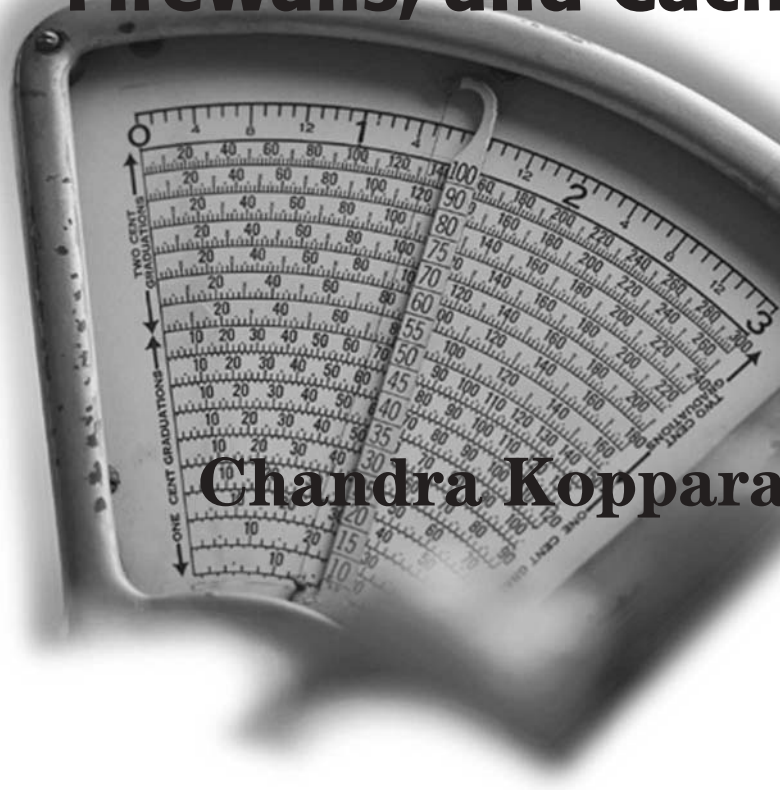


Load Balancing Servers, Firewalls, and Caches



Chandra Kopparapu

Wiley Computer Publishing

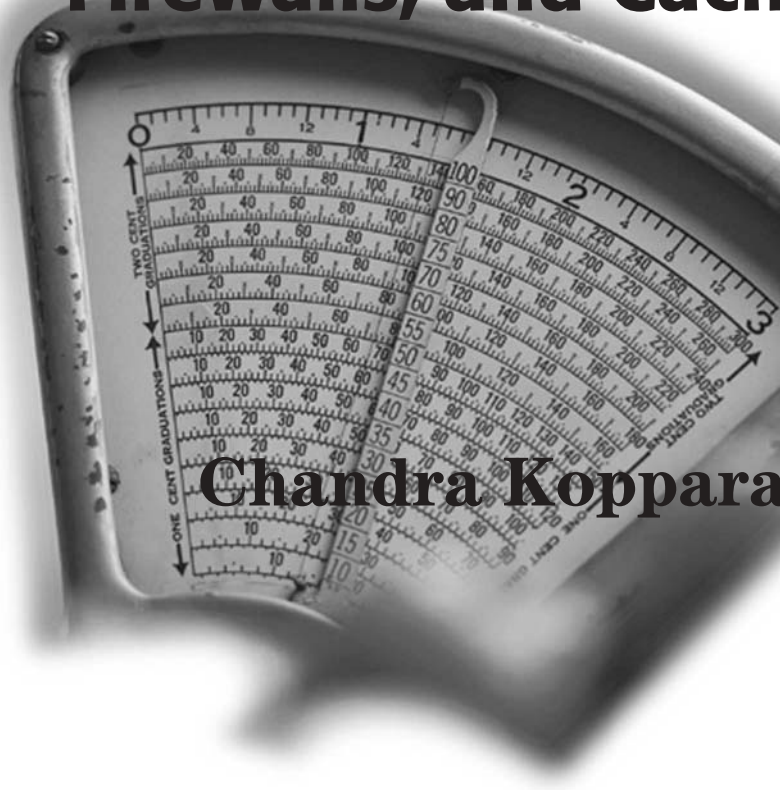


John Wiley & Sons, Inc.

NEW YORK • CHICHESTER • WEINHEIM • BRISBANE • SINGAPORE • TORONTO

Load Balancing Servers, Firewalls, and Caches

Load Balancing Servers, Firewalls, and Caches



Chandra Kopparapu

Wiley Computer Publishing



John Wiley & Sons, Inc.

NEW YORK • CHICHESTER • WEINHEIM • BRISBANE • SINGAPORE • TORONTO

Publisher: Robert Ipsen

Editor: Carol A. Long

Developmental Editor: Adaobi Obi

Managing Editor: Micheline Frederick

Text Design & Composition: Interactive Composition Corporation

Designations used by companies to distinguish their products are often claimed as trademarks. In all instances where John Wiley & Sons, Inc., is aware of a claim, the product names appear in initial capital or ALL CAPITAL LETTERS. Readers, however, should contact the appropriate companies for more complete information regarding trademarks and registration.

