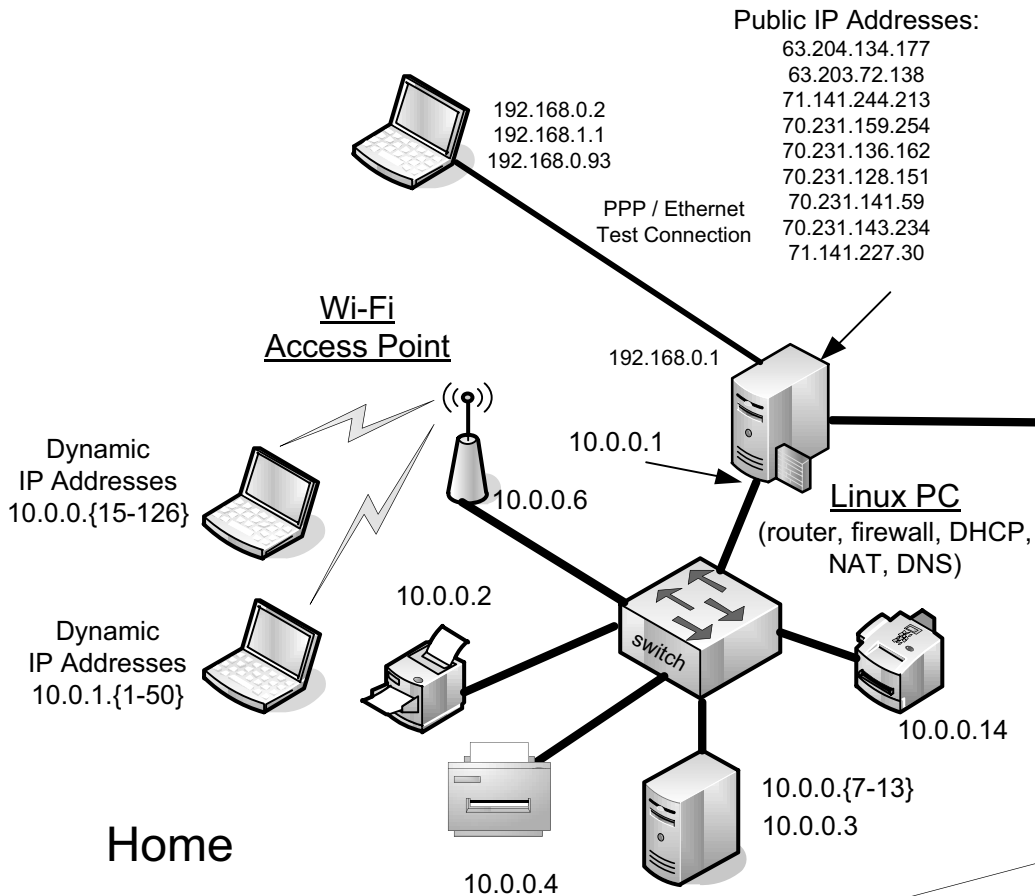


# TCP/IP Illustrated, Volume 1

The Protocols  
**SECOND EDITION**

Kevin R. Fall  
W. Richard Stevens





# **TCP/IP Illustrated, Volume 1: The Protocols**

## **Table of Contents**

Contents

Foreword

Preface to the Second Edition

Adapted Preface to the First Edition

Chapter 1 Introduction

1.1 Architectural Principles

1.1.1 Packets, Connections, and Datagrams

1.1.2 The End-to-End Argument and Fate Sharing

1.1.3 Error Control and Flow Control

1.2 Design and Implementation

1.2.1 Layering

1.2.2 Multiplexing, Demultiplexing, and Encapsulation in Layered Implementations

1.3 The Architecture and Protocols of the TCP/IP Suite

1.3.1 The ARPANET Reference Model

1.3.2 Multiplexing, Demultiplexing, and Encapsulation in TCP/IP

1.3.3 Port Numbers

1.3.4 Names, Addresses, and the DNS

1.4 Internets, Intranets, and Extranets

1.5 Designing Applications

1.5.1 Client/Server

1.5.2 Peer-to-Peer

# **Table of Contents**

1.5.3 Application Programming Interfaces (APIs)

## **1.6 Standardization Process**

1.6.1 Request for Comments (RFC)

1.6.2 Other Standards

## **1.7 Implementations and Software Distributions**

## **1.8 Attacks Involving the Internet Architecture**

## **1.9 Summary**

## **1.10 References**

# **Chapter 2 The Internet Address Architecture**

## **2.1 Introduction**

## **2.2 Expressing IP Addresses**

## **2.3 Basic IP Address Structure**

2.3.1 Classful Addressing

