

Overlay Networks



Toward Information Networking

Sasu Tarkoma



CRC Press
Taylor & Francis Group
AN AUERBACH BOOK

Overlay Networks

Toward Information Networking

Sasu Tarkoma



CRC Press

Taylor & Francis Group

Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
AN AUERBACH BOOK

Auerbach Publications
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2010 by Taylor and Francis Group, LLC
Auerbach Publications is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number: 978-1-4398-1371-3 (Hardback)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (<http://www.copyright.com/>) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Library of Congress Cataloging-in-Publication Data

Tarkoma, Sasu.

Overlay networks : toward information networking / Sasu Tarkoma.
p. cm.

Includes bibliographical references and index.

ISBN 978-1-4398-1371-3 (hardcover : alk. paper)

1. Information networks. 2. Peer-to-peer architecture (Computer networks) 3. Electronic data processing--Distributed processing. 4. Multimedia communications. I. Title.

TK5105.5.T366 2010
004.6'52--dc22

2009046412

Visit the Taylor & Francis Web site at
<http://www.taylorandfrancis.com>

and the Auerbach Web site at
<http://www.auerbach-publications.com>

Overlay Networks

Toward Information Networking

OTHER TELECOMMUNICATIONS BOOKS FROM AUERBACH

Broadband Mobile Multimedia: Techniques and Applications

Yan Zhang, Shiwen Mao, Laurence T. Yang,
and Thomas M Chen
ISBN: 978-1-4200-5184-1

Carrier Ethernet: Providing the Need for Speed

Gilbert Held
ISBN: 978-1-4200-6039-3

Cognitive Radio Networks

Yang Xiao and Fei Hu
ISBN: 978-1-4200-6420-9

Contemporary Coding Techniques and Applications for Mobile Communications

Onur Osman and Osman Nuri Ucan
ISBN: 978-1-4200-5461-3

Converging NGN Wireline and Mobile 3G Networks with IMS: Converging NGN and 3G Mobile

Rebecca Copeland
ISBN: 978-0-8493-9250-4

Cooperative Wireless Communications

Yan Zhang, Hsiao-Hwa Chen, and Mohsen Guizani
ISBN: 978-1-4200-6469-8

Data Scheduling and Transmission Strategies in Asymmetric Telecommunication Environments

Abhishek Roy and Navrati Saxena
ISBN: 978-1-4200-4655-7

Encyclopedia of Wireless and Mobile Communications

Borko Furht
ISBN: 978-1-4200-4326-6

IMS: A New Model for Blending Applications

Mark Wuthnow, Jerry Shih, and Matthew Stafford
ISBN: 978-1-4200-9285-1

The Internet of Things: From RFID to the Next-Generation Pervasive Networked Systems

Lu Yan, Yan Zhang, Laurence T. Yang,
and Huansheng Ning
ISBN: 978-1-4200-5281-7

Introduction to Communications Technologies: A Guide for Non-Engineers, Second Edition

Stephan Jones, Ron Kovac, and Frank M. Groom
ISBN: 978-1-4200-4684-7

Long Term Evolution: 3GPP LTE Radio and Cellular Technology

Borko Furht and Syed A. Ahson
ISBN: 978-1-4200-7210-5

MEMS and Nanotechnology-Based Sensors and Devices for Communications, Medical and Aerospace Applications

A. R. Jha
ISBN: 978-0-8493-8069-3

Millimeter Wave Technology in Wireless PAN, LAN, and MAN

Shao-Qiu Xiao and Ming-Tuo Zhou
ISBN: 978-0-8493-8227-7

Mobile Telemedicine: A Computing and Networking Perspective

Yang Xiao and Hui Chen
ISBN: 978-1-4200-6046-1

Optical Wireless Communications: IR for Wireless Connectivity

Roberto Ramirez-Iniguez, Sevia M. Idrus,
and Ziran Sun
ISBN: 978-0-8493-7209-4

Satellite Systems Engineering in an IPv6 Environment

Daniel Minoli
ISBN: 978-1-4200-7868-8

Security in RFID and Sensor Networks

Yan Zhang and Paris Kitsos
ISBN: 978-1-4200-6839-9

Security of Mobile Communications

Noureddine Boudriga
ISBN: 978-0-8493-7941-3

Unlicensed Mobile Access Technology: Protocols, Architectures, Security, Standards and Applications