This book is printed on acid-free paper. ☹

Copyright © 2002 by Chandra Kopparapu. All rights reserved.

Published by John Wiley & Sons, Inc.

Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4744. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 605 Third Avenue, New York, NY 10158-0012, (212) 850-6011, fax (212) 850-6008, E-Mail: PERMREQ@WILEY.COM.

This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold with the understanding that the publisher is not engaged in professional services. If professional advice or other expert assistance is required, the services of a competent professional person should be sought.

Library of Congress Cataloging-in-Publication Data:

Kopparapu, Chandra.

Load balancing servers, firewalls, and caches / Chandra Kopparapu.

p. cm

Includes bibliographical references and index.

ISBN 0-471-41550-2 (cloth : alk. paper)

1. Client/server computing. 2. Firewalls (Computer security) I. Title.

QA76.9.C55 K67 2001

004.6--dc21

2001046757

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

**To my beloved daughters,
Divya and Nitya,
who bring so much joy to my life.**

TABLE OF CONTENTS

	Acknowledgments	xi
Chapter 1	Introduction	1
	The Need for Load Balancing	2
	The Server Environment	2
	The Network Environment	4
	Load Balancing: Definition and Applications	5
	Load-Balancing Products	7
	The Name Conundrum	7
	How This Book Is Organized	8
	Who Should Read This Book	9
	Summary	9
Chapter 2	Server Load Balancing: Basic Concepts	11
	Networking Fundamentals	12
	Switching Primer	12
	TCP Overview	13
	Web Server Overview	15
	The Server Farm with a Load Balancer	15
	Basic Packet Flow in Load Balancing	19
	Load-Distribution Methods	23
	Stateless Load Balancing	24
	Stateful Load Balancing	26
	Load-Distribution Methods	28
	Health Checks	33
	Basic Health Checks	34
	Application-Specific Health Checks	34
	Application Dependency	35
	Content Checks	35
	Scripting	36
	Agent-Based Checks	36
	The Ultimate Health Check	37

Network-Address Translation	38
Destination NAT	38
Source NAT	39
Reverse NAT	41
Enhanced NAT	42
Port-Address Translation	43
Direct Server Return	44
Summary	47
Chapter 3 Server Load Balancing: Advanced Concepts	49
Session Persistence	49
Defining Session Persistence	50
Types of Session Persistence	52
Source IP-Based Persistence Methods	53
The Megaproxy Problem	58
Delayed Binding	60
Cookie Switching	63
Cookie-Switching Applications	68
Cookie-Switching Considerations	69
SSL Session ID Switching	69
Designing to Deal with Session Persistence	72
HTTP to HTTPS Transition	74
URL Switching	77
Separating Static and Dynamic Content	79
URL Switching Usage Guidelines	80
Summary	82
Chapter 4 Network Design with Load Balancers	83
The Load Balancer as a Layer 2 Switch versus a Router	84
Simple Designs	87
Designing for High Availability	89
Active-Standby Configuration	90
Active-Active Configuration	92
Stateful Failover	96
Multiple VIPs	97
Load-Balancer Recovery	97
High-Availability Design Options	97
Communication between Load Balancers	108
Summary	108
Chapter 5 Global Server Load Balancing	111
The Need for GSLB	112
DNS Overview	113
DNS Concepts and Terminology	113

Local DNS Caching	115
Using Standard DNS for Load Balancing	116
HTTP Redirect	116
DNS-Based GSLB	118
Fitting the Load Balancer into the DNS Framework	118
Selecting the Best Site	122
Limitations of DNS-Based GSLB	133
GSLB Using Routing Protocols	134
Summary	137
Chapter 6 Load-Balancing Firewalls	139
Firewall Concepts	139
The Need for Firewall Load Balancing	140
Load-Balancing Firewalls	141
Traffic-Flow Analysis	142
Load-Distribution Methods	144
Checking the Health of a Firewall	147
Understanding Network Design in Firewall Load Balancing	148
Firewall and Load-Balancer Types	148
Network Design for Layer 3 Firewalls	149
Network Design for Layer 2 Firewalls	150
Advanced Firewall Concepts	151
Synchronized Firewalls	152
Firewalls Performing NAT	152
Addressing High Availability	153
Active-Standby versus Active-Active	155
Interaction between Routers and Load Balancers	155
Interaction between Load Balancers and Firewalls	157
Multizone Firewall Load Balancing	159
VPN Load Balancing	160
Summary	161
Chapter 7 Load-Balancing Caches	163
Cache Definition	163
Cache Types	164
Cache Deployment	165
Forward Proxy	165
Transparent Proxy	167
Reverse Proxy	168
Transparent-Reverse Proxy	169
Cache Load-Balancing Methods	170
Stateless Load Balancing	171

	Stateful Load Balancing	172
	Optimizing Load Balancing for Caches	172
	Content-Aware Cache Switching	175
	Summary	176
Chapter 8	Application Examples	177
	Enterprise Network	178
	Content-Distribution Networks	181
	Enterprise CDNs	182
	Content Provider	183
	CDN Service Providers	183
Chapter 9	The Future of Load-Balancing Technology	185
	Server Load Balancing	186
	The Load Balancer as a Security Device	186
	Cache Load Balancing	187
	SSL Acceleration	187
	Summary	188
Appendix	Standard Reference	189
References		191
Index		193

ACKNOWLEDGMENTS

First and foremost, my gratitude goes to my family. Without the support and understanding of my wife and encouragement from my parents, this book would not have been completed.

