algorithm SHA-1. The parameters used in the DSA signature are
 - p: primes of L bits long. L is a multiple of 64 bits; the range is 512–1024 bits
 - q: prime factors of 160 bits of $p - 1$
 - g: $g = h^{\frac{(p-1)}{q}} \bmod p$, h satisfies $h < p - 1$, $h^{\frac{(p-1)}{q}} \bmod p > 1$
 - x: $x < q$, and x is the private key
 - y: $y = gx \bmod p$, and (p, q, g, y) are the public keys

The signature process is as follows:

- P: generates a random number k, $k < q$
- P: computes $r = (g^k \bmod p) \bmod q$ and $s = (k^{(-1)} (H(m) + xr)) \bmod q$
- The result of the signature is (m, r, s) .

$H(m)$ is the hash value of m, and m is the plaintext to be signed or hash value of the plaintext. The final signature is in integers (r, s) , which is then sent to the authenticator.

- Calculate the encrypted digital signature. Use the SHA-1 algorithm after the reader receives the cipher text. If the signatures are consistent with that provided by the sender, then decrypt the ciphertext with the sub-key generated by the DES algorithm to generate the plaintext.

9.3.11 Advance Encryption Standard (AES)

AES is based on a design principle known as a substitution-permutation network, a combination of both substitution and permutation, and it is fast in both software and hardware. Unlike its predecessor DES, AES does not use a Feistel network. AES uses the Rijndael cipher, which has a fixed block size of 128 bits, a key size of 128, 192, or 256 bits, and is defined in three versions (10, 12 and 14, rounds, respectively).

There are three steps performed in AES

- encryption
- decryption
- key generation

AES Encryption:

Step 1: Get the plaintext and the key

Step 2: Perform the pre-round transformation using the plaintext

Step 3: With “n” key length, perform the transformation for “n” rounds

Step 4: Ciphertext achieved

AES Decryption:

Repeat the steps followed in encryption in reverse order:

Step 1: Cipher text

Step 2: With “n” key length, perform the transformation for “n” rounds

Step 3: Perform the pre-round transformation using the plaintext

Step 4: Get the plaintext and the key

Key Generation

Step 1: Get the key

Step 2: Based upon number of rounds, calculate the required number of words

Step 3: In an array of 4 bytes, the first four words are made from the key

Step 4: Get the next word

Step 5: Repeat step 4 until required number of words is reached.

9.3.12 Requirements for Lightweight Cryptography

- **Size (circuit size, ROM/RAM sizes):** determines the possibility of the implementation in a device
- **Power:** especially important with the RFID and energy harvesting devices
- **Power consumption:** important with battery-driven devices
- **Processing speed (throughput, delay):** A high throughput is necessary for devices with large data transmissions, while a low delay is important for the real-time control processing.

9.3.13 Lightweight Cryptography in the IoT

- **Efficiency of end-to-end communication:** The application of the lightweight symmetric key algorithm allows for a lower energy consumption of end devices.

- **Applicability to lower resource devices:** The lightweight cryptographic primitives open up possibilities of more network connections with lower-resource devices.

9.3.14 Prevention of Attacks on the IoT

Attackers exploit the vulnerabilities on IoT devices in an increasing number of distributed denials of service (DDoS) attacks. The prevention of attacks on the IoT can summarized in five steps:

- **Changing the default credentials of IoT devices:** Devices with weak or default credentials are vulnerable to compromise. IoT devices should be secured with strong authentication to avoid brute force attacks.
- **Disabling universal plug and play (UPnP) on gateway routers:** UPnP allows ports inside a network to be opened easily. Using UPnP, external computers are able to communicate to devices inside the network. To prevent this, we should disable UPnP on gateway routers. Some applications may be affected by disabling UPnP, so reconfiguration could be necessary.
- **Update IoT devices frequently:** Update IoT devices with the latest firmware and patches as soon as possible to ensure that the known vulnerabilities are addressed.
- **Ensure proper firewall configuration and identify malicious traffic:** Configure the firewall to block incoming User Datagram Protocol (UDP) packets because they are used to exploit IoT devices.
- **Review reliance on easily identified Internet connections:** Examine the level of reliance on public-facing web servers that are easy to identify externally and that are used for critical operations. It is important to review incident response procedures as well so that operations are not halted due to a cyber attack. In addition, IoT devices that are on public-facing servers should be secured to prevent unauthorized access.

9.4 SECURITY OF THE SMART GRID

- Smart grids utilize communication technology and information to optimally transmit and distribute electricity from suppliers to consumers. They are the next generation power system.