Yan Zhang, Laurence T. Yang, and Jianhua Ma
ISBN: 978-1-4200-5537-5

Value-Added Services for Next Generation Networks

Thierry Van de Velde
ISBN: 978-0-8493-7318-3

Vehicular Networks: Techniques, Standards, and Applications

Hassnaa Moustafa and Yan Zhang
ISBN: 978-1-4200-8571-6

WiMAX Network Planning and Optimization

Yan Zhang
ISBN: 978-1-4200-6662-3

Wireless Quality of Service: Techniques, Standards, and Applications

Maode Ma and Mieso K. Denko
ISBN: 978-1-4200-5130-8

AUERBACH PUBLICATIONS

www.auerbach-publications.com

To Order Call: 1-800-272-7737 • Fax: 1-800-374-3401

E-mail: orders@crcpress.com

Preface

Data and media delivery have become hugely popular on the Internet, with well over 1 billion Internet users. Therefore scalable and flexible information dissemination solutions are needed. Much of the current development pertaining to services and service delivery happens above the basic network layer and the TCP/IP protocol suite because of the need to be able to rapidly develop and deploy them.

In recent years, various kinds of overlay networking technologies have emerged as an active area of research and development. Overlay systems, especially *peer-to-peer* systems, are technologies that can solve problems in massive information distribution and processing tasks. The key aim of many of these technologies is to be able to offer deployable solution for processing and distributing vast amounts of information, typically petabytes and more, while at the same time keeping the scaling costs low.

The aim of this book is to present the state of the art in overlay technologies, examine the key structures and algorithms used in overlay networks, and discuss their applications. Overlay networks have been a very active area of research and development during the last 10 years, and a substantial amount of scientific literature has formed around this topic.

This book has been inspired by the teaching notes and articles of the author in content-based routing. The book is designed not only as a reference for overlay technologies, but also as a textbook for a course in distributed overlay technologies and information networking at the graduate level.

About the Author

Sasu Tarkoma received his M.Sc. and Ph.D. degrees in Computer Science from the University of Helsinki, Department of Computer Science. He is currently professor at Helsinki University of Technology, Department of Computer Science and Engineering. He has been recently appointed as full professor at University of Helsinki, Department of Computer Science. He has managed and participated in national and international research projects at the University of Helsinki, Helsinki University of Technology, and Helsinki Institute for Information Technology (HIIT). He has worked in the IT industry as a consultant and chief system architect, and he is principal member of research staff at Nokia Research Center. He has over 100 publications, and has also contributed to several books on mobile middleware.

Ms. Nelli Tarkoma produced most of the diagrams used in this book.

Contents

Preface	xi
About the Author	xiii
1 Introduction	1
1.1 Overview	1
1.2 Overlay Technology	2
1.3 Applications	8
1.4 Properties of Data	8
1.5 Structure of the Book	10
2 Network Technologies	13
2.1 Networking	13
2.2 Firewalls and NATs	15
2.3 Naming	17
2.4 Addressing	18
2.5 Routing	19
2.5.1 Overview	19
2.5.2 Interdomain	20
2.5.3 Border Gateway Protocol	21
2.5.4 Current Challenges	21
2.5.5 Compact Routing	22
2.6 Multicast	22
2.6.1 Network-layer Multicast	23
2.6.2 Application-layer Multicast	24
2.6.3 Chaining TCP Connections for Multicast	25
2.7 Network Coordinates	25
2.7.1 Vivaldi Centralized Algorithm	26
2.7.2 Vivaldi Distributed Algorithm	26
2.7.3 Applications	26
2.7.4 Triangle Inequality Violation	27
2.8 Network Metrics	28
2.8.1 Routing Algorithm Invariants	28
2.8.2 Convergence	28
2.8.3 Shortest Path	28
2.8.4 Routing Table Size and Stretch	29
2.8.5 Forwarding Load	29
2.8.6 Churn	30
2.8.7 Other Metrics	30
3 Properties of Networks and Data	33
3.1 Data on the Internet	33
3.1.1 Video Delivery	33
3.1.2 P2P Traffic	35

