THE PRENTICE HALL SERVICE TECHNOLOGY SERIES FROM THOMAS ERL

"This is a great book on the topic of cloud computing."
**Kapil Bakshi,** Architecture and Strategy, Cisco Systems Inc.

"We will recommend this book to Oracle customers, partners, and users for their journey toward cloud computing."
**Jürgen Kress,** Fusion Middleware Partner Adoption, Oracle EMEA

"A cloud computing book that will stand out and survive the test of time. . . . I highly recommend this book. . ."
**Christoph Schittko,** Principal Technology Strategist & Cloud Solution Director, Microsoft Corp.

"...a must-read for any IT professional interested in cloud computing."
**Andre Tost,** Senior Technical Staff Member, IBM Software Group

# Cloud Computing

## Concepts, Technology & Architecture

by Top-Selling Author Thomas Erl
with Zaigham Mahmood and Ricardo Puttini

PRENTICE HALL

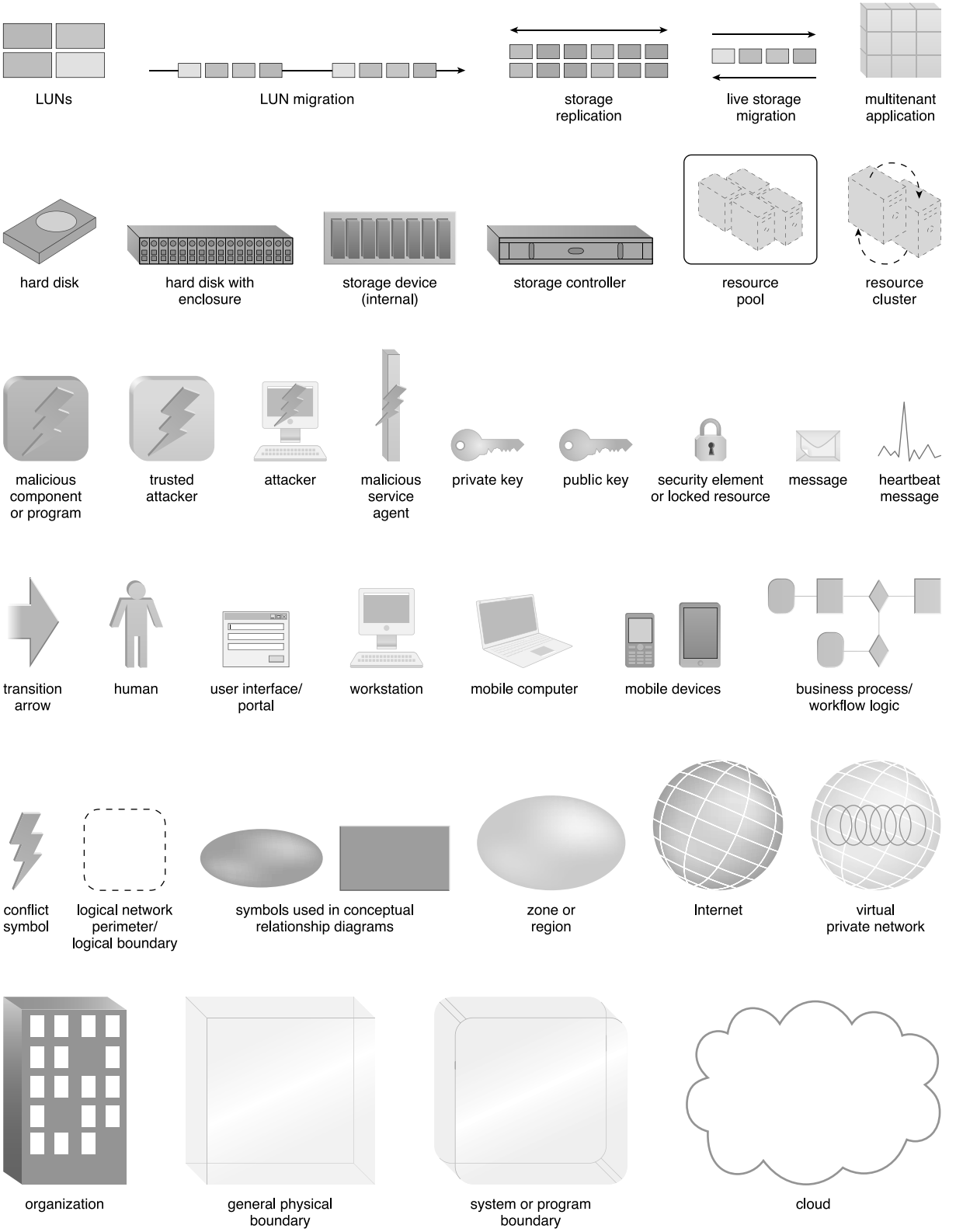
ServiceTech PRESS

Foreword by Pamela J. Wise-Martinez,
Department of Energy, National Nuclear Security Administration
Contributions by Gustavo Azzolin, Amin Naserpour, Vinicius Pacheco, Matthias Ziegler

FREE SAMPLE CHAPTER

SHARE WITH OTHERS

physical server
firewall
CPU
memory
network adapter
physical network device
connection ports or virtual switch
virtual desktops
hypervisor
virtualization platform
virtual server
virtual firewall
virtual network device
VIM
VI manager
live VM migration
router
core switch
top-of-rack switch
schema or data model
policy
general machine processable document
human readable document
ready-made environment
management system
remote administration system
actively processing
component or program
product, system or application
service agent
message queue
repository or storage device
shared storage
state data in memory
service with state data (stateful service)
repository with state data
grid service
service
service composition
service layer
service contract (chorded circle notation)
decoupled service contract
service inventory


LUNs
LUN migration
storage replication
live storage migration
multitenant application
hard disk
hard disk with enclosure
storage device (internal)
storage controller
resource pool
resource cluster
malicious component or program
trusted attacker
attacker
malicious service agent
private key
public key
security element or locked resource
message
heartbeat message
transition arrow
human
user interface/ portal
workstation
mobile computer
mobile devices
business process/ workflow logic
conflict symbol
logical network perimeter/ logical boundary
symbols used in conceptual relationship diagrams
zone or region
Internet
virtual private network
organization
general physical boundary
system or program boundary
cloud

Readers can download high-resolution,
full-color versions of all this book's figures at

www.informit.com/title/9780133387520

and

www.servicetechbooks.com/cloud.

## Praise for this Book

"Cloud computing, more than most disciplines in IT, suffers from too much talk and not enough practice. Thomas Erl has written a timely book that condenses the theory and buttresses it with real-world examples that demystify this important technology. An important guidebook for your journey into the cloud."

*—Scott Morrison, Chief Technology Officer, Layer 7 Technologies*

"An excellent, extremely well-written, lucid book that provides a comprehensive picture of cloud computing, covering multiple dimensions of the subject. The case studies presented in the book provide a real-world, practical perspective on leveraging cloud computing in an organization. The book covers a wide range of topics, from technology aspects to the business value provided by cloud computing. This is the best, most comprehensive book on the subject—a must-read for any cloud computing practitioner or anyone who wants to get an in-depth picture of cloud computing concepts and practical implementation."

*—Suzanne D'Souza, SOA/BPM Practice Lead, KBACE Technologies*

"This book offers a thorough and detailed description of cloud computing concepts, architectures, and technologies. It serves as a great reference for both newcomers and experts and is a must-read for any IT professional interested in cloud computing."

*—Andre Tost, Senior Technical Staff Member, IBM Software Group*

"This is a great book on the topic of cloud computing. It is impressive how the content spans from taxonomy, technology, and architectural concepts to important business considerations for cloud adoption. It really does provide a holistic view to this technology paradigm."

*—Kapil Bakshi, Architecture and Strategy, Cisco Systems Inc.*

"I have read every book written by Thomas Erl and *Cloud Computing* is another excellent publication and demonstration of Thomas Erl's rare ability to take the most complex topics and provide critical core concepts and technical information in a logical and understandable way."

*—Melanie A. Allison, Principal, Healthcare Technology Practice, Integrated Consulting Services*

"Companies looking to migrate applications or infrastructure to the cloud are often misled by buzzwords and industry hype. This work cuts through the hype and provides a detailed look, from investigation to contract to implementation to termination, at what it takes for an organization to engage with cloud service providers. This book really lays out the benefits and struggles with getting a company to an IaaS, PaaS, or SaaS solution."

—*Kevin Davis, Ph.D., Solutions Architect*

"Thomas, in his own distinct and erudite style, provides a comprehensive and a definitive book on cloud computing. Just like his previous masterpiece, *Service-Oriented Architecture: Concepts*, *Technology*, *and Design*, this book is sure to engage CxOs, cloud architects, and the developer community involved in delivering software assets on the cloud. Thomas and his authoring team have taken great pains in providing great clarity and detail in documenting cloud architectures, cloud delivery models, cloud governance, and economics of cloud, without forgetting to explain the core of cloud computing that revolves around Internet architecture and virtualization. As a reviewer for this outstanding book, I must admit I have learned quite a lot while reviewing the material. A 'must have' book that should adorn everybody's desk!"

—*Vijay Srinivasan, Chief Architect - Technology, Cognizant Technology Solutions*

"This book provides comprehensive and descriptive vendor-neutral coverage of cloud computing technology, from both technical and business aspects. It provides a deep-down analysis of cloud architectures and mechanisms that capture the real-world moving parts of cloud platforms. Business aspects are elaborated on to give readers a broader perspective on choosing and defining basic cloud computing business models. Thomas Erl's *Cloud Computing: Concepts*, *Technology & Architecture* is an excellent source of knowledge of fundamental and in-depth coverage of cloud computing."

—*Masykur Marhendra Sukmanegara, Communication Media & Technology, Consulting Workforce Accenture*

"The richness and depth of the topics discussed are incredibly impressive. The depth and breadth of the subject matter are such that a reader could become an expert in a short amount of time."

—*Jamie Ryan, Solutions Architect, Layer 7 Technologies*

"Demystification, rationalization, and structuring of implementation approaches have always been strong parts in each and every one of Thomas Erl's books. This book is no exception. It provides the definitive, essential coverage of cloud computing and, most importantly, presents this content in a very comprehensive manner. Best of all, this book follows the conventions of the previous service technology series titles, making it read like a natural extension of the library. I strongly believe that this will be another best-seller from one of the top-selling IT authors of the past decade."