2.3.2 Subnet Addressing

2.3.3 Subnet Masks

2.3.4 Variable-Length Subnet Masks (VLSM)

2.3.5 Broadcast Addresses

2.3.6 IPv6 Addresses and Interface Identifiers

## **2.4 CIDR and Aggregation**

2.4.1 Prefixes

2.4.2 Aggregation

## **2.5 Special-Use Addresses**

2.5.1 Addressing IPv4/IPv6 Translators

2.5.2 Multicast Addresses

2.5.3 IPv4 Multicast Addresses

2.5.4 IPv6 Multicast Addresses

2.5.5 Anycast Addresses

## **2.6 Allocation**

# **Table of Contents**

2.6.1 Unicast

2.6.2 Multicast

## **2.7 Unicast Address Assignment**

2.7.1 Single Provider/No Network/Single Address

2.7.2 Single Provider/Single Network/Single Address

2.7.3 Single Provider/Multiple Networks/Multiple Addresses

2.7.4 Multiple Providers/Multiple Networks/Multiple Addresses (Multihoming)

## **2.8 Attacks Involving IP Addresses**

## **2.9 Summary**

## **2.10 References**

# **Chapter 3 Link Layer**

## **3.1 Introduction**

## **3.2 Ethernet and the IEEE 802 LAN/MAN Standards**

3.2.1 The IEEE 802 LAN/MAN Standards

3.2.2 The Ethernet Frame Format

3.2.3 802.1p/q: Virtual LANs and QoS Tagging

3.2.4 802.1AX: Link Aggregation (Formerly 802.3ad)

## **3.3 Full Duplex, Power Save, Autonegotiation, and 802.1X Flow Control**

3.3.1 Duplex Mismatch

3.3.2 Wake-on LAN (WoL), Power Saving, and Magic Packets

3.3.3 Link-Layer Flow Control

## **3.4 Bridges and Switches**

3.4.1 Spanning Tree Protocol (STP)

3.4.2 802.1ak: Multiple Registration Protocol (MRP)