3.1.3	Trends in Networking	35
3.2	Zipf's Law	36
3.2.1	Overview	36
3.2.2	Zipf's Law and the Internet	37
3.2.3	Implications for P2P	37
3.3	Scale-free Networks	37
3.4	Robustness	39
3.5	Small Worlds	40
4	Unstructured Overlays	43
4.1	Overview	43
4.2	Early Systems	44
4.3	Locating Data	44
4.4	Napster	45
4.5	Gnutella	46
4.5.1	Overview	46
4.5.2	Searching the Network	48
4.5.3	Efficient Keyword Lists	49
4.6	Skype	50
4.7	BitTorrent	50
4.7.1	Torrents and Swarms	53
4.7.2	Networking	54
4.7.3	Choking Mechanism	54
4.7.4	Antisnubbing	55
4.7.5	End Game	55
4.7.6	Trackerless Operation	55
4.7.7	BitTorrent Vulnerabilities	56
4.7.8	Service Capacity	56
4.7.9	Fluid Models for Performance Evaluation	57
4.8	Cross-ISP BitTorrent	58
4.9	Freenet	60
4.9.1	Overview	60
4.9.2	Bootstrapping	62
4.9.3	Identifier keys	62
4.9.4	Key-based Routing	64
4.9.5	Indirect Files	65
4.9.6	API	65
4.9.7	Security	66
4.10	Comparison	67
5	Foundations of Structured Overlays	71
5.1	Overview	71
5.2	Geometries	72
5.2.1	Trees	73
5.2.2	Hypercubes and Tori	73
5.2.3	Butterflies	74
5.2.4	de Bruijn graph	74
5.2.5	Rings	75
5.2.6	XOR Geometry	76
5.2.7	Summary	76
5.3	Consistent Hashing	77

5.4	Distributed Data Structures for Clusters	78
5.4.1	Linear Hashing	78
5.4.2	SDDS Taxonomy	79
5.4.3	LH* Overview	80
5.4.4	Ninja	82
6	Distributed Hash Tables	85
6.1	Overview	85
6.2	APIs	86
6.3	Plaxton's Algorithm	87
6.3.1	Routing	87
6.3.2	Performance	88
6.4	Chord	89
6.4.1	Joining the Network	90
6.4.2	Leaving the Network	90
6.4.3	Routing	91
6.4.4	Performance	92
6.5	Pastry	92
6.5.1	Joining and Leaving the Network	93
6.5.2	Routing	93
6.5.3	Performance	95
6.5.4	Bamboo	96
6.6	Koorde	96
6.6.1	Routing	96
6.6.2	Performance	97
6.7	Tapestry	97
6.7.1	Joining and Leaving the Network	98
6.7.2	Routing	99
6.7.3	Performance	100
6.8	Kademlia	101
6.8.1	Joining and Leaving the Network	101
6.8.2	Routing	101
6.8.3	Performance	102
6.9	Content Addressable Network	103
6.9.1	Joining the Network	103
6.9.2	Leaving the Network	104
6.9.3	Routing	105
6.9.4	Performance	105
6.10	Viceroy	106
6.10.1	Joining the Network	106
6.10.2	Leaving the Network	107
6.10.3	Routing	107
6.10.4	Performance	107
6.11	Skip Graph	108
6.12	Comparison	109
6.12.1	Geometries	110
6.12.2	Routing Function	110
6.12.3	Churn	110
6.12.4	Asymptotic Trade-offs	110
6.12.5	Network Proximity	112
6.12.6	Adding Hierarchy to DHTs	112

6.12.7	Experimenting with Overlays	113
6.12.8	Criticism	114
7	Probabilistic Algorithms	115
7.1	Overview of Bloom Filters	115
7.2	Bloom Filters	116
7.2.1	False Positive Probability	118
7.2.2	Operations	119
7.2.3	d-left Counting Bloom Filter	120
7.2.4	Compressed Bloom Filter	121
7.2.5	Counting Bloom Filters	121
7.2.6	Hierarchical Bloom Filters	123
7.2.7	Spectral Bloom Filters	124
7.2.8	Bloomier Filters	124
7.2.9	Approximate State Machines	125
7.2.10	Perfect Hashing Scheme	125
7.2.11	Summary	125
7.3	Bloom Filters in Distributed Computing	127
7.3.1	Caching	127
7.3.2	P2P Networks	128
7.3.3	Packet Routing and Forwarding	128
7.3.4	Measurement	129
7.4	Gossip Algorithms	130
7.4.1	Overview	130
7.4.2	Design Considerations	132
7.4.3	Basic Models	132
7.4.4	Basic Shuffling	133
7.4.5	Enhanced Shuffling	135
7.4.6	Flow Control and Fairness	135
7.4.7	Gossip for Structured Overlays	136
8	Content-based Networking and Publish/Subscribe	137
8.1	Overview	137
8.2	DHT-based Data-centric Communications	138
8.2.1	Scribe	138
8.2.2	Bayeux	139
8.2.3	SplitStream	139
8.2.4	Overcast	140
8.2.5	Meghdoot	141
8.2.6	MEDYM	141
8.2.7	Internet Indirection Infrastructure	141
8.2.8	Data-oriented Network Architecture	141
8.2.9	Semantic Search	142
8.2.10	Distributed Segment Tree	142
8.2.11	Semantic Queries	143
8.3	Content-based Routing	144
8.4	Router Configurations	145
8.4.1	Basic Configuration	145
8.4.2	Structured DHT-based Overlays	146
8.4.3	Interest Propagation	147
8.5	Siena and Routing Structures	148