*—Sergey Popov, Senior Enterprise Architect SOA/Security, Liberty Global International*

"A must-read for anyone involved in cloud design and decision making! This insightful book provides in-depth, objective, vendor-neutral coverage of cloud computing concepts, architecture models, and technologies. It will prove very valuable to anyone who needs to gain a solid understanding of how cloud environments work and how to design and migrate solutions to clouds."

*—Gijs in 't Veld, Chief Architect, Motion10*

"A reference book covering a wide range of aspects related to cloud providers and cloud consumers. If you would like to provide or consume a cloud service and need to know how, this is your book. The book has a clear structure to facilitate a good understanding of the various concepts of cloud."

*—Roger Stoffers, Solution Architect*

"Cloud computing has been around for a few years, yet there is still a lot of confusion around the term and what it can bring to developers and deployers alike. This book is a great way of finding out what's behind the cloud, and not in an abstract or high-level manner: It dives into all of the details that you'd need to know in order to plan for developing applications on cloud and what to look for when using applications or services hosted on a cloud. There are very few books that manage to capture this level of detail about the evolving cloud paradigm as this one does. It's a must for architects and developers alike."

*—Dr. Mark Little, Vice President, Red Hat*

"This book provides a comprehensive exploration of the concepts and mechanics behind clouds. It's written for anyone interested in delving into the details of how cloud environments function, how they are architected, and how they can impact business. This is the book for any organization seriously considering adopting cloud computing. It will pave the way to establishing your cloud computing roadmap."

*—Damian Maschek, SOA Architect, Deutsche Bahn*

"One of the best books on cloud computing I have ever read. It is complete yet vendor technology neutral and successfully explains the major concepts in a well-structured and disciplined way. It goes through all the definitions and provides many hints for organizations or professionals who are approaching and/or assessing cloud solutions. This book gives a complete list of topics playing fundamental roles in the cloud computing discipline. It goes through a full list of definitions very clearly stated. Diagrams are simple to understand and self-contained. Readers with different skill sets, expertise, and backgrounds will be able to understand the concepts seamlessly."

*—Antonio Bruno, Infrastructure and Estate Manager, UBS AG*

"*Cloud Computing: Concepts, Technology & Architecture* is a comprehensive book that focuses on what cloud computing is really all about…. This book will become the foundation on which many organizations will build successful cloud adoption projects. It is a must-read reference for both IT infrastructure and application architects interested in cloud computing or involved in cloud adoption projects. It contains extremely useful and comprehensive information for those who need to build cloud-based architectures or need to explain it to customers thinking about adopting cloud computing technology in their organization."

*—Johan Kumps, SOA Architect, RealDolmen*

"This book defines the basic terminology and patterns for the topic—a useful reference for the cloud practitioner. Concepts from multitenancy to hypervisor are presented in a succinct and clear manner. The underlying case studies provide wonderful real-worldness."

*—Dr. Thomas Rischbeck, Principal Architect, ipt*

"The book provides a good foundation to cloud services and issues in cloud service design. Chapters highlight key issues that need to be considered in learning how to think in cloud technology terms; this is highly important in today's business and technology environments where cloud computing plays a central role in connecting user services with virtualized resources and applications."

*—Mark Skilton, Director, Office of Strategy and Technology, Global Infrastructure Services, Capgemini*

"The book is well organized and covers basic concepts, technologies, and business models about cloud computing. It defines and explains a comprehensive list of terminologies and glossaries about cloud computing so cloud computing experts can speak and communicate with the same set of standardized language. The book is easy to understand and consistent with early published books from Thomas Erl.… It is a must-read for both beginners and experienced professionals."

*—Jian "Jeff" Zhong, Chief Technology Officer (Acting) and Chief Architect for SOA and Cloud Computing, Futrend Technology Inc.*

"Students of the related specialties can fulfill their educational process with very easily understood materials that are broadly illustrated and clearly described. Professors of different disciplines, from business analysis to IT implementation—even legal and financial monitoring—can use the book as an on-table lecturing manual. IT specialists of all ranks and fields of application will find the book as a practical and useful support for sketching solutions unbound to any particular vendor or brand."

*—Alexander Gromoff, Director of Science & Education, Center of Information Control Technologies, Chairman of BPM Chair in Business Informatics Department, National Research University "Higher School of Economics"*

"*Cloud Computing: Concepts, Technology & Architecture* is a comprehensive compendium of all the relevant information about the transformative cloud technology. Erl's latest title concisely and clearly illustrates the origins and positioning of the cloud paradigm as the next-generation computing model. All the chapters are carefully written and arranged in an easy-to-understand manner. This book will be immeasurably beneficial for business and IT professionals. It is set to shake up and help organize the world of cloud computing."

*—Pethuru Raj, Ph.D., Enterprise Architecture Consultant, Wipro*

"A cloud computing book that will stand out and survive the test of time, even in one of the fastest evolving areas of technology. This book does a great job breaking down the high level of complexity of cloud computing into easy-to-understand pieces. It goes beyond the basic, often repeated, explanations. It examines the fundamental concepts and the components, as well as the mechanisms and architectures that make up cloud computing environments. The approach gradually builds the reader's understanding from the ground up.

"In a rapidly evolving area like cloud computing, it's easy to focus on details and miss the big picture. The focus on concepts and architectural models instead of vendor-specific details allows readers to quickly gain essential knowledge of complex topics. The concepts come together in the last part of the book, which should be required reading for any decision maker evaluating when and how to start a transition to cloud computing. Its thorough, comprehensive coverage of fundamentals and advanced topics makes the book a valuable resource to keep on your desk or your eBook reader, regardless if you're new to the topic or you already have cloud experience.

"I highly recommend the book to those looking to implement or evaluate cloud environments, or simply looking to educate themselves in a field that will shape IT over the next decade."

—*Christoph Schittko, Principal Technology Strategist & Cloud Solution Director, Microsoft*

"*Cloud Computing: Concepts, Technology & Architecture* is an excellent resource for IT professionals and managers who want to learn and understand cloud computing, and who need to select or build cloud systems and solutions. It lays the foundation for cloud concepts, models, technologies, and mechanisms. As the book is vendor-neutral, it will remain valid for many years. We will recommend this book to Oracle customers, partners, and users for their journey toward cloud computing. This book has the potential to become the basis for a cloud computing manifesto, comparable to what was accomplished with the SOA manifesto."

—*Jürgen Kress, Fusion Middleware Partner Adoption, Oracle EMEA*

# Cloud Computing

*Concepts, Technology & Architecture*

Thomas Erl,
Zaigham Mahmood,
and Ricardo Puttini

PRENTICE HALL

PRENTICE HALL
UPPER SADDLE RIVER, NJ • BOSTON • INDIANAPOLIS • SAN FRANCISCO
NEW YORK • TORONTO • MONTREAL • LONDON • MUNICH • PARIS • MADRID
CAPE TOWN • SYDNEY • TOKYO • SINGAPORE • MEXICO CITY

ServiceTech PRESS

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and the publisher was aware of a trademark claim, the designations have been printed with initial capital letters or in all capitals.

The authors and publisher have taken care in the preparation of this book, but make no expressed or implied warranty of any kind and assume no responsibility for errors or omissions. No liability is assumed for incidental or consequential damages in connection with or arising out of the use of the information or programs contained herein.

The publisher offers excellent discounts on this book when ordered in quantity for bulk purchases or special sales, which may include electronic versions and/or custom covers and content particular to your business, training goals, marketing focus, and branding interests. For more information, please contact:

U.S. Corporate and Government Sales
(800) 382-3419
corpsales@pearsontechgroup.com

For sales outside the United States, please contact:

International Sales
international@pearsoned.com

Visit us on the Web: informit.com/ph

The Library of Congress Cataloging-in-Publication data is on file.

Copyright © 2013 Arcitura Education Inc.

All rights reserved. Printed in the United States of America. This publication is protected by copyright, and permission must be obtained from the publisher prior to any prohibited reproduction, storage in a retrieval system, or transmission in any form or by any means, electronic, mechanical, photocopying, recording, or likewise. To obtain permission to use material from this work, please submit a written request to Pearson Education, Inc., Permissions Department, One Lake Street, Upper Saddle River, New Jersey 07458, or you may fax your request to (201) 236-3290.

ISBN-13: 978-0-13-338752-0
ISBN-10: 0-13-338752-6

Text printed in the United States on recycled paper at Courier in Westford, Massachusetts.

First printing: May 2013

**Editor-in-Chief**
Mark L. Taub

**Managing Editor**
Kristy Hart

**Senior Project Editor**
Betsy Gratner

**Copy Editor and Development Editor**
Maria Lee

**Senior Indexer**
Cheryl Lenser

**Proofreaders**
Maria Lee
Williams Woods Publishing

**Publishing Coordinator**
Kim Boedigheimer

**Research Assistant**
Briana Lee

**Cover Designer**
Thomas Erl

**Compositor**
Bumpy Design

**Photos**
Thomas Erl
Dominika Sládkovičová

**Graphics**
KK Lui
Briana Lee

*To my family and friends*
—Thomas Erl

*To Zoya, Hanya, and Ozair with love*
—Zaigham Mahmood

*To Silvia, Luiza, Isadora, and Lucas*
—Ricardo Puttini

# Contents at a Glance

Foreword . . . . . xxix
Chapter 1: Introduction . . . . . 1
Chapter 2: Case Study Background . . . . . 13

**PART I: FUNDAMENTAL CLOUD COMPUTING**
Chapter 3: Understanding Cloud Computing . . . . . 25
Chapter 4: Fundamental Concepts and Models . . . . . 51
Chapter 5: Cloud-Enabling Technology. . . . . 79
Chapter 6: Fundamental Cloud Security . . . . . 117

**PART II: CLOUD COMPUTING MECHANISMS**
Chapter 7: Cloud Infrastructure Mechanisms . . . . . 139
Chapter 8: Specialized Cloud Mechanisms . . . . . 169
Chapter 9: Cloud Management Mechanisms . . . . . 213
Chapter 10: Cloud Security Mechanisms . . . . . 229

**PART III: CLOUD COMPUTING ARCHITECTURE**
Chapter 11: Fundamental Cloud Architectures . . . . . 255
Chapter 12: Advanced Cloud Architectures . . . . . 281
Chapter 13: Specialized Cloud Architectures . . . . . 323

**PART IV: WORKING WITH CLOUDS**
Chapter 14: Cloud Delivery Model Considerations . . . . . 359
Chapter 15: Cost Metrics and Pricing Models . . . . . 379
Chapter 16: Service Quality Metrics and SLAs . . . . . 403