## **3.5 Wireless LANs IEEE 802.11 (Wi-Fi)**

3.5.1 802.11 Frames

3.5.2 Power Save Mode and the Time Sync Function (TSF)

# **Table of Contents**

3.5.3 802.11 Media Access Control

3.5.4 Physical-Layer Details: Rates, Channels, and Frequencies

3.5.5 Wi-Fi Security

3.5.6 Wi-Fi Mesh (802.11s)

## **3.6 Point-to-Point Protocol (PPP)**

3.6.1 Link Control Protocol (LCP)

3.6.2 Multilink PPP (MP)

3.6.3 Compression Control Protocol (CCP)

3.6.4 PPP Authentication

3.6.5 Network Control Protocols (NCPs)

3.6.6 Header Compression

3.6.7 Example

## **3.7 Loopback**

## **3.8 MTU and Path MTU**

## **3.9 Tunneling Basics**

3.9.1 Unidirectional Links

## **3.10 Attacks on the Link Layer**

## **3.11 Summary**

## **3.12 References**

# **Chapter 4 ARP: Address Resolution Protocol**

## **4.1 Introduction**

## **4.2 An Example**

4.2.1 Direct Delivery and ARP

## **4.3 ARP Cache**

## **4.4 ARP Frame Format**

## **4.5 ARP Examples**

4.5.1 Normal Example

4.5.2 ARP Request to a Nonexistent Host

# **Table of Contents**

4.6 ARP Cache Timeout

4.7 Proxy ARP

4.8 Gratuitous ARP and Address Conflict Detection (ACD)

4.9 The arp Command

4.10 Using ARP to Set an Embedded Devices IPv4 Address

4.11 Attacks Involving ARP

4.12 Summary

4.13 References

## **Chapter 5 The Internet Protocol (IP)**

5.1 Introduction

5.2 IPv4 and IPv6 Headers

5.2.1 IP Header Fields

5.2.2 The Internet Checksum

5.2.3 DS Field and ECN (Formerly Called the ToS Byte or IPv6 Traffic Class)

5.2.4 IP Options

5.3 IPv6 Extension Headers

5.3.1 IPv6 Options

5.3.2 Routing Header

5.3.3 Fragment Header

5.4 IP Forwarding

5.4.1 Forwarding Table

5.4.2 IP Forwarding Actions

5.4.3 Examples

5.4.4 Discussion

5.5 Mobile IP

5.5.1 The Basic Model: Bidirectional Tunneling

5.5.2 Route Optimization (RO)

# **Table of Contents**

5.5.3 Discussion

## **5.6 Host Processing of IP Datagrams**

5.6.1 Host Models

5.6.2 Address Selection

## **5.7 Attacks Involving IP**

## **5.8 Summary**

## **5.9 References**

# **Chapter 6 System Configuration: DHCP and Autoconfiguration**

## **6.1 Introduction**

## **6.2 Dynamic Host Configuration Protocol (DHCP)**

6.2.1 Address Pools and Leases

6.2.2 DHCP and BOOTP Message Format

6.2.3 DHCP and BOOTP Options

6.2.4 DHCP Protocol Operation

6.2.5 DHCPv6

6.2.6 Using DHCP with Relays

6.2.7 DHCP Authentication

6.2.8 Reconfigure Extension

6.2.9 Rapid Commit

6.2.10 Location Information (LCI and LoST)