- The grid environment requires security mechanisms that have the following characteristics:
 - cross multiple administrative domains
 - have a high scalability in terms of a large and dynamic user population
 - support a large and dynamic pool of resources, each with different authentication and authorization policies
 - have the ability of grid applications to acquire and release resources dynamically during execution

9.4.1 Smart Grid Challenges

- the network congestion and safety related factors
- the lack of pervasive and effective communications, monitoring, fault diagnostics, and automation
- power grid integration, system stability, energy storage, which are introduced by the adaptation of renewable and alternative energy sources.

9.4.2 Smart Grid Layers

- Master Station System Layer
- Remote Communication Network Layer
- Terminal Layer
- Cross (Life Cycle of Information Systems) Layer
- Security Management Layer

9.4.3 Information Security Risks and Demands of Smart Grids

- **Master Station System Layer**
 - Physical layer attacks and protection
 - Network layer attacks and protection
 - Host layer attacks and protection
 - Application layer attacks and protection

- Data leaks and prevention; backup and recovery
 - Cloud computing application and its risks
 - Intercept and anti-intercept
- **Remote Communication Network Layer**
 - Monitor and anti-monitor
 - Tamper and anti-tamper
 - Encrypted communication channel
 - Fake terminal
 - Fake master station
 - Terminal integrity
 - Terminal network security
- **Terminal Layer**
 - Lack of computing, storage, and process resources to implement security schemes
 - Internet of Things application and its risks
 - Lack of security considerations in the system planning and system analysis stages
 - Lack of security design in the system design stage
- **Cross (Life Cycle of Information Systems) Layer**
 - Code security and secure system development in the system implementation stage
 - Security management in the system running and maintenance stage
 - Sensitive information processing in the system obsolescence stage
 - Social engineering attacks
 - How to build an effective information system
- **Security Management Layer:**
 - Security Management Systems for smart grids

9.4.4 Smart Grid Security Objectives

- **Availability:** Ensuring timely and reliable access to and use of information is of the most importance in the smart grid.
- **Integrity:** Guarding against improper information modification or destruction ensures information non-repudiation and authenticity.
- **Confidentiality:** Preserving authorized restrictions on information access and disclosure is mainly to protect personal privacy and proprietary information.

9.4.5 The Smart Grid System: Three Major Systems

- Smart Infrastructure System
- Smart Management System
- Smart Protection System

These major systems are also subdivided in other subsystems, applications and/or objectives, as shown in Figure 9.3.

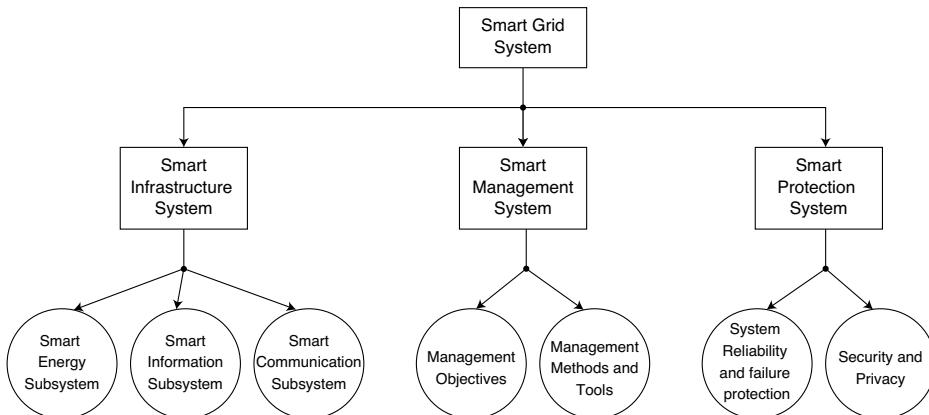


FIGURE 9.3 Smart Grid System Classifications

9.4.6 Types of Security Attacks that can Compromise the Smart Grid Security

- *Passive attacks* learn and use the system information without affecting the system resources. The attack target is only the transmitted information in order to learn the system configuration, architecture, and normal operation behavior.

- *Active attacks* affect the system operation through data modification or introducing false information into the system.