**PART V: APPENDICES**
Appendix A: Case Study Conclusions . . . . . 421
Appendix B: Industry Standards Organizations . . . . . 427
Appendix C: Mapping Mechanisms to Characteristics . . . . . 433
Appendix D: Data Center Facilities (TIA-942) . . . . . 437
Appendix E: Emerging Technologies . . . . . 443
Appendix F: Cloud Provisioning Contracts . . . . . 449
Appendix G: Cloud Business Case Template . . . . . 461

About the Authors . . . . . 465
About the Foreword Contributor . . . . . 467
About the Contributors. . . . . 469

Index . . . . . 471

*This page intentionally left blank*

# Contents

**Foreword** . . . . . xxix

**Acknowledgments** . . . . . xxxiii

**CHAPTER 1: Introduction** . . . . . 1

1.1 Objectives of This Book . . . . . 3

1.2 What This Book Does Not Cover . . . . . 4

1.3 Who This Book Is For . . . . . 4

1.4 How This Book Is Organized . . . . . 4

Part I: Fundamental Cloud Computing . . . . . 5

*Chapter 3: Understanding Cloud Computing* . . . . . 5

*Chapter 4: Fundamental Concepts and Models* . . . . . 5

*Chapter 5: Cloud-Enabling Technology* . . . . . 5

*Chapter 6: Fundamental Cloud Security* . . . . . 5

Part II: Cloud Computing Mechanisms . . . . . 5

*Chapter 7: Cloud Infrastructure Mechanisms* . . . . . 6

*Chapter 8: Specialized Cloud Mechanisms* . . . . . 6

*Chapter 9: Cloud Management Mechanisms* . . . . . 6

*Chapter 10: Cloud Security Mechanisms* . . . . . 6

Part III: Cloud Computing Architecture . . . . . 6

*Chapter 11: Fundamental Cloud Architectures* . . . . . 6

*Chapter 12: Advanced Cloud Architectures* . . . . . 7

*Chapter 13: Specialized Cloud Architectures* . . . . . 7

Part IV: Working with Clouds . . . . . 7

*Chapter 14: Cloud Delivery Model Considerations* . . . . . 7

*Chapter 15: Cost Metrics and Pricing Models* . . . . . 8

*Chapter 16: Service Quality Metrics and SLAs* . . . . . 8

Part V: Appendices . . . . . 8

*Appendix A: Case Study Conclusions* . . . . . 8

*Appendix B: Industry Standards Organizations* . . . . . 8

*Appendix C: Mapping Mechanisms to Characteristics* . . . . . 8

*Appendix D: Data Center Facilities (TIA-942)* . . . . . 8

*Appendix E: Emerging Technologies* . . . . . 8

*Appendix F: Cloud Provisioning Contracts* . . . . . 9

*Appendix G: Cloud Business Case Template* . . . . . 9

1.5 Conventions . . . . . 9
Symbols and Figures . . . . . 9
Summary of Key Points . . . . . 9
1.6 Additional Information . . . . . 9
Updates, Errata, and Resources (www.servicetechbooks.com) . . . 9
Referenced Specifications (www.servicetechspecs.com). . . . . 10
The Service Technology Magazine (www.servicetechmag.com) . 10
International Service Technology Symposium (www.servicetechsymposium.com) . . . . . 10
What Is Cloud? (www.whatiscloud.com) . . . . . 10
What Is REST? (www.whatisrest.com) . . . . . 10
Cloud Computing Design Patterns (www.cloudpatterns.org) . . . . 10
Service-Orientation (www.serviceorientation.com) . . . . . 11
CloudSchool.com™ Certified Cloud (CCP) Professional (www.cloudschool.com). . . . . 11
SOASchool.com® SOA Certified (SOACP) Professional (www.soaschool.com) . . . . . 11
Notification Service . . . . . 11

**CHAPTER 2: Case Study Background . . . . . 13**

2.1 Case Study #1: ATN . . . . . 14
Technical Infrastructure and Environment . . . . . 14
Business Goals and New Strategy. . . . . 15
Roadmap and Implementation Strategy . . . . . 15
2.2 Case Study #2: DTGOV . . . . . 16
Technical Infrastructure and Environment . . . . . 17
Business Goals and New Strategy. . . . . 18
Roadmap and Implementation Strategy . . . . . 19
2.3 Case Study #3: Innovartus Technologies Inc.. . . . . 20
Technical Infrastructure and Environment . . . . . 20
Business Goals and Strategy. . . . . 20
Roadmap and Implementation Strategy . . . . . 21

# PART I: FUNDAMENTAL CLOUD COMPUTING


## CHAPTER 3: Understanding Cloud Computing . . . . . 25

3.1 Origins and Influences . . . . . 26
A Brief History . . . . . 26
Definitions . . . . . 27
Business Drivers . . . . . 28
*Capacity Planning . . . . . 28*
*Cost Reduction . . . . . 29*
*Organizational Agility . . . . . 30*
Technology Innovations . . . . . 30
*Clustering . . . . . 31*
*Grid Computing . . . . . 31*
*Virtualization . . . . . 32*
*Technology Innovations vs. Enabling Technologies . . . . . 32*

3.2 Basic Concepts and Terminology . . . . . 33
Cloud . . . . . 33
IT Resource . . . . . 34
On-Premise . . . . . 36
Cloud Consumers and Cloud Providers . . . . . 36
Scaling . . . . . 37
*Horizontal Scaling . . . . . 37*
*Vertical Scaling . . . . . 37*
Cloud Service . . . . . 38
Cloud Service Consumer . . . . . 40

3.3 Goals and Benefits . . . . . 40
Reduced Investments and Proportional Costs . . . . . 41
Increased Scalability . . . . . 42
Increased Availability and Reliability . . . . . 43

3.4 Risks and Challenges . . . . . 45
Increased Security Vulnerabilities . . . . . 45
Reduced Operational Governance Control . . . . . 45
Limited Portability Between Cloud Providers . . . . . 47
Multi-Regional Compliance and Legal Issues . . . . . 48

**CHAPTER 4: Fundamental Concepts and Models . . . . . 51**

4.1 Roles and Boundaries . . . . . 52
Cloud Provider . . . . . 52
Cloud Consumer . . . . . 52
Cloud Service Owner . . . . . 53
Cloud Resource Administrator . . . . . 54
Additional Roles . . . . . 56
Organizational Boundary . . . . . 56
Trust Boundary . . . . . 57

4.2 Cloud Characteristics . . . . . 58
On-Demand Usage . . . . . 59
Ubiquitous Access . . . . . 59
Multitenancy (and Resource Pooling) . . . . . 59
Elasticity . . . . . 61
Measured Usage . . . . . 61
Resiliency . . . . . 61

4.3 Cloud Delivery Models . . . . . 63
Infrastructure-as-a-Service (IaaS) . . . . . 64
Platform-as-a-Service (PaaS). . . . . 65
Software-as-a-Service (SaaS) . . . . . 66
Comparing Cloud Delivery Models . . . . . 67
Combining Cloud Delivery Models . . . . . 69
*IaaS + PaaS . . . . . 69*
*IaaS + PaaS + SaaS . . . . . 72*

4.4 Cloud Deployment Models . . . . . 73
Public Clouds . . . . . 73
Community Clouds. . . . . 74
Private Clouds . . . . . 75
Hybrid Clouds . . . . . 77
Other Cloud Deployment Models. . . . . 78

**CHAPTER 5: Cloud-Enabling Technology . . . . . 79**

5.1 Broadband Networks and Internet Architecture . . . . . 80
Internet Service Providers (ISPs) . . . . . 80
Connectionless Packet Switching (Datagram Networks) . . . . . 83
Router-Based Interconnectivity . . . . . 83
*Physical Network . . . . . 84*
*Transport Layer Protocol . . . . . 84*
*Application Layer Protocol . . . . . 85*
Technical and Business Considerations . . . . . 85
*Connectivity Issues . . . . . 85*
*Network Bandwidth and Latency Issues . . . . . 88*
*Cloud Carrier and Cloud Provider Selection . . . . . 89*

5.2 Data Center Technology . . . . . 90
Virtualization . . . . . 90
Standardization and Modularity . . . . . 90
Automation . . . . . 91
Remote Operation and Management . . . . . 92
High Availability . . . . . 92
Security-Aware Design, Operation, and Management . . . . . 92
Facilities . . . . . 92
Computing Hardware . . . . . 93
Storage Hardware . . . . . 93
Network Hardware . . . . . 95
*Carrier and External Networks Interconnection . . . . . 95*
*Web-Tier Load Balancing and Acceleration . . . . . 95*
*LAN Fabric . . . . . 95*
*SAN Fabric . . . . . 95*
*NAS Gateways . . . . . 95*
Other Considerations . . . . . 96

5.3 Virtualization Technology . . . . . 97
Hardware Independence . . . . . 98
Server Consolidation . . . . . 98
Resource Replication . . . . . 98
Operating System-Based Virtualization . . . . . 99
Hardware-Based Virtualization . . . . . 101
Virtualization Management . . . . . 102
Other Considerations . . . . . 102

5.4 Web Technology . . . . . 103
Basic Web Technology . . . . . 104
Web Applications . . . . . 104
5.5 Multitenant Technology . . . . . 106
5.6 Service Technology . . . . . 108
Web Services . . . . . 109
REST Services . . . . . 110
Service Agents . . . . . 111
Service Middleware . . . . . 112
5.7 Case Study Example . . . . . 113

**CHAPTER 6: Fundamental Cloud Security . . . . . 117**

6.1 Basic Terms and Concepts . . . . . 118
Confidentiality . . . . . 118
Integrity . . . . . 119
Authenticity . . . . . 119
Availability . . . . . 119
Threat . . . . . 120
Vulnerability . . . . . 120
Risk . . . . . 120
Security Controls . . . . . 120
Security Mechanisms . . . . . 121
Security Policies . . . . . 121
6.2 Threat Agents . . . . . 121
Anonymous Attacker . . . . . 122
Malicious Service Agent . . . . . 123
Trusted Attacker . . . . . 123
Malicious Insider . . . . . 123
6.3 Cloud Security Threats . . . . . 124
Traffic Eavesdropping . . . . . 124
Malicious Intermediary . . . . . 124
Denial of Service . . . . . 126
Insufficient Authorization . . . . . 127
Virtualization Attack . . . . . 127
Overlapping Trust Boundaries . . . . . 129