6.2.11 Mobility and Handoff Information (MoS and ANDSF)

6.2.12 DHCP Snooping

## **6.3 Stateless Address Autoconfiguration (SLAAC)**

6.3.1 Dynamic Configuration of IPv4 Link-Local Addresses

6.3.2 IPv6 SLAAC for Link-Local Addresses

## **6.4 DHCP and DNS Interaction**

## **6.5 PPP over Ethernet (PPPoE)**



# **Table of Contents**

6.6 Attacks Involving System Configuration

6.7 Summary

6.8 References

## **Chapter 7 Firewalls and Network Address Translation (NAT)**

7.1 Introduction

7.2 Firewalls

7.2.1 Packet-Filtering Firewalls

7.2.2 Proxy Firewalls

7.3 Network Address Translation (NAT)

7.3.1 Traditional NAT: Basic NAT and NAPT

7.3.2 Address and Port Translation Behavior

7.3.3 Filtering Behavior

7.3.4 Servers behind NATs

7.3.5 Hairpinning and NAT Loopback

7.3.6 NAT Editors

7.3.7 Service Provider NAT (SPNAT) and Service Provider IPv6 Transition

7.4 NAT Traversal

7.4.1 Pinholes and Hole Punching

7.4.2 UNilateral Self-Address Fixing (UNSAF)

7.4.3 Session Traversal Utilities for NAT (STUN)

7.4.4 Traversal Using Relays around NAT (TURN)

7.4.5 Interactive Connectivity Establishment (ICE)

7.5 Configuring Packet-Filtering Firewalls and NATs

7.5.1 Firewall Rules

7.5.2 NAT Rules

7.5.3 Direct Interaction with NATs and Firewalls: UPnP, NAT-PMP, and

# **Table of Contents**

## PCP

### 7.6 NAT for IPv4/IPv6 Coexistence and Transition

#### 7.6.1 Dual-Stack Lite (DS-Lite)

#### 7.6.2 IPv4/IPv6 Translation Using NATs and ALGs

### 7.7 Attacks Involving Firewalls and NATs

### 7.8 Summary

### 7.9 References

## Chapter 8 ICMPv4 and ICMPv6: Internet Control Message Protocol

### 8.1 Introduction

#### 8.1.1 Encapsulation in IPv4 and IPv6

### 8.2 ICMP Messages

#### 8.2.1 ICMPv4 Messages

#### 8.2.2 ICMPv6 Messages

#### 8.2.3 Processing of ICMP Messages

### 8.3 ICMP Error Messages

#### 8.3.1 Extended ICMP and Multipart Messages

#### 8.3.2 Destination Unreachable (ICMPv4 Type 3, ICMPv6 Type 1) and Packet Too Big (ICMPv6 Type 2)

#### 8.3.3 Redirect (ICMPv4 Type 5, ICMPv6 Type 137)

#### 8.3.4 ICMP Time Exceeded (ICMPv4 Type 11, ICMPv6 Type 3)

#### 8.3.5 Parameter Problem (ICMPv4 Type 12, ICMPv6 Type 4)

### 8.4 ICMP Query/Informational Messages

#### 8.4.1 Echo Request/Reply (ping) (ICMPv4 Types 0/8, ICMPv6 Types 129/128)

#### 8.4.2 Router Discovery: Router Solicitation and Advertisement (ICMPv4 Types 9, 10)

#### 8.4.3 Home Agent Address Discovery Request/Reply (ICMPv6 Types

# **Table of Contents**

144/145)