9.4.7 Cybersecurity Attacks on a Smart Grid

- *Eavesdropping* is a passive attack described as an unauthorized interception of an on-going communication without the consent of the communicating parties.
- *Traffic analysis* is similar to eavesdropping attack, but the attacker monitors the traffic patterns in order to infer useful information from it.
- *Replay* consists of captured transmitted messages and their retransmission in order to cause an unauthorized effect. The retransmitted messages are normally valid except the timestamp field.
- *Message modification* is similar to the replay attack, but the message is modified to cause unwanted behavior in the system. This attack can also involve message delay and reordering a message stream.
- *Impersonation* is when the intruder pretends to be an authorized entity or device.
- *Denial of Service* aims to suspend or interrupt the system communications. To accomplish this effect, the attacker can flood the communication network with messages to disable the physical components access, inhibiting the system's normal operation.
- *Malware* exploits internal weaknesses of the system with the goal of steal, modify, and destroy information and/or physical components of the system. It can also obtain unauthorized access to the system.

9.5 SECURITY OF SCADA CONTROL SYSTEMS

- A Supervisory Control and Data Acquisition (SCADA) system is a common process automation system that is used to gather data from sensors and instruments located at remote sites and to transmit data at a central site for either control or monitoring purposes.
- SCADA systems evolved from hardware and software that include standard PCs and operating systems, TCP/IP communications, and Internet access. SCADA systems can monitor and control hundreds to thousands of I/O points.

- SCADA systems differ from Distributed Control Systems (DCSs); DCSs cover plant sites while SCADA systems cover much larger geographic areas.
- SCADA architecture supports TCP/IP, UDP, or other IP-based communications protocols as well as strictly industrial protocols, such as Modbus TCP, Modbus over TCP, or Modbus over UDP, all working over private radio, cellular, or satellite networks.

9.5.1 Components of SCADA Systems

- Instruments that sense process variables
- Operating equipment connected to instruments
- Local processors that collect data and communicate with the site's instruments and operating equipment such as the Programmable Logic Controller (PLC), Remote Terminal Unit (RTU), Intelligent Electronic Device (IED), or Process Automation Controller (PAC)
- Short range communications between local processors, instruments, and operating equipment
- Host computers as central point of human monitoring and control of the processes, storing databases, and display of statistical control charts, and reports. A host computer is also known as the Master Terminal Unit (MTU), the SCADA server, or a PC with Human Machine Interface (HMI)
- Long range communications between local processors and host computers using wired and/or wireless network connections

9.5.2 SCADA System Layers

- *Supervisory Control Layer:* The supervisory control layer (the control center) is responsible for monitoring the operation of SCADA systems by gathering data from field devices, performing control and supervisory tasks, and sending control commands to the field controllers through the communication network. The supervisory control layer of a SCADA system consists of following elements:
 - SCADA server
 - builder server

- communication server
 - database server
 - diagnostic server
 - application server
 - human-machine interface
 - system operators
- *Automatic Control Layer:* The automatic control layer (regulatory control layer) is in charge of regulating the operation of physical processes based on control commands from the control center and sensor measurements from field devices. The control signals are then transmitted to field devices through the communication network. Various system variables, including control commands, sensor measurements, and control signals, are gathered within the control center for supervisory and management purposes. The automatic control layer of SCADA system consists of following elements:
 - master terminal units (MTUs)
 - remote terminal units (RTUs)
 - programmable logic controllers (PLCs)
 - intelligent electronic devices
 - *Physical Layer:* The physical processes (e.g., electric power grids, gas pipelines, and water networks) are equipped with actuators (e.g., motors, compressors, pumps, and valves), sensors (e.g., temperature sensors, pressure sensors, flow sensors, level sensors, and speed sensors), and protection devices (e.g., circuit breakers and protective relays) to realize technological goals. The physical elements are controlled and monitored by the control center through the automatic control layer and the communication network.

9.5.3 Requirements and Features for the Security of Control Systems

- Critical path protection
- Strong safety policies and procedures

- Knowledge management
- System development skills
- Enhanced security for device
- Sensor networks solutions
- Operating system based on microkernel architecture
- Increasing quality of software with security features
- Security requirements early in the software development cycle
- Compliance to standards for software development
- Integration of different technologies
- Vulnerability analysis based on proactive, discovery, and adaptation solutions
- Innovative risk management approaches
- Ensure authentication, confidentiality, integrity, availability, and non-repudiation
- Calculate risk in regards to security and safety

9.5.4 Categories of Security Threats to Modern SCADA Systems

- Insiders
- Hackers
- Hidden criminal groups
- Nation-states

9.6 SECURITY OF WIRELESS SENSOR NETWORKS (WSNS)

- A Wireless Sensor Networks (WSN) is an infrastructure-less network composed of large number of sensor nodes. These cooperatively sense and control the environment to enable its interaction with people or devices.
- Wireless Sensor Networks (WSNs) are considered one of the core technologies in implementing Cyber Physical Systems (CPSs).