6.4 Additional Considerations . . . . . 131
Flawed Implementations . . . . . 131
Security Policy Disparity . . . . . 132
Contracts . . . . . 132
Risk Management . . . . . 133
6.5 Case Study Example . . . . . 135

## PART II: CLOUD COMPUTING MECHANISMS

**CHAPTER 7: Cloud Infrastructure Mechanisms . . . . . 139**

7.1 Logical Network Perimeter . . . . . 140
Case Study Example . . . . . 142
7.2 Virtual Server . . . . . 144
Case Study Example . . . . . 145
7.3 Cloud Storage Device . . . . . 149
Cloud Storage Levels . . . . . 149
Network Storage Interfaces . . . . . 150
Object Storage Interfaces . . . . . 151
Database Storage Interfaces . . . . . 151
*Relational Data Storage . . . . . 151*
*Non-Relational Data Storage . . . . . 152*
Case Study Example . . . . . 152
7.4 Cloud Usage Monitor . . . . . 155
Monitoring Agent . . . . . 155
Resource Agent . . . . . 155
Polling Agent . . . . . 157
Case Study Example . . . . . 157
7.5 Resource Replication . . . . . 161
Case Study Example . . . . . 162
7.6 Ready-Made Environment . . . . . 166
Case Study Example . . . . . 167

**CHAPTER 8: Specialized Cloud Mechanisms . . . . . 169**

8.1 Automated Scaling Listener . . . . . 170
Case Study Example . . . . . 172

8.2 Load Balancer . . . . . 176
Case Study Example . . . . . 177

8.3 SLA Monitor . . . . . 178
Case Study Example . . . . . 180
*SLA Monitor Polling Agent . . . . . 180*
*SLA Monitoring Agent . . . . . 180*

8.4 Pay-Per-Use Monitor . . . . . 184
Case Study Example . . . . . 187

8.5 Audit Monitor . . . . . 189
Case Study Example . . . . . 189

8.6 Failover System . . . . . 191
Active-Active . . . . . 191
Active-Passive . . . . . 194
Case Study Example . . . . . 196

8.7 Hypervisor . . . . . 200
Case Study Example . . . . . 201

8.8 Resource Cluster . . . . . 203
Case Study Example . . . . . 206

8.9 Multi-Device Broker . . . . . 208
Case Study Example . . . . . 209

8.10 State Management Database . . . . . 210
Case Study Example . . . . . 211

**CHAPTER 9: Cloud Management Mechanisms . . . . . 213**

9.1 Remote Administration System . . . . . 214
Case Study Example . . . . . 219
9.2 Resource Management System . . . . . 219
Case Study Example . . . . . 221
9.3 SLA Management System. . . . . 222
Case Study Example . . . . . 224
9.4 Billing Management System . . . . . 225
Case Study Example . . . . . 227

**CHAPTER 10: Cloud Security Mechanisms . . . . . 229**

10.1 Encryption . . . . . 230
Symmetric Encryption . . . . . 231
Asymmetric Encryption . . . . . 231
Case Study Example . . . . . 233
10.2 Hashing . . . . . 234
Case Study Example . . . . . 235
10.3 Digital Signature . . . . . 236
Case Study Example . . . . . 238
10.4 Public Key Infrastructure (PKI) . . . . . 240
Case Study Example . . . . . 242
10.5 Identity and Access Management (IAM) . . . . . 243
Case Study Example . . . . . 244
10.6 Single Sign-On (SSO) . . . . . 244
Case Study Example . . . . . 246
10.7 Cloud-Based Security Groups . . . . . 247
Case Study Example . . . . . 249
10.8 Hardened Virtual Server Images . . . . . 251
Case Study Example . . . . . 252

## PART III: CLOUD COMPUTING ARCHITECTURE

**CHAPTER 11: Fundamental Cloud Architectures . . . . . .255**

11.1 Workload Distribution Architecture . . . . . .256
11.2 Resource Pooling Architecture . . . . . .257
11.3 Dynamic Scalability Architecture . . . . . .262
11.4 Elastic Resource Capacity Architecture. . . . . .265
11.5 Service Load Balancing Architecture . . . . . .268
11.6 Cloud Bursting Architecture. . . . . .271
11.7 Elastic Disk Provisioning Architecture . . . . . .272
11.8 Redundant Storage Architecture . . . . . .275
11.9 Case Study Example . . . . . .277

**CHAPTER 12: Advanced Cloud Architectures . . . . . .281**

12.1 Hypervisor Clustering Architecture . . . . . .282
12.2 Load Balanced Virtual Server Instances Architecture. . .288
12.3 Non-Disruptive Service Relocation Architecture . . . . . .293
12.4 Zero Downtime Architecture . . . . . .298
12.5 Cloud Balancing Architecture . . . . . .299
12.6 Resource Reservation Architecture . . . . . .301
12.7 Dynamic Failure Detection and Recovery Architecture . .306
12.8 Bare-Metal Provisioning Architecture . . . . . .309
12.9 Rapid Provisioning Architecture . . . . . .312
12.10 Storage Workload Management Architecture . . . . . .315
12.11 Case Study Example . . . . . .321

**CHAPTER 13: Specialized Cloud Architectures . . . . . .323**

13.1 Direct I/O Access Architecture . . . . . .324

13.2 Direct LUN Access Architecture . . . . . .326

13.3 Dynamic Data Normalization Architecture. . . . . .329

13.4 Elastic Network Capacity Architecture . . . . . .330

13.5 Cross-Storage Device Vertical Tiering Architecture . . . .332

13.6 Intra-Storage Device Vertical Data Tiering Architecture .337

13.7 Load Balanced Virtual Switches Architecture . . . . . .340

13.8 Multipath Resource Access Architecture . . . . . .342

13.9 Persistent Virtual Network Configuration Architecture . . .344

13.10 Redundant Physical Connection for Virtual Servers Architecture . . . . . .347

13.11 Storage Maintenance Window Architecture . . . . . .350

**PART IV: WORKING WITH CLOUDS**

**CHAPTER 14: Cloud Delivery Model Considerations. . . . . .359**

14.1 Cloud Delivery Models: The Cloud Provider Perspective . . . . . .360

Building IaaS Environments . . . . . 360

*Data Centers . . . . . 361*

*Scalability and Reliability . . . . . .363*

*Monitoring . . . . . .363*

*Security . . . . . .364*

Equipping PaaS Environments . . . . . 364

*Scalability and Reliability . . . . . .365*

*Monitoring . . . . . .367*

*Security . . . . . .367*

Optimizing SaaS Environments . . . . . 367

*Security . . . . . 370*

14.2 Cloud Delivery Models: The Cloud Consumer Perspective . . . . . 370
Working with IaaS Environments . . . . . 370
*IT Resource Provisioning Considerations . . . . . 372*
Working with PaaS Environments . . . . . 373
*IT Resource Provisioning Considerations . . . . . 373*
Working with SaaS Services . . . . . 374
14.3 Case Study Example . . . . . 375

**CHAPTER 15: Cost Metrics and Pricing Models . . . . . 379**

15.1 Business Cost Metrics . . . . . 380
Up-Front and On-Going Costs . . . . . 380
Additional Costs . . . . . 381
Case Study Example . . . . . 382
Product Catalog Browser . . . . . 382
*On-Premise Up-Front Costs . . . . . 382*
*On-Premise On-Going Costs . . . . . 383*
*Cloud-Based Up-Front Costs . . . . . 383*
*Cloud-Based On-Going Costs . . . . . 383*
Client Database . . . . . 385
*On-Premise Up-Front Costs . . . . . 385*
*On-Premise On-Going Costs . . . . . 385*
*Cloud-Based Up-Front Costs . . . . . 385*
*Cloud-Based On-Going Costs . . . . . 385*
15.2 Cloud Usage Cost Metrics . . . . . 387
Network Usage . . . . . 387
*Inbound Network Usage Metric . . . . . 387*
*Outbound Network Usage Metric . . . . . 388*
*Intra-Cloud WAN Usage Metric . . . . . 388*
Server Usage . . . . . 389
*On-Demand Virtual Machine Instance Allocation Metric . . . . . 389*
*Reserved Virtual Machine Instance Allocation Metric . . . . . 389*
Cloud Storage Device Usage . . . . . 390
*On-Demand Storage Space Allocation Metric . . . . . 390*
*I/O Data Transferred Metric . . . . . 390*

Cloud Service Usage . . . . . 390
*Application Subscription Duration Metric . . . . . 390*
*Number of Nominated Users Metric . . . . . 391*
*Number of Transactions Users Metric . . . . . 391*

15.3 Cost Management Considerations . . . . . 391
Pricing Models . . . . . 393
Additional Considerations . . . . . 395
Case Study Example . . . . . 396
Virtual Server On-Demand Instance Allocation . . . . . 397
Virtual Server Reserved Instance Allocation . . . . . 399
Cloud Storage Device . . . . . 401
WAN Traffic . . . . . 401

**CHAPTER 16: Service Quality Metrics and SLAs . . . . . 403**

16.1 Service Quality Metrics . . . . . 404
Service Availability Metrics . . . . . 405
*Availability Rate Metric . . . . . 405*
*Outage Duration Metric . . . . . 406*
Service Reliability Metrics . . . . . 407
*Mean-Time Between Failures (MTBF) Metric . . . . . 407*
*Reliability Rate Metric . . . . . 407*
Service Performance Metrics . . . . . 407
*Network Capacity Metric . . . . . 408*
*Storage Device Capacity Metric . . . . . 408*
*Server Capacity Metric . . . . . 408*
*Web Application Capacity Metric . . . . . 408*
*Instance Starting Time Metric . . . . . 409*
*Response Time Metric . . . . . 409*
*Completion Time Metric . . . . . 409*
Service Scalability Metrics . . . . . 409
*Storage Scalability (Horizontal) Metric . . . . . 410*
*Server Scalability (Horizontal) Metric . . . . . 410*
*Server Scalability (Vertical) Metric . . . . . 410*
Service Resiliency Metrics . . . . . 411
*Mean-Time to Switchover (MTSO) Metric . . . . . 411*
*Mean-Time System Recovery (MTSR) Metric . . . . . 412*