8.5.1	Routing Blocks	148
8.5.2	Definitions	150
8.5.3	Siena Filters Poset	150
8.5.4	Advertisements	152
8.5.5	Poset-derived Forest.....	152
8.5.6	Filter Merging.....	154
8.6	Hermes	156
8.7	Formal Specification of Content-based Routing Systems	158
8.7.1	Valid Routing Configuration.....	158
8.7.2	Weakly Valid Routing Configuration.....	159
8.7.3	Mobility-Safety	159
8.8	Pub/sub Mobility	160
9	Security	165
9.1	Overview	165
9.2	Attacks and Threats	166
9.2.1	Worms	166
9.2.2	Sybil Attack	166
9.2.3	Eclipse Attack.....	166
9.2.4	File Poisoning	167
9.2.5	Man-in-the-Middle Attack.....	168
9.2.6	DoS Attack	168
9.3	Securing Data	169
9.3.1	Self-Certifying Data	169
9.3.2	Merkle Trees	170
9.3.3	Information Dispersal	171
9.3.4	Secret-sharing Schemes.....	171
9.3.5	Smartcards for Bootstrapping Trust.....	171
9.3.6	Distributed Steganographic File Systems.....	172
9.3.7	Erasure Coding	173
9.3.8	Censorship Resistance	173
9.4	Security Issues in P2P Networks.....	174
9.4.1	Overview	174
9.4.2	Insider Attacks	176
9.4.3	Outsider Attacks	176
9.4.4	SybilGuard	177
9.4.5	Reputation Management with EigenTrust.....	178
9.5	Anonymous Routing	180
9.5.1	Mixes	180
9.5.2	Onion Routing	181
9.5.3	Tor	181
9.5.4	P2P Anonymization System	182
9.5.5	Censorship-resistant Lookup: Achord	184
9.5.6	Crowds	184
9.5.7	Hordes	184
9.5.8	Mist.....	185
9.6	Security Issues in Pub/Sub Networks	186
9.6.1	Hermes	186
9.6.2	EventGuard	187
9.6.3	QUIP	187

- 10 Applications**.....189
 - 10.1 Amazon Dynamo189
 - 10.1.1 Architecture.....191
 - 10.1.2 Ring Membership.....193
 - 10.1.3 Partitioning Algorithm.....193
 - 10.1.4 Replication.....194
 - 10.1.5 Data Versioning.....194
 - 10.1.6 Vector Clocks195
 - 10.1.7 Coping with Failures.....196
 - 10.2 Overlay Video Delivery197
 - 10.2.1 Live Streaming.....197
 - 10.2.2 Video-on-Demand198
 - 10.3 SIP and P2PSIP200
 - 10.4 CDN Solutions.....203
 - 10.4.1 Overview203
 - 10.4.2 Akamai207
 - 10.4.3 Limelight208
 - 10.4.4 Coral.....208
 - 10.4.5 Comparison.....211
- 11 Conclusions**.....213
- References**.....217
- Index**.....235

1

Introduction

1.1 Overview

In recent years, various kinds of overlay networking technologies have emerged as an active area of research and development. Overlay systems, especially *peer-to-peer* (P2P) systems, are technologies that can solve problems in massive information distribution and processing tasks. The key aim of many of these technologies is to be able to offer deployable solution for processing and distributing vast amounts of information, typically petabytes and more, while at the same time keeping the scaling costs low.

Data and media delivery have become hugely popular on the Internet. Currently there are over 1.4 billion Internet users, well over 3 billion mobile phones, and 4 billion mobile subscriptions. By 2000 the Google index reached the 1 billion indexed web resources mark, and by 2008 it reached the trillion mark.

Multimedia content, especially videos, are paving the way for truly versatile network services that both compete with and extend existing broadcast-based medias. As a consequence, new kinds of social collaboration and advertisement mechanisms are being introduced both in the fixed Internet and also in the mobile world. This trend is heightened by the ubiquitous nature of digital cameras. Indeed, this has created a lot of interest in community-based services, in which users create their own content and make it available to others.