- WSNs include sensor nodes, actuator nodes, gateways and clients. A large number of sensor nodes deployed randomly inside of or near the monitoring area form networks through self-organization.
- The data is captured at the level of the sensor node, compressed, and transmitted to the gateway. Through the gateway connection, data is then passed by the base station to a server.
- Sensor nodes monitor the collected data to transmit along to other sensor nodes by hopping. During the process of transmission, the monitored data may be handled by multiple nodes to get to gateway node after multi-hop routing, and finally reach the management node through the Internet or satellite.
- It is the user who configures and manages the WSN with the management node, publishes monitoring missions, and collects the monitored data.

9.6.1 WSN layers

- The *transport layer* is responsible for managing the end-to-end connections and reliable transport of data.
- The *network and routing layer* is responsible for the routing of sensors based on addressing and location awareness, sensor networking, power efficiency, and topology management. It provides more effective routing for the data from node to node, node to sink, node to base station and node to cluster head, and vice versa. Due to the broadcast method, every node works as a router.
- The *data link layer* is responsible for the multiplexing of data streams, data frame detection, medium access, and error control
- The *physical layer* is responsible for frequency selection, carrier frequency generation, signal detection, modulation, and data encryption

9.6.2 Security Requirements in WSNs

- **Data Confidentiality:** To provide the data confidentiality, encrypted data is used so that only the recipient decrypts the data to its original form. Only the authorized sensor nodes can get the content of the messages.
- **Data Integrity:** Data received by the receiver should not be altered or modified. Original data is changed by intruder or the environment.

The intruder changes the data and sends this new data to the receiver. Thus, any received message has not been modified in sent by unauthorized parties.

- **Data Authentication:** This is the procedure of confirmation that the communicating node is the one that it claims to be. The receiver node needs to verify that the data is received from an authenticate node.
- **Data Availability:** The services are available all the time.
- **Source Localization:** For data transmission, some applications use location information of the sink node. It is essential to give security to the location information. Non-secured data can be controlled by the malicious node.
- **Self-Organization:** In WSNs, no fixed infrastructure exists, hence, every node is independent, having properties of adaptation to the different situations and maintains self-organizing and self-healing properties.
- **Data Freshness:** Each message transmitted over the channel is new and fresh. It guarantees that the old messages cannot be replayed by any node. This can be solved by adding some time related counter to check the freshness of the data.

9.6.3 WSN Attack Categories

- **Outsider vs Insider Attacks:** Outsider attacks are external attacks and the insider attacks are internal attacks. An outsider attack comes from outside the WSN. In an outsider attack, bad data is inserted in the network for the services interruption. An insider attack is also known as the internal attack, these attacks come from the inside of the WSN.
- **Passive vs. Active Attacks:** Passive attacks include eavesdropping on or monitoring packets exchanged within the WSN; in active attacks, an attacker has the capability to remove or modify the messages during the transmission on the network.
- **Monte-class vs. laptop-class attacks:** In Monte-class attacks, an adversary attacks a WSN by utilizing a few nodes with similar capabilities to the network nodes; in laptop-class attacks, an adversary can use more powerful devices to attack a WSN. These devices have greater transmission range, processing power, and energy reserves than the network nodes.

Table 9.2 Attacks and Defenses in WSNs at Different Layers.

Layer	Possible Type of Attack	Defense
Transport	Flooding Desynchronization	Client puzzles Authentication
Network and Routing	Spoofed, altered or replayed routing information Selective forwarding Sinkhole Sybil Wormholes Hello flood attacks Acknowledgment spoofing	Egress filtering, authentication, monitoring Redundancy, probing Authentication, monitoring, redundancy Authentication, probing Authentication, packet leashes by using geographic and temporal information Authentication, verify the bidirectional link Authentication
Data Link	Collision Exhaustion Unfairness	Error-correcting code Rate limitation Small frames
Physical	Jamming Tampering	Spread-spectrum, priority messages, lower duty cycle, region mapping, mode change Tamper-proofing, hiding

9.6.4 Security Protocols in WSNs

- **Sensor Protocols for Information via Negotiation (SPINs)**
 - SPIN is an adaptive routing protocol, which transmits the information first by negotiating.
 - The SPIN transmission is data-centric; it is only transmitted to the nodes that have interest in the data. This process continues until the data reaches the sink node. SPIN reduces both the network overhead and the energy consumption in the transmission. There will not be

duplicate messages in the network since the nodes negotiate before transmitting the data.