8.4.4 Mobile Prefix Solicitation/Advertisement (ICMPv6 Types  
146/147)

8.4.5 Mobile IPv6 Fast Handover Messages (ICMPv6 Type 154)

8.4.6 Multicast Listener Query/Report/Done (ICMPv6 Types  
130/131/132)

8.4.7 Version 2 Multicast Listener Discovery (MLDv2) (ICMPv6 Type  
143)

8.4.8 Multicast Router Discovery (MRD) (IGMP Types 48/49/50, ICMPv6  
Types 151/152/153)

## **8.5 Neighbor Discovery in IPv6**

8.5.1 ICMPv6 Router Solicitation and Advertisement (ICMPv6 Types  
133, 134)

8.5.2 ICMPv6 Neighbor Solicitation and Advertisement (IMCPv6 Types  
135, 136)

8.5.3 ICMPv6 Inverse Neighbor Discovery Solicitation/Advertisement  
(ICMPv6 Types 141/142)

8.5.4 Neighbor Unreachability Detection (NUD)

8.5.5 Secure Neighbor Discovery (SEND)

8.5.6 ICMPv6 Neighbor Discovery (ND) Options

## **8.6 Translating ICMPv4 and ICMPv6**

8.6.1 Translating ICMPv4 to ICMPv6

8.6.2 Translating ICMPv6 to ICMPv4

## **8.7 Attacks Involving ICMP**

## **8.8 Summary**

## **8.9 References**

# **Chapter 9 Broadcasting and Local Multicasting (IGMP and MLD)**

## **9.1 Introduction**

# **Table of Contents**

## **9.2 Broadcasting**

9.2.1 Using Broadcast Addresses

9.2.2 Sending Broadcast Datagrams

## **9.3 Multicasting**

9.3.1 Converting IP Multicast Addresses to 802 MAC/Ethernet  
Addresses

9.3.2 Examples

9.3.3 Sending Multicast Datagrams

9.3.4 Receiving Multicast Datagrams

9.3.5 Host Address Filtering

## **9.4 The Internet Group Management Protocol (IGMP) and Multicast Listener Discovery Protocol (MLD)**

9.4.1 IGMP and MLD Processing by Group Members (Group Member  
Part)

9.4.2 IGMP and MLD Processing by Multicast Routers (Multicast  
Router Part)

9.4.3 Examples

9.4.4 Lightweight IGMPv3 and MLDv2

9.4.5 IGMP and MLD Robustness

9.4.6 IGMP and MLD Counters and Variables

9.4.7 IGMP and MLD Snooping

## **9.5 Attacks Involving IGMP and MLD**

## **9.6 Summary**

## **9.7 References**

# **Chapter 10 User Datagram Protocol (UDP) and IP Fragmentation**

## **10.1 Introduction**

## **10.2 UDP Header**

# **Table of Contents**

10.3 UDP Checksum

10.4 Examples

10.5 UDP and IPv6

10.5.1 Teredo: Tunneling IPv6 through IPv4 Networks

10.6 UDP-Lite

10.7 IP Fragmentation

10.7.1 Example: UDP/IPv4 Fragmentation

10.7.2 Reassembly Timeout

10.8 Path MTU Discovery with UDP

10.8.1 Example

10.9 Interaction between IP Fragmentation and ARP/ND

10.10 Maximum UDP Datagram Size

10.10.1 Implementation Limitations

10.10.2 Datagram Truncation

10.11 UDP Server Design

10.11.1 IP Addresses and UDP Port Numbers

10.11.2 Restricting Local IP Addresses

10.11.3 Using Multiple Addresses

10.11.4 Restricting Foreign IP Address

10.11.5 Using Multiple Servers per Port

10.11.6 Spanning Address Families: IPv4 and IPv6

10.11.7 Lack of Flow and Congestion Control

10.12 Translating UDP/IPv4 and UDP/IPv6 Datagrams

10.13 UDP in the Internet

10.14 Attacks Involving UDP and IP Fragmentation

10.15 Summary

10.16 References

Chapter 11 Name Resolution and the Domain Name System

# **Table of Contents**

## **(DNS)**

### **11.1 Introduction**

### **11.2 The DNS Name Space**

#### **11.2.1 DNS Naming Syntax**

### **11.3 Name Servers and Zones**

### **11.4 Caching**

### **11.5 The DNS Protocol**

#### **11.5.1 DNS Message Format**

#### **11.5.2 The DNS Extension Format (EDNS0)**

#### **11.5.3 UDP or TCP**

#### **11.5.4 Question (Query) and Zone Section Format**

#### **11.5.5 Answer, Authority, and Additional Information Section Formats**

#### **11.5.6 Resource Record Types**

#### **11.5.7 Dynamic Updates (DNS UPDATE)**

#### **11.5.8 Zone Transfers and DNS NOTIFY**

### **11.6 Sort Lists, Round-Robin, and Split DNS**

### **11.7 Open DNS Servers and DynDNS**

### **11.8 Transparency and Extensibility**

### **11.9 Translating DNS from IPv4 to IPv6 (DNS64)**

### **11.10 LLMNR and mDNS**

### **11.11 LDAP**

### **11.12 Attacks on the DNS**

### **11.13 Summary**

### **11.14 References**

## **Chapter 12 TCP: The Transmission Control Protocol (Preliminaries)**

### **12.1 Introduction**

# **Table of Contents**

12.1.1 ARQ and Retransmission