These developments have had a profound impact on network requirements and performance. Video delivery has become one of the recent services on the Web with the advent of YouTube [67] and other social media Web sites. Moreover, the network impact is heightened by various P2P services. Estimates of P2P share of network traffic range from 50% to 70%. Cisco's latest traffic forecast for 2009–2013 indicates that annual global IP traffic will reach 667 exabytes in 2013, two-thirds of a zettabyte [79]. An exabyte (EB) is an SI unit of information, and 1 EB equals 10^{18} bytes. Exabyte is followed by the zettabyte (1 Z = 10^{21}) and yottabyte (1 Y = 10^{24}). The traffic is expected to increase some 40% each year. Much of this increase comes from the delivery of video data in various forms. Video delivery on the Internet will see a huge increase, and the volume of video delivery in 2013 is expected to be 700 times the capacity of the US Internet backbone in 2000. The study anticipates that video traffic will account for 91% of all consumer traffic in 2013.

According to the study, P2P traffic will continue to grow but will become a smaller component of Internet traffic in terms of its current share. The current P2P systems in 2009 are transferring 3.3 EB data per month. The recent study indicates that the P2P share of consumer Internet traffic will drop to 20% by 2013, down from the current 50% (at the end of 2008). Even though the P2P share may drop, most video delivery solutions, accounting for much of the traffic increase, will utilize overlay technologies, which makes this area crucial for ensuring efficient and scalable services.

A P2P network consists of nodes that cooperate in order to provide services to each other. A pure P2P network consists of equal peers that are simultaneously clients and servers. The P2P model differs from the *client-server* model, where clients access services provided by logically centralized servers.

To date, P2P delivery has not been successfully combined with browser-based operation and media sites such as YouTube. Nevertheless, a number of businesses have realized the importance of scalable data delivery. For example, the game company Blizzard uses P2P technology to distribute patches for the *World of Warcraft* game. Given the heavy use of network, P2P protocols such as BitTorrent offer to reduce network load by peer-assisted data delivery. This means that peer users cooperate to transfer large files over the network.

1.2 Overlay Technology

Data structures and algorithms are central for today's data communications. We may consider circuit switching technology as an example of how information processing algorithms are vital for products and how innovation changes markets. Early telephone systems were based on manual circuit switching. Everything was done using human hands. Later systems used electromechanical devices to connect calls, but they required laborious preconfiguration of telephone numbers and had limited scalability. Modern digital circuit switching algorithms evolved from these older semiautomatic systems and optimize the number of connections in a switch. The nonblocking minimal spanning tree algorithm enabled the optimization of these automatic switches. Any algorithm used to connect millions of calls must be proven to be correct and efficient. The latest development changes the fundamentals of telephone switching, because information is forwarded as packets on a hop-by-hop basis and not via preestablished physical circuits. Today, this complex machinery enables end-to-end connectivity irrespective of time and location.

Data structures are at the heart of the Internet. Network-level routers use efficient algorithms for matching data packets to outgoing interfaces based on prefixes. Internet backbone routers have to manage 200,000 routes and more in order to route packets between systems. The matching algorithms include *suffix trees* and *ternary content addressable memories (TCAMs)* [268], which have to balance between matching efficiency and router memory. Therefore, just as with telephone switches, optimization plays a major role in the development of routers and routing systems.

The current generation of networks is being developed on top of TCP/IPs network-layer (layer 3 in the *open systems interconnection (OSI)* stack). These so-called overlay networks come in various shapes and forms. Overlays make many implementation issues easier, because network-level routers do not need to be changed. In many ways, overlay networks represent a fundamental paradigm shift compared to older technologies such as circuit switching and hierarchical routing.

Overlay networks are useful both in control and content plane scenarios. This division of traffic into control and content is typical of current telecommunications solutions such as the *session initiation protocol (SIP)*; however, this division does not exist on the current Internet as such. As control plane elements, overlays can be used to route control messages and connect different entities. As content plane elements, they can participate in data forwarding and dissemination.

An *overlay network* is a network that is built on top of an existing network. The overlay therefore relies on the so-called *underlay* network for basic networking functions, namely routing and forwarding. Today, most overlay networks are built in the application layer on top of the TCP/IP networking suite. Overlay technologies can be used to overcome some of the limitations of the underlay, at the same time offering new routing and forwarding features without changing the routers. The nodes in an overlay network are connected via logical links that can span many physical links. A link between two overlay nodes may take several hops in the underlying network.