- SPIN makes use of the metadata of the actual data to be sent. Metadata contains the description of the message that the node wants to send. The actual data will be transmitted only if the node wishes to receive it. SPIN makes use of three messages namely,
 1. ADV: Before sending a message, a node first generates the descriptor of the message to be sent. This metadata is exchanged by making use of ADV message. ADV message informs the size, contents and requirements of the message. This helps the receiving node on deciding transmission of the message.
 2. REQUEST: After receiving the ADV message, the receiver node verifies the descriptor (whether the message is a duplicate and whether the receiver node's battery capabilities are enough to transmit the data). If the node is interested in the data, it replies with a REQUEST message to the sender node.
 3. DATA: If the sender node receives a REQUEST message, it starts the actual transmission of the data by making use of the DATA message. This is the actual data transfer phase.

• **Localized Encryption and Authentication Protocol (LEAP)**

- LEAP is a protocol with a key management scheme that is very efficient with its security mechanisms and it is used for large scale distributed sensor networks. It is designed to support in-network processing, such as data aggregation. In-network processing results in reduction of the energy consumption in network. To provide the confidentiality and authentication to the data packet, LEAP uses multiple key mechanisms. For each node, four symmetric keys are used as follows:
 - *Individual Key*: Used for the communication between source node and the sink node
 - *Pair wise Key*: Shared with another sensor nodes
 - *Cluster Key*: Used for locally broadcast messages and shares it between the node and all its surrounding neighboring nodes
 - *Group Key*: Used by all network nodes

- **TINYSEC**

- TINYSEC is a lightweight protocol and link layer security architecture for WSNs.
- It supports integrity, confidentiality and authentication. To achieve confidentiality, encryption is done by using CBC (Cipher-Block Chaining) mode with ciphertext stealing, and authentication is done using CBC-MAC. No counters are used in TINYSEC. Hence, it doesn't check the data freshness. Authorized senders and receivers share a secret key to compute a MAC.
- TINYSEC has two different security options. One is for authenticated and encrypted messages (TINYSEC-AE) and another is for authenticated messages (TINYSEC-Auth).
- In TINYSEC-AE, the data payload is encrypted and the received data packet is authenticated with a MAC.
- In TINYSEC-Auth mode, the entire packet is authenticated with a MAC, but on the other hand the data payload is not encrypted.
- In CBC, Initialization Vector (IV) is used to achieve semantic security. Some of the messages are same with only little variation. In that case IV adds the variation to the encrypted process. To decrypt the message receiver must use the IV. IVs are not secret and are included in the same packet with the encrypted data.

- **ZIGBEE**

- ZIGBEE is a worldwide open standard for wireless radio networks in the monitoring and control fields.
- IEEE 802.15.4 is a standard used for ZIGBEE. The IEEE 802.15.4 standard defines the characteristics of the physical and MAC layers for Low-Rate Wireless Personal Area Networks (LR-WPAN).
- To implement the security mechanism ZIGBEE uses 128-bit keys. A trust center is used in ZIGBEE that authenticates and allows other devices/nodes to join the network and also distribute the keys.
- Three different roles in ZIGBEE are:
 - Trust Manager: It authenticates the devices that request to join the network.

- Network Manager: It manages the network keys and helps to maintain and distribute the network keys.
 - Configuration Manager: It configures the security mechanism and enables end-to-end security between devices.
- A ZIGBEE network can adopt one of the three topologies: star, tree, and mesh.

EXERCISES

1. Define big data and big data analytics.
2. What are the characteristics of big data?
3. What are the big data processing groups?
4. Draw the big data flow.
5. What are the security issues and privacy challenges big data analytics?
6. Define cloud computing.
7. What are the two ends of cloud computing?
8. What are the cloud computing deployment models?
9. What are the three layers of the cloud computing services model and its services?
10. What are the security concerns and challenges of cloud computing?
11. Define the Internet of Things and draw its concept topology.
12. What are the building blocks of the IoT?
13. Make a map comparison between the IoT and M2M.
14. What are the layered models of the IoT?
15. What are the applications of the IoT for automation, health care, smart manufacturing, and smart cities?
16. What are the new challenges created by the IoT?
17. What are the security requirements of the IoT?
18. What are the requirements for lightweight cryptography?