12.1.2 Windows of Packets and Sliding Windows

12.1.3 Variable Windows: Flow Control and Congestion Control

12.1.4 Setting the Retransmission Timeout

## **12.2 Introduction to TCP**

12.2.1 The TCP Service Model

12.2.2 Reliability in TCP

## **12.3 TCP Header and Encapsulation**

## **12.4 Summary**

## **12.5 References**

# **Chapter 13 TCP Connection Management**

## **13.1 Introduction**

## **13.2 TCP Connection Establishment and Termination**

13.2.1 TCP Half-Close

13.2.2 Simultaneous Open and Close

13.2.3 Initial Sequence Number (ISN)

13.2.4 Example

13.2.5 Timeout of Connection Establishment

13.2.6 Connections and Translators

## **13.3 TCP Options**

13.3.1 Maximum Segment Size (MSS) Option

13.3.2 Selective Acknowledgment (SACK) Options

13.3.3 Window Scale (WSCALE or WSOPT) Option

13.3.4 Timestamps Option and Protection against Wrapped Sequence  
Numbers (PAWS)

13.3.5 User Timeout (UTO) Option

13.3.6 Authentication Option (TCP-AO)

## **13.4 Path MTU Discovery with TCP**

# **Table of Contents**

13.4.1 Example

## **13.5 TCP State Transitions**

13.5.1 TCP State Transition Diagram

13.5.2 TIME\_WAIT (2MSL Wait) State

13.5.3 Quiet Time Concept

13.5.4 FIN\_WAIT\_2 State

13.5.5 Simultaneous Open and Close Transitions

## **13.6 Reset Segments**

13.6.1 Connection Request to Nonexistent Port

13.6.2 Aborting a Connection

13.6.3 Half-Open Connections

13.6.4 TIME-WAIT Assassination (TWA)

## **13.7 TCP Server Operation**

13.7.1 TCP Port Numbers

13.7.2 Restricting Local IP Addresses

13.7.3 Restricting Foreign Endpoints

13.7.4 Incoming Connection Queue

## **13.8 Attacks Involving TCP Connection Management**

## **13.9 Summary**

## **13.10 References**

# **Chapter 14 TCP Timeout and Retransmission**

## **14.1 Introduction**

## **14.2 Simple Timeout and Retransmission Example**

## **14.3 Setting the Retransmission Timeout (RTO)**

14.3.1 The Classic Method

14.3.2 The Standard Method

14.3.3 The Linux Method

14.3.4 RTT Estimator Behaviors



# **Table of Contents**

14.3.5 RTTM Robustness to Loss and Reordering

## **14.4 Timer-Based Retransmission**

14.4.1 Example

## **14.5 Fast Retransmit**

14.5.1 Example

## **14.6 Retransmission with Selective Acknowledgments**

14.6.1 SACK Receiver Behavior

14.6.2 SACK Sender Behavior

14.6.3 Example

## **14.7 Spurious Timeouts and Retransmissions**

14.7.1 Duplicate SACK (DSACK) Extension

14.7.2 The Eifel Detection Algorithm

14.7.3 Forward-RTO Recovery (F-RTO)

14.7.4 The Eifel Response Algorithm

## **14.8 Packet Reordering and Duplication**

14.8.1 Reordering

14.8.2 Duplication

## **14.9 Destination Metrics**

## **14.10 Repacketization**

## **14.11 Attacks Involving TCP Retransmission**

## **14.12 Summary**

## **14.13 References**

# **Chapter 15 TCP Data Flow and Window Management**

## **15.1 Introduction**

## **15.2 Interactive Communication**

## **15.3 Delayed Acknowledgments**

## **15.4 Nagle Algorithm**

# **Table of Contents**

15.4.1 Delayed ACK and Nagle Algorithm Interaction

15.4.2 Disabling the Nagle Algorithm

## **15.5 Flow Control and Window Management**

15.5.1 Sliding Windows

15.5.2 Zero Windows and the TCP Persist Timer

15.5.3 Silly Window Syndrome (SWS)

15.5.4 Large Buffers and Auto-Tuning

## **15.6 Urgent Mechanism**

15.6.1 Example

## **15.7 Attacks Involving Window Management**

## **15.8 Summary**

## **15.9 References**

# **Chapter 16 TCP Congestion Control**

## **16.1 Introduction**

16.1.1 Detection of Congestion in TCP