16.2 Case Study Example . . . . . 412

16.3 SLA Guidelines . . . . . 413

16.4 Case Study Example . . . . . 416

*Scope and Applicability . . . . . 416*

*Service Quality Guarantees . . . . . 416*

*Definitions . . . . . 417*

*Usage of Financial Credits . . . . . 417*

*SLA Exclusions . . . . . 418*

## PART V: APPENDICES

**Appendix A: Case Study Conclusions . . . . . 421**

A.1 ATN . . . . . 422

A.2 DTGOV . . . . . 422

A.3 Innovartus . . . . . 424

**Appendix B: Industry Standards Organizations . . . . . 427**

B.1 National Institute of Standards and Technology (NIST) . . 428

B.2 Cloud Security Alliance (CSA) . . . . . 429

B.3 Distributed Management Task Force (DMTF) . . . . . 429

B.4 Storage Networking Industry Association (SNIA) . . . . . 430

B.5 Organization for the Advancement of Structured Information Standards (OASIS) . . . . . 430

B.6 The Open Group . . . . . 430

B.7 Open Cloud Consortium (OCC) . . . . . 431

B.8 European Telecommunications Standards Institute (ETSI) . . . . . 431

B.9 Telecommunications Industry Association (TIA) . . . . . 431

B.10 Liberty Alliance . . . . . 432

B.11 Open Grid Forum (OGF) . . . . . 432

**Appendix C: Mapping Mechanisms to Characteristics. . .433**

**Appendix D: Data Center Facilities (TIA-942). . . . . .437**

D.1 Primary Rooms . . . . . .438
Electrical Room . . . . . 438
Mechanical Room . . . . . 438
Storage and Staging . . . . . 438
Offices, Operations Center, and Support. . . . . 438
Telecommunications Entrance . . . . . 438
Computer Room. . . . . 439

D.2 Environmental Controls. . . . . .440
External Electrical Power Provider Interconnection . . . . . 440
Power Distribution . . . . . .441
Uninterruptible Power Source (UPS) . . . . . .441
Power Engine-Generator . . . . . .441

D.3 Infrastructure Redundancy Summary. . . . . .442

**Appendix E: Emerging Technologies. . . . . .443**

E.1 Autonomic Computing. . . . . .444

E.2 Grid Computing . . . . . .445
Service Grids . . . . . 446

**Appendix F: Cloud Provisioning Contracts . . . . . .449**

F.1 Cloud Provisioning Contract Structure . . . . . .450
Terms of Service. . . . . 452
*Service Usage Policy . . . . . 452*
*Security and Privacy Policy . . . . . 453*
*Warranties and Liabilities. . . . . 455*
*Rights and Responsibilities. . . . . 455*
*Termination and Renewal . . . . . 456*
Specifications and SLAs . . . . . 456
Pricing and Billing. . . . . 457

Other Issues . . . . . 457
*Legal and Compliance Issues . . . . . 457*
*Auditability and Accountability . . . . . 457*
*Changes in the Contract Terms and Conditions . . . . . 457*
F.2 Cloud Provider Selection Guidelines . . . . . 458
Cloud Provider Viability . . . . . 458

**Appendix G: Cloud Business Case Template . . . . . 461**

G.1 Business Case Identification . . . . . 462
G.2 Business Needs . . . . . 462
G.3 Target Cloud Environment . . . . . 463
G.4 Technical Issues . . . . . 464
G.5 Economic Factors . . . . . 464

**About the Authors . . . . . 465**

Thomas Erl . . . . . 465
Zaigham Mahmood . . . . . 465
Ricardo Puttini . . . . . 466

**About the Foreword Contributor . . . . . 467**

Pamela J. Wise-Martinez, MSc . . . . . 467

**About the Contributors . . . . . 469**

Gustavo Azzolin, BSc, MSc . . . . . 469
Amin Naserpour . . . . . 469
Vinícius Pacheco, MSc . . . . . 470
Matthias Ziegler . . . . . 470

**Index . . . . . 471**

# Foreword by Pamela J. Wise-Martinez

The idea of cloud computing isn't new, or overly complicated from a technology resources and internetworking perspective. What's new is the growth and maturity of cloud computing methods, and strategies that enable the goals of business agility.

Looking back, the phrase "utility computing" didn't captivate or create the stir in the information industry as the term "cloud computing" has in recent years. Nevertheless, appreciation of readily available resources has arrived and the utilitarian or servicing features are what are at the heart of *outsourcing* the access of information technology resources and services. In this light, cloud computing represents a flexible, cost-effective, and proven delivery platform for business and consumer information services over the Internet. Cloud computing has become an industry game changer as businesses and information technology leaders realize the potential in *combining and sharing* computing resources as opposed to *building and maintaining* them.

There's seemingly no shortage of views regarding the benefits of cloud computing nor is there a shortage of vendors willing to offer services in either open source or promising commercial solutions. Beyond the hype, there are many aspects of the cloud that have earned new consideration due to their increased service capability and potential efficiencies. The ability to demonstrate transforming results in cloud computing to resolve traditional business problems using information technology management best

practices now exists. In the case of economic impacts, the principle of *pay-as-you-go* and *computer agnostic services* are concepts ready for prime time. We can measure performance as well as calculate the economic and environmental effects of cloud computing today.

The architectural change from *client-server* to *service orientation* led to an evolution of composable and reusable code; though the practice had been around for many years, it is now the de facto approach used to lower cost and identify best practices and patterns for increasing business agility. This has advanced the computer software industry's design methods, components, and engineering. Comparatively, the wide acceptance and adoption of cloud computing is revolutionizing information and technology resource management. We now have the ability to outsource hardware and software capabilities on a large-scale to fulfill end-to-end business automation requirements. Marks and Lozano understood this emergence and the need for better software design: *"…we now have the ability to collect, transport, process, store, and access data nearly anywhere in nearly arbitrary volume."* The limitations depend largely on how "cloudy" or cloud-aware the service/component is, and hence the need for better software architecture. (Eric A. Marks and Roberto Lozano [*Executive Guide to Cloud Computing*]).

The reusable evolution through service architecture reinforces a focus on business objectives as opposed to the number of computing platforms to support. As a viable resource management alternative, cloud computing is fundamentally changing the way we think about computing solutions in retail, education, and public sectors. The use of cloud computing architecture and standards are driving unique ways in which computing solutions are delivered, as well as platform diversity to meet bottom-line business objectives.

Thomas Erl's body of work on service technology guided the technology industry through eloquent illustrations and literature over the past decade. Thomas' brilliant efforts on principles, concepts, patterns, and expressions gave the information technology community an *evolved* software architecture approach that now forms a foundation for cloud computing goals to be successfully fulfilled in practice. This is a key assertion, as cloud computing is no longer a far-reaching concept of the future, but rather a dominant information technology service option and resource delivery presence.

Thomas' *Cloud Computing: Concepts, Technology & Architecture* takes the industry beyond the definitions of cloud computing and juxtaposes virtualization, grid, and sustainment strategies as contrasted in day to day operations. Thomas and his team of authors take the reader from beginning to end with the essential elements of cloud computing,

its history, innovation, and demand. Through case studies and architectural models they articulate service requirements, infrastructure, security, and outsourcing of salient computing resources.

Thomas again enlightens the industry with poignant analysis and reliable architecture-driven practices and principles. No matter the level of interest or experience, the reader will find clear value in this in-depth, vendor-neutral study of cloud computing.

Pamela J. Wise-Martinez,
Inventor and Chief Architect
Department of Energy, National Nuclear Security Administration

*(Disclaimer: The views expressed are the personal views of the author and are not intended to reflect either the views of the U.S. Government, the U.S. Department of Energy, or the National Nuclear Security Administration.)*

*This page intentionally left blank*

# Acknowledgments

In alphabetical order by last name:

- Ahmed Aamer, AlFaisaliah Group
- Randy Adkins, Modus21
- Melanie Allison, Integrated Consulting Services
- Gabriela Inacio Alves, University of Brasilia
- Marcelo Ancelmo, IBM Rational Software Services
- Kapil Bakshi, Cisco Systems
- Toufic Boubez, Metafor Software
- Antonio Bruno, UBS AG
- Dr. Paul Buhler, Modus21
- Pethuru Raj Cheliah, Wipro
- Kevin Davis, Ph.D.
- Suzanne D'Souza, KBACE Technologies
- Alexander Gromoff, Center of Information Control Technologies
- Chris Haddad, WSO2
- Richard Hill, University of Derby
- Michaela Iorga, Ph.D.
- Johan Kumps, RealDolmen
- Gijs in 't Veld, Motion10
- Masykur Marhendra, Consulting Workforce Accenture
- Damian Maschek, Deutshe Bahn
- Claynor Mazzarolo, IBTI
- Steve Millidge, C2B2
- Jorge Minguez, Thales Deutschland
- Scott Morrison, Layer 7
- Amin Naserpour, HP

- Vicente Navarro, European Space Agency
- Laura Olson, IBM WebSphere
- Tony Pallas, Intel
- Cesare Pautasso, University of Lugano
- Sergey Popov, Liberty Global International
- Olivier Poupeney, Dreamface Interactive
- Alex Rankov, EMC
- Dan Rosanova, West Monroe Partners
- Jaime Ryan, Layer 7
- Filippos Santas, Credit Suisse
- Christoph Schittko, Microsoft
- Guido Schmutz, Trivadis
- Mark Skilton, Capgemini
- Gary Smith, CloudComputingArchitect.com
- Vijay Srinivasan, Cognizant
- Daniel Starcevich, Raytheon
- Roger Stoffers, HP
- Andre Toffanello, IBTI
- Andre Tost, IBM Software Group
- Bernd Trops, talend
- Clemens Utschig, Boehringer Ingelheim Pharma
- Ignaz Wanders, Archimiddle
- Philip Wik, Redflex
- Jorge Williams, Rackspace
- Dr. Johannes Maria Zaha
- Jeff Zhong, Futrend Technologies

Special thanks to the CloudSchool.com research and development team that produced the CCP course modules upon which this book is based.

# Chapter 3

# Understanding Cloud Computing


**3.1** Origins and Influences

**3.2** Basic Concepts and Terminology

**3.3** Goals and Benefits

**3.4** Risks and Challenges

This is the first of two chapters that provide an overview of introductory cloud computing topics. It begins with a brief history of cloud computing along with short descriptions of its business and technology drivers. This is followed by definitions of basic concepts and terminology, in addition to explanations of the primary benefits and challenges of cloud computing adoption.