19. What are the three primary targets of attacks against the IoT?
20. What are the five steps for the prevention of attacks on the IoT?
21. Define smart grids.
22. What are the requirements of security mechanisms' characteristics of the grid environment?
23. What are the challenges of the smart grid?
24. What are the smart grid's layers?
25. What are the information security risks and demands of a smart grid?
26. Draw the smart grid system classifications.
27. List the possible cybersecurity attacks in a smart grid.
28. Define SCADA.
29. What are the components of SCADA systems?
30. What are the SCADA system layers?
31. What are the categories for security threats to modern SCADA systems?
32. Define WSN.
33. What are the WSN Layers?
34. What are the security requirements in WSNs?
35. What are the attack categories in WSNs?
36. List the attacks and defenses in WSNs at different layers.
37. What are the security protocols in WSNs?
38. Define the ZIGBEE protocol.

INDEX

A

Access Control, 8, 182
Access Networks, 236
Adaptive equalizers, 235
Address Resolution Protocol, 83, 258
Addressing, 70–71
Admission Control, 253
Aggressive Mode, 298–299
Analytics, Big Data, 312–313
Applet and ActiveX Controls, 15–17
Application Layer, 284, 320
Application Software, 72
Approximation, 41, 231
Architecture Models, 282, 283
ARP Spoofing, 83
Asymmetric Algorithms, 61–62
Asymmetric Encryption, 58, 162
Asymmetric Public Key, 57–58
Attacker Substitutes, 153
Attacker’s Physical Presence, 47
Attacks, 47–49, 76
Attempts Originating, 111
Authentication, 6–7, 55, 182, 196, 317
Authentication Header, 272–273
Authentication Mechanism, 6, 45
Authentication Methods, 35
Automatic Rekeying, 38–39
Auxiliary Information, 29

B

Backbone Service Provider, 51
Backup and Fault Tolerance, 170
Basic Receiver, 264
Best Effort Service, 250
Big Data, 312
Biometrics, 39–42
Biometrics Industry, 39
Block Ciphers, 184–185
Blocking Broadcast Traffic, 103–104
Breaking Ciphers, 185
Bridge Protocol Data Units, 79
Buying Versus Building, 140–142

C

Centralized Certificates, 32–34
Certification Authorities, 169–170
Certification Authority Hierarchies, 168–170
Channel Data Rate, 247
Cisco Discovery Protocol, 52–53
Cisco Systems, 51, 244, 246, 258
Cisco’s Point-to-Point, 244–245
Cloud Computing, 314–315
Code Division Multiple Access, 227
Combination of Routers, 131
Committed Information Rate, 250
Communication Session, 78
Compression Parameter Index, 304
Computing Environment, 197, 317

Confidentiality, 6, 74, 181, 323, 332
 Content Addressable Memory Table, 80–81
 Control Your Secrets, 159
 Controlled Point, 131
 Conventional Encryption, 34, 38
 Cookies, 17–19, 296
 Core Networks, 236
 Correct Configuration, 137
 Count the Cost, 158
 Countermeasures, 100–101
 Cryptanalysis, 34, 186
 Cryptographic Hash, 206, 208
 Cryptographically, 31, 35, 206
 Cryptographically Strong, 173, 174
 Cryptography, 53, 181–183
 Cultural Issues, 5

D

Data Confidentiality, 267, 337
 Data Encryption Standard, 189–190
 Data Link Layer, 77, 78, 285, 286, 337
 Data Origin Authentication, 268
 Data Packets, 71
 Denial of Service, 114–116, 333
 Denial of Service Attack, 12, 86, 116,
 148–149, 152–153
 Denial of Service (DoS), 114–116, 333
 Deployment, 237–238
 Designated Port, 94–95, 101
 Dictionary Program, 150
 Difficult to be Traced, 76
 Diffie-Hellman, 192–193
 Diffie-Hellman (DH) Group, 295–298
 Digital Certificates, 33, 191, 196, 201, 269
 Digital Signatures, 30–31, 294
 Digital Subscriber Line, 69
 Directory Services, 61
 Distribute Network, 146, 147
 DNS Spoofing, 23, 203–210
 Domain Name System, 23, 72
 DoS Attack, 12, 86, 116, 148–149, 152–153

Duplexing Techniques, 248
 Duplications of Payload, 226

E

Ease Network Administration, 168
 Email Security, 36–37, 163–164
 Encode Information, 2
 Encrypted Digital Signature, 57, 58
 Encryption, 53, 187, 325
 Encryption Process, 274, 289
 Enormous Destruction, 74
 ESP Implementations, 292
 ESP Packet Fields, 289–291
 Ethernet Network, 118
 Example Implementation, 239
 Extensible Authentication Protocol, 302
 External Authentication, 294–295
 External Routers, 157
 External Sources, 150, 222