An overlay network therefore consists of a set of distributed nodes, typically client devices or servers, that are deployed on the Internet. The nodes are expected to meet the following requirements:

1. Support the execution of one or more distributed applications by providing infrastructure for them.
2. Participate in and support high-level routing and forwarding tasks. The overlay is expected to provide data-forwarding capabilities that are different from those that are part of the basic Internet.
3. Deploy across the Internet in such a way that third parties can participate in the organization and operation of the overlay network.

Figure 1.1 presents a layered view to overlay networks. The view starts from the underlay, the network that offers the basic primitives of sending and receiving messages (packets). The two obvious choices today are UDP and TCP as the transport layer protocols. TCP is favored due to its connection-oriented nature, congestion control, and reliability.

After the underlay layer, we have the custom routing, forwarding, rendezvous, and discovery functions of the overlay architecture. Routing pertains to the process of building and maintaining routing tables. Forwarding is the process of sending messages toward their destination, and rendezvous is a function that is used to resolve issues regarding some identifier or node—for example, by offering indirection support in the case of mobility. Discovery is an integral part of this layer and is needed to populate the routing table by discovering both physically and logically nearby neighbors.

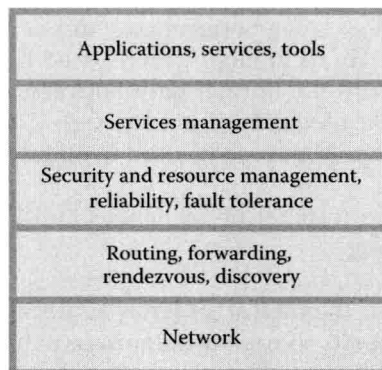


FIGURE 1.1
Layered view to overlay networks.

The next layer introduces additional functions, such as security and resource management, reliability support, and fault tolerance. These are typically built on top of the basic overlay functions mentioned above. Security pertains to the way node identities are assigned and controlled, and messages and packets are secured. Security encompasses multiple protocol layers and is responsible for ensuring that peers can maintain sufficient level of trust toward the system. Resource management is about taking content demand and supply into account and ensuring that certain performance and reliability requirements are met. For example, relevant issues are data placement and replication rate. Data replication is also a basic mechanism for ensuring fault-tolerance. If one node fails, another can take its place and, given that the data was replicated, there is no loss of information.

Above this layer, we have the services management for both monitoring and controlling service lifecycles. When a service is deployed on top of an overlay, there need to be functions for administering it and controlling various issues such as administrative boundaries, and data replication and access control policies.

Finally, in the topmost layer we have the actual applications and services that are executed on top of the layered overlay architecture. The applications rely on the overlay architecture for scalable and resilient data discovery and exchange.

An overlay network offers a number of advantages over both centralized solutions and solutions that introduce changes in routers. These include the following three key advantages:

Incremental deployment: Overlay networks do not require changes to the existing routers. This means that an overlay network can be grown node by node, and with more nodes it is possible to both monitor and control routing paths across the Internet from one overlay node to another. An overlay network can be built based on standard network protocols and existing APIs—for example, the Sockets API of the TCP/IP protocol stack.

Adaptable: The overlay algorithm can utilize a number of metrics when making routing and forwarding decisions. Thus the overlay can take application-specific concerns into account that are not currently offered by the Internet infrastructure. Key metrics include latency, bandwidth, and security.

Robust: An overlay network is robust to node and network failures due to its adaptable nature. With a sufficient number of nodes in the overlay, the network may be able to offer multiple independent (router-disjoint) paths to the same destination. At best, overlay networks are able to route around faults.

The designers of an early overlay system called *resilient overlay network (RON)* [361] used the idea of alternative paths to improve performance and to route around network faults. Figure 1.2 illustrates how overlay technology can be used to route around faults. In this example, there is a problem with the normal path between A and B across the Internet. Now, the overlay can use a so-called *detour path* through C to send traffic to B. This will result in some networking overhead but can be used to maintain communications between A and B.

Overlay networks face also a number of challenges and limitations. The three central challenges include the following:

- **The real world:** In practice, the typical underlay protocol, IP, does not provide universal end-to-end connectivity due to the ubiquitous nature of firewalls and *network address translation (NAT)* devices. This means that special solutions are needed to overcome reachability issues. In addition, many overlay networks are oblivious to the current organizational and management structures that exist in applications