## 3.1 Origins and Influences

### A Brief History

The idea of computing in a "cloud" traces back to the origins of utility computing, a concept that computer scientist John McCarthy publicly proposed in 1961:

> *"If computers of the kind I have advocated become the computers of the future, then computing may someday be organized as a public utility just as the telephone system is a public utility. … The computer utility could become the basis of a new and important industry."*

In 1969, Leonard Kleinrock, a chief scientist of the Advanced Research Projects Agency Network or ARPANET project that seeded the Internet, stated:

> *"As of now, computer networks are still in their infancy, but as they grow up and become sophisticated, we will probably see the spread of 'computer utilities' …".*

The general public has been leveraging forms of Internet-based computer utilities since the mid-1990s through various incarnations of search engines (Yahoo!, Google), e-mail services (Hotmail, Gmail), open publishing platforms (MySpace, Facebook, YouTube), and other types of social media (Twitter, LinkedIn). Though consumer-centric, these services popularized and validated core concepts that form the basis of modern-day cloud computing.

In the late 1990s, Salesforce.com pioneered the notion of bringing remotely provisioned services into the enterprise. In 2002, Amazon.com launched the Amazon Web Services (AWS) platform, a suite of enterprise-oriented services that provide remotely provisioned storage, computing resources, and business functionality.

A slightly different evocation of the term "Network Cloud" or "Cloud" was introduced in the early 1990s throughout the networking industry. It referred to an abstraction layer derived in the delivery methods of data across heterogeneous public and semi-public networks that were primarily packet-switched, although cellular networks used the "Cloud" term as well. The networking method at this point supported the transmission of data from one end-point (local network) to the "Cloud" (wide area network) and then further decomposed to another intended end-point. This is relevant, as the networking industry still references the use of this term, and is considered an early adopter of the concepts that underlie utility computing.

It wasn't until 2006 that the term "cloud computing" emerged in the commercial arena. It was during this time that Amazon launched its Elastic Compute Cloud (EC2) services that enabled organizations to "lease" computing capacity and processing power to run their enterprise applications. Google Apps also began providing browser-based enterprise applications in the same year, and three years later, the Google App Engine became another historic milestone.

### Definitions

A Gartner report listing cloud computing at the top of its strategic technology areas further reaffirmed its prominence as an industry trend by announcing its formal definition as:

> *"…a style of computing in which scalable and elastic IT-enabled capabilities are delivered as a service to external customers using Internet technologies."*

This is a slight revision of Gartner's original definition from 2008, in which "massively scalable" was used instead of "scalable and elastic." This acknowledges the importance of scalability in relation to the ability to scale vertically and not just to enormous proportions.

Forrester Research provided its own definition of cloud computing as:

> *"…a standardized IT capability (services, software, or infrastructure) delivered via Internet technologies in a pay-per-use, self-service way."*

The definition that received industry-wide acceptance was composed by the National Institute of Standards and Technology (NIST). NIST published its original definition back in 2009, followed by a revised version after further review and industry input that was published in September of 2011:

> *"Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model is composed of five essential characteristics, three service models, and four deployment models."*

This book provides a more concise definition:

> *"Cloud computing is a specialized form of distributed computing that introduces utilization models for remotely provisioning scalable and measured resources."*

This simplified definition is in line with all of the preceding definition variations that were put forth by other organizations within the cloud computing industry. The characteristics, service models, and deployment models referenced in the NIST definition are further covered in Chapter 4.

## Business Drivers

Before delving into the layers of technologies that underlie clouds, the motivations that led to their creation by industry leaders must first be understood. Several of the primary business drivers that fostered modern cloud-based technology are presented in this section.

The origins and inspirations of many of the characteristics, models, and mechanisms covered throughout subsequent chapters can be traced back to the upcoming business drivers. It is important to note that these influences shaped clouds and the overall cloud computing market from both ends. They have motivated organizations to adopt cloud computing in support of their business automation requirements. They have correspondingly motivated other organizations to become providers of cloud environments and cloud technology vendors in order to create and meet the demand to fulfill consumer needs.

### *Capacity Planning*

Capacity planning is the process of determining and fulfilling future demands of an organization's IT resources, products, and services. Within this context, *capacity* represents the maximum amount of work that an IT resource is capable of delivering in a given period of time. A discrepancy between the capacity of an IT resource and its demand can result in a system becoming either inefficient (over-provisioning) or unable

to fulfill user needs (under-provisioning). Capacity planning is focused on minimizing this discrepancy to achieve predictable efficiency and performance.

Different capacity planning strategies exist:

- *Lead Strategy* – adding capacity to an IT resource in anticipation of demand
- *Lag Strategy* – adding capacity when the IT resource reaches its full capacity
- *Match Strategy* – adding IT resource capacity in small increments, as demand increases

Planning for capacity can be challenging because it requires estimating usage load fluctuations. There is a constant need to balance peak usage requirements without unnecessary over-expenditure on infrastructure. An example is outfitting IT infrastructure to accommodate maximum usage loads which can impose unreasonable financial investments. In such cases, moderating investments can result in under-provisioning, leading to transaction losses and other usage limitations from lowered usage thresholds.

#### *Cost Reduction*

A direct alignment between IT costs and business performance can be difficult to maintain. The growth of IT environments often corresponds to the assessment of their maximum usage requirements. This can make the support of new and expanded business automations an ever-increasing investment. Much of this required investment is funneled into infrastructure expansion because the usage potential of a given automation solution will always be limited by the processing power of its underlying infrastructure.

Two costs need to be accounted for: the cost of acquiring new infrastructure, and the cost of its ongoing ownership. Operational overhead represents a considerable share of IT budgets, often exceeding up-front investment costs.

Common forms of infrastructure-related operating overhead include the following:

- technical personnel required to keep the environment operational
- upgrades and patches that introduce additional testing and deployment cycles
- utility bills and capital expense investments for power and cooling
- security and access control measures that need to be maintained and enforced to protect infrastructure resources
- administrative and accounts staff that may be required to keep track of licenses and support arrangements

The on-going ownership of internal technology infrastructure can encompass burdensome responsibilities that impose compound impacts on corporate budgets. An IT department can consequently become a significant—and at times overwhelming—drain on the business, potentially inhibiting its responsiveness, profitability, and overall evolution.

### *Organizational Agility*

Businesses need the ability to adapt and evolve to successfully face change caused by both internal and external factors. Organizational agility is the measure of an organization's responsiveness to change.

An IT enterprise often needs to respond to business change by scaling its IT resources beyond the scope of what was previously predicted or planned for. For example, infrastructure may be subject to limitations that prevent the organization from responding to usage fluctuations—even when anticipated—if previous capacity planning efforts were restricted by inadequate budgets.

In other cases, changing business needs and priorities may require IT resources to be more available and reliable than before. Even if sufficient infrastructure is in place for an organization to support anticipated usage volumes, the nature of the usage may generate runtime exceptions that bring down hosting servers. Due to a lack of reliability controls within the infrastructure, responsiveness to consumer or customer requirements may be reduced to a point whereby a business' overall continuity is threatened.

On a broader scale, the up-front investments and infrastructure ownership costs that are required to enable new or expanded business automation solutions may themselves be prohibitive enough for a business to settle for IT infrastructure of less-than-ideal quality, thereby decreasing its ability to meet real-world requirements.

Worse yet, the business may decide against proceeding with an automation solution altogether upon review of its infrastructure budget, because it simply cannot afford to. This form of inability to respond can inhibit an organization from keeping up with market demands, competitive pressures, and its own strategic business goals.

## Technology Innovations

Established technologies are often used as inspiration and, at times, the actual foundations upon which new technology innovations are derived and built. This section briefly describes the pre-existing technologies considered to be the primary influences on cloud computing.

### *Clustering*

A cluster is a group of independent IT resources that are interconnected and work as a single system. System failure rates are reduced while availability and reliability are increased, since redundancy and failover features are inherent to the cluster.

A general prerequisite of hardware clustering is that its component systems have reasonably identical hardware and operating systems to provide similar performance levels when one failed component is to be replaced by another. Component devices that form a cluster are kept in synchronization through dedicated, high-speed communication links.

The basic concept of built-in redundancy and failover is core to cloud platforms. Clustering technology is explored further in Chapter 8 as part of the *Resource Cluster* mechanism description.

### *Grid Computing*

A computing grid (or "computational grid") provides a platform in which computing resources are organized into one or more logical pools. These pools are collectively coordinated to provide a high performance distributed grid, sometimes referred to as a "super virtual computer." Grid computing differs from clustering in that grid systems are much more loosely coupled and distributed. As a result, grid computing systems can involve computing resources that are heterogeneous and geographically dispersed, which is generally not possible with cluster computing-based systems.

Grid computing has been an on-going research area in computing science since the early 1990s. The technological advancements achieved by grid computing projects have influenced various aspects of cloud computing platforms and mechanisms, specifically in relation to common feature-sets such as networked access, resource pooling, and scalability and resiliency. These types of features can be established by both grid computing and cloud computing, in their own distinctive approaches.

For example, grid computing is based on a middleware layer that is deployed on computing resources. These IT resources participate in a grid pool that implements a series of workload distribution and coordination functions. This middle tier can contain load balancing logic, failover controls, and autonomic configuration management, each having previously inspired similar—and several more sophisticated—cloud computing technologies. It is for this reason that some classify cloud computing as a descendant of earlier grid computing initiatives.

### *Virtualization*

Virtualization represents a technology platform used for the creation of virtual instances of IT resources. A layer of virtualization software allows physical IT resources to provide multiple virtual images of themselves so that their underlying processing capabilities can be shared by multiple users.

Prior to the advent of virtualization technologies, software was limited to residing on and being coupled with static hardware environments. The virtualization process severs this software-hardware dependency, as hardware requirements can be simulated by emulation software running in virtualized environments.

Established virtualization technologies can be traced to several cloud characteristics and cloud computing mechanisms, having inspired many of their core features. As cloud computing evolved, a generation of *modern* virtualization technologies emerged to overcome the performance, reliability, and scalability limitations of traditional virtualization platforms.

As a foundation of contemporary cloud technology, modern virtualization provides a variety of virtualization types and technology layers that are discussed separately in Chapter 5.

### *Technology Innovations vs. Enabling Technologies*

It is essential to highlight several other areas of technology that continue to contribute to modern-day cloud-based platforms. These are distinguished as *cloud-enabling technologies*, the following of which are covered in Chapter 5:

- Broadband Networks and Internet Architecture
- Data Center Technology
- (Modern) Virtualization Technology
- Web Technology
- Multitenant Technology
- Service Technology

Each of these cloud-enabling technologies existed in some form prior to the formal advent of cloud computing. Some were refined further, and on occasion even redefined, as a result of the subsequent evolution of cloud computing.