16.1.2 Slowing Down a TCP Sender

## **16.2 The Classic Algorithms**

16.2.1 Slow Start

16.2.2 Congestion Avoidance

16.2.3 Selecting between Slow Start and Congestion Avoidance

16.2.4 Tahoe, Reno, and Fast Recovery

16.2.5 Standard TCP

## **16.3 Evolution of the Standard Algorithms**

16.3.1 NewReno

16.3.2 TCP Congestion Control with SACK

16.3.3 Forward Acknowledgment (FACK) and Rate Halving

16.3.4 Limited Transmit

16.3.5 Congestion Window Validation (CWV)

# **Table of Contents**

## 16.4 Handling Spurious RTOs the Eifel Response Algorithm

## 16.5 An Extended Example

### 16.5.1 Slow Start Behavior

### 16.5.2 Sender Pause and Local Congestion (Event 1)

### 16.5.3 Stretch ACKs and Recovery from Local Congestion

### 16.5.4 Fast Retransmission and SACK Recovery (Event 2)

### 16.5.5 Additional Local Congestion and Fast Retransmit Events

### 16.5.6 Timeouts, Retransmissions, and Undoing cwnd Changes

### 16.5.7 Connection Completion

## 16.6 Sharing Congestion State

## 16.7 TCP Friendliness

## 16.8 TCP in High-Speed Environments

### 16.8.1 HighSpeed TCP (HSTCP) and Limited Slow Start

### 16.8.2 Binary Increase Congestion Control (BIC and CUBIC)

## 16.9 Delay-Based Congestion Control

### 16.9.1 Vegas

### 16.9.2 FAST

### 16.9.3 TCP Westwood and Westwood+

### 16.9.4 Compound TCP

## 16.10 Buffer Bloat

## 16.11 Active Queue Management and ECN

## 16.12 Attacks Involving TCP Congestion Control

## 16.13 Summary

## 16.14 References

## Chapter 17 TCP Keepalive

### 17.1 Introduction

### 17.2 Description

# **Table of Contents**

17.2.1 Keepalive Examples

17.3 Attacks Involving TCP Keepalives

17.4 Summary

17.5 References

## **Chapter 18 Security: EAP, IPsec, TLS, DNSSEC, and DKIM**

18.1 Introduction

18.2 Basic Principles of Information Security

18.3 Threats to Network Communication

18.4 Basic Cryptography and Security Mechanisms

18.4.1 Cryptosystems

18.4.2 Rivest, Shamir, and Adleman (RSA) Public Key Cryptography

18.4.3 Diffie-Hellman-Merkle Key Agreement (aka Diffie-Hellman or DH)

18.4.4 Signcryption and Elliptic Curve Cryptography (ECC)

18.4.5 Key Derivation and Perfect Forward Secrecy (PFS)

18.4.6 Pseudorandom Numbers, Generators, and Function Families

18.4.7 Nonces and Salt

18.4.8 Cryptographic Hash Functions and Message Digests

18.4.9 Message Authentication Codes (MACs, HMAC, CMAC, and GMAC)

18.4.10 Cryptographic Suites and Cipher Suites

18.5 Certificates, Certificate Authorities (CAs), and PKIs

18.5.1 Public Key Certificates, Certificate Authorities, and X.509

18.5.2 Validating and Revoking Certificates

18.5.3 Attribute Certificates

18.6 TCP/IP Security Protocols and Layering

18.7 Network Access Control: 802.1X, 802.1AE, EAP, and PANA

18.7.1 EAP Methods and Key Derivation

# **Table of Contents**

18.7.2 The EAP Re-authentication Protocol (ERP)

18.7.3 Protocol for Carrying Authentication for Network Access  
(PANA)

## **18.8 Layer 3 IP Security (IPsec)**

18.8.1 Internet Key Exchange (IKEv2) Protocol

18.8.2 Authentication Header (AH)

18.8.3 Encapsulating Security Payload (ESP)

18.8.4 Multicast

18.8.5 L2TP/IPsec

18.8.6 IPsec NAT Traversal

18.8.7 Example

## **18.9 Transport Layer Security (TLS and DTLS)**

18.9.1 TLS 1.2

18.9.2 TLS with Datagrams (DTLS)

## **18.10 DNS Security (DNSSEC)**

18.10.1 DNSSEC Resource Records

18.10.2 DNSSEC Operation

18.10.3 Transaction Authentication (TSIG, TKEY, and SIG(0))

18.10.4 DNSSEC with DNS64

## **18.11 DomainKeys Identified Mail (DKIM)**

18.11.1 DKIM Signatures

18.11.2 Example

## **18.12 Attacks on Security Protocols**

## **18.13 Summary**

## **18.14 References**

## **Glossary of Acronyms**

A

B

# **Table of Contents**

C

D

E

F

G

H

I

K

L

M

N

O

P

Q

R

S

T

U

V

W

X

Z

Index