F

Fabrication, 7, 11
 Face recognition, 40
 Failure to Receive Hello BPDUS, 96–97
 Famous Software Vendors, 124
 Firewalls, 114, 130–135, 138
 Frame and Slot Format, 251

G

Gateway-to-gateway, 285–286
 Gateway-to-gateway Architecture, 279–280
 Geographic Divisions, 170
 Great Opportunities, 144
 Group Exchange, 302

H

Handshake Authentication Protocol, 37
 Hardware Components, 131, 245, 246
 Hash Functions, 31–32, 62–63, 193

Hashing Algorithm, 194, 212
 High-Gain Receiver, 264
 Hijacking, 86, 150
 Host Security, 4, 127–128
 Host-to-gateway, 286
 Host-to-gateway Architecture, 280–282
 How ESP Works, 291–292
 Human Factors, 159

I

Identify Any Assumptions, 158
 Identity Protection, 275
 IKE Protocol, 274–276, 293
 Illusion of Security, 137
 Improved Bandwidth Usage, 103
 Informational Exchange, 301
 Infrastructure for Certifying, 196
 Infrastructure Security, 74
 Initialization Vector, 274, 290, 341
 Innovative Purchasing, 127
 Installed on the Server, 163
 Interception, 6, 57
 Interconnected Computers, 66
 Internal Security Measures, 135
 Internet Data Access, 241
 Internet Firewall, 107–108, 130–132
 Internet Infrastructure, 65–68, 74
 Internet Key Exchange (IKE), 269, 272, 293
 Internet of Things (IoT), 318
 Internet of Things, Applications, 321–323
 Internet of Things, Models, 320–321
 Internet Services, 66, 116–117
 Intrusion, 113–114
 IP Security Encryption Router, 50–53
 IP Spoofing Attack, 148, 152
 IP Wireless Open Standards, 245–246
 IP Wireless System Advantages, 239–241
 IPSEC Network Security, 269–272
 IPSec Protocols, 270–276
 IPSec Security Protocol, 300

J

Java Application Security, 21
 Java Security, 20–22
 Joy Riders, 119–120

K

Key Distribution, 59–61
 Key Management Security Association, 275
 Key Recovery Entry, 171–172
 Key Recovery Methodologies, 171
 Key-based Cryptosystems, 62–63
 Key-based Methodology, 54
 Keystroke Intervals, 42
 Know Your Weaknesses, 159–160

L

Larger Scale of Modification, 75–76
 Lightweight Certificate Authority, 172
 Limit the Scope of Access, 160
 Limit Your Trust, 160
 Local Area Network (LAN), 67, 221

M

MAC Flooding, 79–81
 MAC Flooding Attacks, 80, 81
 Make Security Pervasive, 160
 Man-in-the-Middle Attack, 86–87
 Management of Routers, 50
 Management Problems, 130
 Marvelous Opportunities, 132
 Mathematical Functions, 215, 217
 Mathematical Preliminaries, 214–218
 Mathematical Scheme, 30
 Medium Businesses, 239, 241–242
 Message Contents, 10
 Message Integrity, 164, 193–194
 Model Comparison, 282–283
 Modification, 7, 11
 Modular Exponentiation, 218
 Modular Multiplication, 216–217

Multifactor Authentication, 44–45
 Multiple Access Technique, 248–251
 Multiple Applications, 147
 Multiple Perimeter Networks, 156
 Mutual Authentication, 44

N

NAT-Traversal, 276–277
 Negligible Probability, 29
 Network Access Point (NAP), 66–67
 Network Infrastructure, 67–68
 Network Layer, 284–285, 320
 Network Management, 236–237
 Network Packet Sniffers, 146–148, 151–152
 Network Security, 4, 128–132
 Network Taps, 116–118
 Node-to-node Encryption, 286–287

O

One-way Functions, 27–30
 Open Systems Interconnection, 70
 Organizational Divisions, 169
 Orthogonal Frequency-Division Multiplexing, 246–247, 251
 Outdoor Unit (ODU), 243, 258

P

Packet Sniffing, 22
 Padding Length, 290
 Passive Attacks, 10, 178, 332, 338
 Password Attacks, 149–150, 152
 Payroll Department, 148
 Peer Authorization Database, 302
 Perimeter Networks, 156–158
 Periodic Reviews, 5
 Persistent TCN Messages, 99
 PGP Versus PEM, 166
 Phase One Exchange, 293
 Phase Two Exchange, 299–301
 Physical Coaxial, 257, 258
 Plaintext Attack, 185–186