**SUMMARY OF KEY POINTS**

- The primary business drivers that exposed the need for cloud computing and led to its formation include capacity planning, cost reduction, and organizational agility.
- The primary technology innovations that influenced and inspired key distinguishing features and aspects of cloud computing include clustering, grid computing, and traditional forms of virtualization.

## 3.2 Basic Concepts and Terminology

This section establishes a set of basic terms that represent the fundamental concepts and aspects pertaining to the notion of a cloud and its most primitive artifacts.

### Cloud

A *cloud* refers to a distinct IT environment that is designed for the purpose of remotely provisioning scalable and measured IT resources. The term originated as a metaphor for the Internet which is, in essence, a network of networks providing remote access to a set of decentralized IT resources. Prior to cloud computing becoming its own formalized IT industry segment, the symbol of a cloud was commonly used to represent the Internet in a variety of specifications and mainstream documentation of Web-based architectures. This same symbol is now used to specifically represent the boundary of a cloud environment, as shown in Figure 3.1.

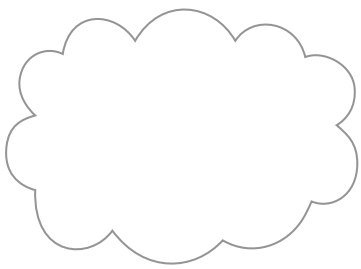


**Figure 3.1**
The symbol used to denote the boundary of a cloud environment.

It is important to distinguish the term "cloud" and the cloud symbol from the Internet. As a specific environment used to remotely provision IT resources, a cloud has a finite boundary. There are many individual clouds that are accessible via the Internet.

Whereas the Internet provides open access to many Web-based IT resources, a cloud is typically privately owned and offers access to IT resources that is metered.

Much of the Internet is dedicated to the access of content-based IT resources published via the World Wide Web. IT resources provided by cloud environments, on the other hand, are dedicated to supplying back-end processing capabilities and user-based access to these capabilities. Another key distinction is that it is not necessary for clouds to be Web-based even if they are commonly based on Internet protocols and technologies. Protocols refer to standards and methods that allow computers to communicate with each other in a pre-defined and structured manner. A cloud can be based on the use of any protocols that allow for the remote access to its IT resources.

**NOTE**

Diagrams in this book depict the Internet using the globe symbol.

## IT Resource

An *IT resource* is a physical or virtual IT-related artifact that can be either software-based, such as a virtual server or a custom software program, or hardware-based, such as a physical server or a network device (Figure 3.2).

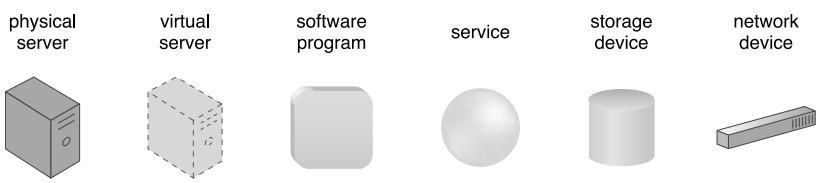



**Figure 3.2**
Examples of common IT resources and their corresponding symbols.

Figure 3.3 illustrates how the cloud symbol can be used to define a boundary for a cloud-based environment that hosts and provisions a set of IT resources. The displayed IT resources are consequently considered to be cloud-based IT resources.

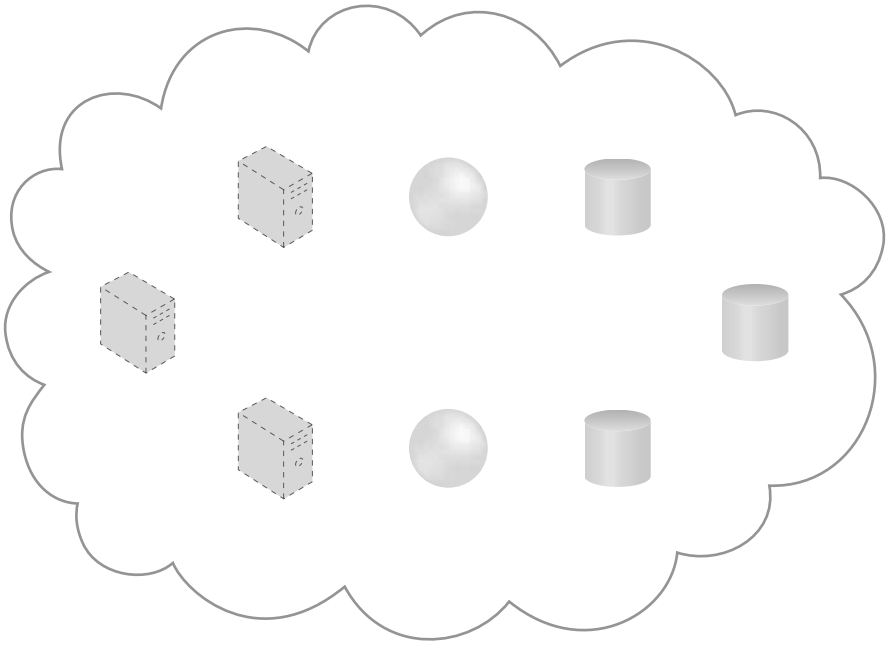


**Figure 3.3**
A cloud is hosting eight IT resources: three virtual servers, two cloud services, and three storage devices.

Technology architectures and various interaction scenarios involving IT resources are illustrated in diagrams like the one shown in Figure 3.3. It is important to note the following points when studying and working with these diagrams:

- The IT resources shown within the boundary of a given cloud symbol usually do not represent all of the available IT resources hosted by that cloud. Subsets of IT resources are generally highlighted to demonstrate a particular topic.
- Focusing on the relevant aspects of a topic requires many of these diagrams to intentionally provide abstracted views of the underlying technology architectures. This means that only a portion of the actual technical details are shown.

Furthermore, some diagrams will display IT resources outside of the cloud symbol. This convention is used to indicate IT resources that are not cloud-based.

**NOTE**

The virtual server IT resource displayed in Figure 3.2 is further discussed in Chapters 5 and 7. Physical servers are sometimes referred to as *physical hosts* (or just *hosts*) in reference to the fact that they are responsible for hosting virtual servers.

### On-Premise

As a distinct and remotely accessible environment, a cloud represents an option for the deployment of IT resources. An IT resource that is hosted in a conventional IT enterprise within an organizational boundary (that does not specifically represent a cloud) is considered to be located on the premises of the IT enterprise, or *on-premise* for short. In other words, the term "on-premise" is another way of stating "on the premises of a controlled IT environment that is not cloud-based." This term is used to qualify an IT resource as an alternative to "cloud-based." An IT resource that is on-premise cannot be cloud-based, and vice-versa.

Note the following key points:

- An on-premise IT resource can access and interact with a cloud-based IT resource.
- An on-premise IT resource can be moved to a cloud, thereby changing it to a cloud-based IT resource.
- Redundant deployments of an IT resource can exist in both on-premise and cloud-based environments.

If the distinction between on-premise and cloud-based IT resources is confusing in relation to private clouds (described in the *Cloud Deployment Models* section of Chapter 4), then an alternative qualifier can be used.

### Cloud Consumers and Cloud Providers

The party that provides cloud-based IT resources is the *cloud provider*. The party that uses cloud-based IT resources is the *cloud consumer*. These terms represent roles usually assumed by organizations in relation to clouds and corresponding cloud provisioning contracts. These roles are formally defined in Chapter 4, as part of the *Roles and Boundaries* section.

## Scaling

Scaling, from an IT resource perspective, represents the ability of the IT resource to handle increased or decreased usage demands.

The following are types of scaling:

- *Horizontal Scaling* – scaling out and scaling in
- *Vertical Scaling* – scaling up and scaling down

The next two sections briefly describe each.

### *Horizontal Scaling*

The allocating or releasing of IT resources that are of the same type is referred to as *horizontal scaling* (Figure 3.4). The horizontal allocation of resources is referred to as *scaling out* and the horizontal releasing of resources is referred to as *scaling in*. Horizontal scaling is a common form of scaling within cloud environments.

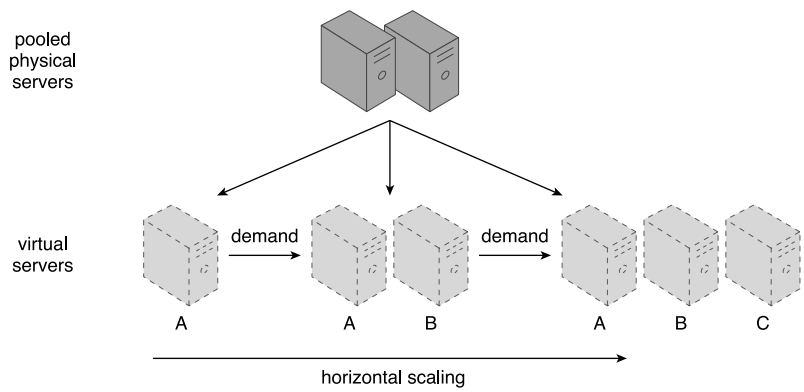



**Figure 3.4**

An IT resource (Virtual Server A) is scaled out by adding more of the same IT resources (Virtual Servers B and C).

### *Vertical Scaling*

When an existing IT resource is replaced by another with higher or lower capacity, *vertical scaling* is considered to have occurred (Figure 3.5). Specifically, the replacing of an IT resource with another that has a higher capacity is referred to as *scaling up* and the replacing an IT resource with another that has a lower capacity is considered *scaling down*. Vertical scaling is less common in cloud environments due to the downtime required while the replacement is taking place.

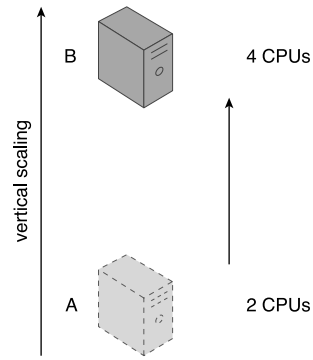



**Figure 3.5**
An IT resource (a virtual server with two CPUs) is scaled up by replacing it with a more powerful IT resource with increased capacity for data storage (a physical server with four CPUs).