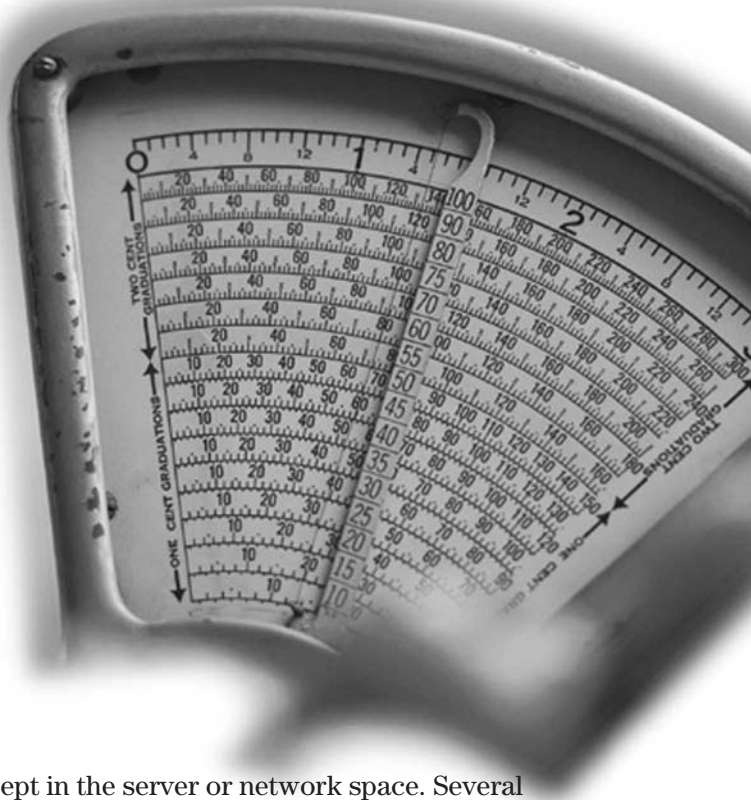
Rajkumar Jalan, principal architect for load balancers at Foundry Networks, was of invaluable help to me in understanding many load-balancing concepts when I was new to this technology. Many thanks go to Matthew Naugle, systems engineer at Foundry Networks, for encouraging me to write this book, giving me valuable feedback, and reviewing some of the chapters. Matt patiently spent countless hours with me, discussing several high-availability designs, and contributed valuable insight based on several customers he worked with. Terry Rolon, who used to work as a systems engineer at Foundry Networks, was also particularly helpful to me in coming up to speed on load-balancing products and network designs.

I would like to thank Mark Hoover of Acuitive Consulting for his thorough review and valuable analysis on Chapters 1, 2, 3, and 9. Mark has been very closely involved with the evolution of load-balancing products as an industry consultant and guided some load-balancing vendors in their early days. Many thanks to Brian Jacoby from America Online, who reviewed many of the chapters in this book from a customer perspective and provided valuable feedback.

Countless thanks to my colleagues at Foundry Networks, who worked with me over the last few years in advancing load-balancing product functionality and designing customer networks. I worked with many developers, systems engineers, customers, and technical support engineers to gain valuable insight into how load balancers are deployed and used by customers. Special thanks to Srin Ramadurai, David Cheung, Joe Tomasello, Ivy Hsu, Ron Szeto, and Ritesh Rekhi for helping me understand various aspects of load balancing functionality. I would also like to thank Ken Cheng, VP of Marketing at Foundry, for being supportive of this effort, and Bobby Johnson, Foundry's CEO, for giving me the opportunity to work with Foundry's load-balancing product line.

CHAPTER 1

Introduction



Load balancing is not a new concept in the server or network space. Several products perform different types of load balancing. For example, routers can distribute traffic across multiple paths to the same destination, balancing the load across different network resources. A server load balancer, on the other hand, distributes traffic among server resources rather than network resources. While load balancers started with simple load balancing, they soon evolved to perform a variety of functions: load balancing, traffic engineering, and intelligent traffic switching. Load balancers can perform sophisticated health checks on servers, applications, and content to improve availability and manageability. Because load balancers are deployed as the front end of a server farm, they also protect the servers from malicious users, and enhance security. Based on information in the IP packets or content in application requests, load balancers make intelligent decisions to direct the traffic appropriately—to the right data center, server, firewall, cache, or application.