Point of Presence (POP), 66, 236
 Point-to-Point Protocol, 37, 77
 Poisoned Cache, 204
 Potential Security Hole, 158
 Predictable Regularity, 184
 Premises Networks, 235–236
 Pretty Good Privacy, 165, 190
 Primary Signal, 232–234
 Principles of Security, 5–9
 Privacy-Enhanced Mail, 166
 Private Network, 144
 Protected Connection, 281, 300
 Protected Resources, 22, 46
 Protective Device, 109–113
 Pseudorandom Noise Code, 227
 Public Key Cryptography, 30, 43–44, 191, 278
 Public Key Encryption, 294
 Public Key Information, 43
 Public Key Infrastructure, 168, 196–197

Q

Quittner-Slatalla, 114, 115

R

Radio Resource Management, 255
 Random Data, 303
 Random Key Generation, 34–35
 Real-time Polling Service, 249–250
 Registration Information, 125
 Remember Physical Security, 160
 Replay Attacks, 11–12
 Requirement to Cell Radius, 253–256
 Resource Record Sets, 206
 Resources, 111
 Restrictive Definitions, 139
 Rooftop Unit, 263–264
 Root Bridge, 91–92, 94, 95, 101
 Root Claim, 98–99
 Root Guard, 101–102
 Routers, 50, 69–70

Routing Loop, 73
 RSA Algorithm, 214–219
 RSA Description, 218–219
 Rudimentary Cipher, 183, 184

S

Scorekeeper, 120–121
 Secret Key Exchange, 197–201
 Secure Email Protocols, 164–167
 Secure File System, 208
 Secure Naming, 202–203
 Secure Sockets Layer, 161–163, 178, 210–214
 Security Algorithms, 265
 Security and Performance Tradeoff, 76
 Security Approaches, 4–5
 Security, Firewall, 118–119
 Security Holes, 2, 125
 Security Mechanisms, 2, 10, 155, 330, 340
 Security Objectives, 143–150
 Security Parameter Index, 217, 273, 274
 Security Perimeter, 155–156
 Security Policy, 4, 155, 156, 267–269
 Security Problem, 127, 139, 202
 Security Through Obscurity, 4, 124–127
 Self-certifying Names, 208–210
 Sensitive Information, 147, 148, 150, 153
 Sequence Spread Spectrum, 226
 Service Disruption, 74
 Shared-bandwidth, 242
 Site-to-site Encryption, 287–288
 Smart Grids, 329
 Smart Grids, Attacks, 332–333
 Social Engineering, 148, 187
 Software Piracy, 113
 Sometimes Compression, 271
 Source Software, 141
 Spanning Tree Protocol, 79, 89
 Spatial Diversity, 229
 Spectrum Management, 255
 Spies Industrial and Otherwise, 121–122
 Spoofed Address, 22

Spoofing WAN Traffic, 86–88
 Spread Spectrum, 226–227
 Spreading Security Decisions, 133
 Static ARP Entries, 88
 STP Attack Scenarios, 97–98
 Stupidity and Accidents, 116
 Supervisory Control and Data Acquisition (SCADA) Control Systems, 333–336
 Symmetric Key Cryptosystems, 188, 191
 Symmetric Key Encryption, 188–189
 Symmetric Session Key, 56, 61
 System Security, 71–73

T

Tapping a Network, 117
 TCP/IP Network Security Protocol, 283–286
 Technology Offline Message Keys, 223–224
 Theoretical Attacks, 123
 Ticket-Granting Server, 38
 Time of Day Protocol, 259
 Time-Slotted Upstream, 255–256
 TINYSEC, 341
 Topology Change, 89–90
 Traffic Policing, 256–257
 Transmission Facility, 233
 Transmitting the Ticket, 200
 Transport Layer Products, 262–264
 Trapdoor Functions, 27

U

Understand Your Environment, 160
 Unknown Networks, 155
 Unsolicited Grant Service, 249, 250
 Untrusted Networks, 155
 User Security Component, 171

V

VAN with Firewall Security, 129
 Verification of Sender, 164
 Verification Template, 40

- Virtual Private Networking (VPN), 277–279
Viruses, 13
Virus Protection, 135, 137
Visible Interruption, 130
- Web-based Email Services, 167–168
Wide Area Network (WAN), 67, 280
Wireless Protocol Stack, 258–260
Wireless Sensor Network, 336
Wireless Technology, 222
WLAN Standards Comparison, 265–266

W

- Web Security, 201–203
Web Security Considerations, 175–178
Web Traffic Security Approaches, 178

Z

- ZIGBEE, 341–342