Table 3.1 provides a brief overview of common pros and cons associated with horizontal and vertical scaling.

| Horizontal Scaling | Vertical Scaling |
|---|---|
| less expensive (through commodity hardware components) | more expensive (specialized servers) |
| IT resources instantly available | IT resources normally instantly available |
| resource replication and automated scaling | additional setup is normally needed |
| additional IT resources needed | no additional IT resources needed |
| not limited by hardware capacity | limited by maximum hardware capacity |

**Table 3.1**
A comparison of horizontal and vertical scaling.

## Cloud Service

Although a cloud is a remotely accessible environment, not all IT resources residing within a cloud can be made available for remote access. For example, a database or a physical server deployed within a cloud may only be accessible by other IT resources that are within the same cloud. A software program with a published API may be deployed specifically to enable access by remote clients.

A *cloud service* is any IT resource that is made remotely accessible via a cloud. Unlike other IT fields that fall under the service technology umbrella—such as service-oriented architecture—the term "service" within the context of cloud computing is especially broad. A cloud service can exist as a simple Web-based software program with a technical interface invoked via the use of a messaging protocol, or as a remote access point for administrative tools or larger environments and other IT resources.

In Figure 3.6, the yellow circle symbol is used to represent the cloud service as a simple Web-based software program. A different IT resource symbol may be used in the latter case, depending on the nature of the access that is provided by the cloud service.

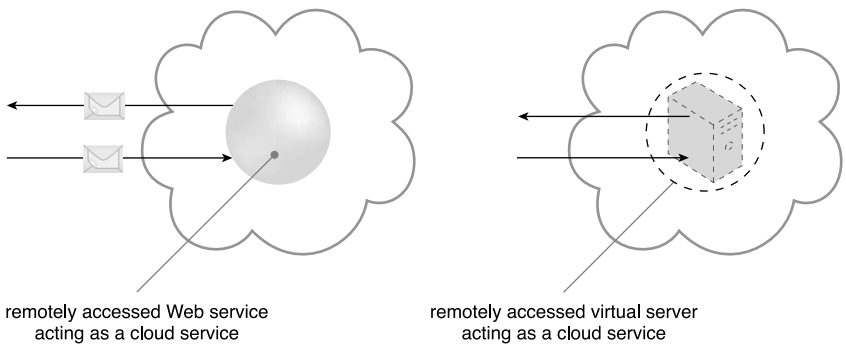



**Figure 3.6**

A cloud service with a published technical interface is being accessed by a consumer outside of the cloud (left). A cloud service that exists as a virtual server is also being accessed from outside of the cloud's boundary (right). The cloud service on the left is likely being invoked by a consumer program that was designed to access the cloud service's published technical interface. The cloud service on the right may be accessed by a human user that has remotely logged on to the virtual server.

The driving motivation behind cloud computing is to provide IT resources as services that encapsulate other IT resources, while offering functions for clients to use and leverage remotely. A multitude of models for generic types of cloud services have emerged, most of which are labeled with the "as-a-service" suffix.

**NOTE**

Cloud service usage conditions are typically expressed in a service-level agreement (SLA) that is the human-readable part of a service contract between a cloud provider and cloud consumer that describes QoS features, behaviors, and limitations of a cloud-based service or other provisions.

An SLA provides details of various measurable characteristics related to IT outcomes, such as uptime, security characteristics, and other specific QoS features, including availability, reliability, and performance. Since the implementation of a service is hidden from the cloud consumer, an SLA becomes a critical specification. SLAs are covered in detail in Chapter 16.

### Cloud Service Consumer

The *cloud service consumer* is a temporary runtime role assumed by a software program when it accesses a cloud service.

As shown in Figure 3.7, common types of cloud service consumers can include software programs and services capable of remotely accessing cloud services with published service contracts, as well as workstations, laptops and mobile devices running software capable of remotely accessing other IT resources positioned as cloud services.

**Figure 3.7**

Examples of cloud service consumers. Depending on the nature of a given diagram, an artifact labeled as a cloud service consumer may be a software program or a hardware device (in which case it is implied that it is running a software program capable of acting as a cloud service consumer).

## 3.3 Goals and Benefits

The common benefits associated with adopting cloud computing are explained in this section.

**NOTE**

The following sections make reference to the terms “public cloud” and “private cloud.” These terms are described in the *Cloud Deployment Models* section in Chapter 4.

### Reduced Investments and Proportional Costs

Similar to a product wholesaler that purchases goods in bulk for lower price points, public cloud providers base their business model on the mass-acquisition of IT resources that are then made available to cloud consumers via attractively priced leasing packages. This opens the door for organizations to gain access to powerful infrastructure without having to purchase it themselves.

The most common economic rationale for investing in cloud-based IT resources is in the reduction or outright elimination of up-front IT investments, namely hardware and software purchases and ownership costs. A cloud's Measured Usage characteristic represents a feature-set that allows measured operational expenditures (directly related to business performance) to replace anticipated capital expenditures. This is also referred to as *proportional costs*.

This elimination or minimization of up-front financial commitments allows enterprises to start small and accordingly increase IT resource allocation as required. Moreover, the reduction of up-front capital expenses allows for the capital to be redirected to the core business investment. In its most basic form, opportunities to decrease costs are derived from the deployment and operation of large-scale data centers by major cloud providers. Such data centers are commonly located in destinations where real estate, IT professionals, and network bandwidth can be obtained at lower costs, resulting in both capital and operational savings.

The same rationale applies to operating systems, middleware or platform software, and application software. Pooled IT resources are made available to and shared by multiple cloud consumers, resulting in increased or even maximum possible utilization. Operational costs and inefficiencies can be further reduced by applying proven practices and patterns for optimizing cloud architectures, their management, and their governance.

Common measurable benefits to cloud consumers include:

- On-demand access to pay-as-you-go computing resources on a short-term basis (such as processors by the hour), and the ability to release these computing resources when they are no longer needed.
- The perception of having unlimited computing resources that are available on demand, thereby reducing the need to prepare for provisioning.
- The ability to add or remove IT resources at a fine-grained level, such as modifying available storage disk space by single gigabyte increments.

- Abstraction of the infrastructure so applications are not locked into devices or locations and can be easily moved if needed.

For example, a company with sizable batch-centric tasks can complete them as quickly as their application software can scale. Using 100 servers for one hour costs the same as using one server for 100 hours. This "elasticity" of IT resources, achieved without requiring steep initial investments to create a large-scale computing infrastructure, can be extremely compelling.

Despite the ease with which many identify the financial benefits of cloud computing, the actual economics can be complex to calculate and assess. The decision to proceed with a cloud computing adoption strategy will involve much more than a simple comparison between the cost of leasing and the cost of purchasing. For example, the financial benefits of dynamic scaling and the risk transference of both over-provisioning (under-utilization) and under-provisioning (over-utilization) must also be accounted for. Chapter 15 explores common criteria and formulas for performing detailed financial comparisons and assessments.

**NOTE**

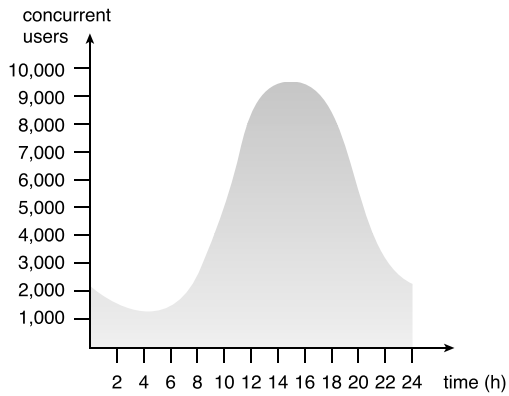
Another area of cost savings offered by clouds is the "as-a-service" usage model, whereby technical and operational implementation details of IT resource provisioning are abstracted from cloud consumers and packaged into "ready-to-use" or "off-the-shelf" solutions. These services-based products can simplify and expedite the development, deployment, and administration of IT resources when compared to performing equivalent tasks with on-premise solutions. The resulting savings in time and required IT expertise can be significant and can contribute to the justification of adopting cloud computing.

### Increased Scalability

By providing pools of IT resources, along with tools and technologies designed to leverage them collectively, clouds can instantly and dynamically allocate IT resources to cloud consumers, on-demand or via the cloud consumer's direct configuration. This empowers cloud consumers to scale their cloud-based IT resources to accommodate processing fluctuations and peaks automatically or manually. Similarly, cloud-based IT resources can be released (automatically or manually) as processing demands decrease.

A simple example of usage demand fluctuations throughout a 24 hour period is provided in Figure 3.8.

**Figure 3.8**
An example of an organization's changing demand for an IT resource over the course of a day.




The inherent, built-in feature of clouds to provide flexible levels of scalability to IT resources is directly related to the aforementioned proportional costs benefit. Besides the evident financial gain to the automated reduction of scaling, the ability of IT resources to always meet and fulfill unpredictable usage demands avoids potential loss of business that can occur when usage thresholds are met.

**NOTE**

When associating the benefit of Increased Scalability with the capacity planning strategies introduced earlier in the *Business Drivers* section, the Lag and Match Strategies are generally more applicable due to a cloud's ability to scale IT resources on-demand.

### Increased Availability and Reliability

The availability and reliability of IT resources are directly associated with tangible business benefits. Outages limit the time an IT resource can be "open for business" for its customers, thereby limiting its usage and revenue generating potential. Runtime failures that are not immediately corrected can have a more significant impact during high-volume usage periods. Not only is the IT resource unable to respond to customer requests, its unexpected failure can decrease overall customer confidence.

A hallmark of the typical cloud environment is its intrinsic ability to provide extensive support for increasing the availability of a cloud-based IT resource to minimize or even eliminate outages, and for increasing its reliability so as to minimize the impact of runtime failure conditions.

Specifically:

- An IT resource with increased availability is accessible for longer periods of time (for example, 22 hours out of a 24 hour day). Cloud providers generally offer "resilient" IT resources for which they are able to guarantee high levels of availability.
- An IT resource with increased reliability is able to better avoid and recover from exception conditions. The modular architecture of cloud environments provides extensive failover support that increases reliability.

It is important that organizations carefully examine the SLAs offered by cloud providers when considering the leasing of cloud-based services and IT resources. Although many cloud environments are capable of offering remarkably high levels of availability and reliability, it comes down to the guarantees made in the SLA that typically represent their actual contractual obligations.

### SUMMARY OF KEY POINTS

- Cloud environments are comprised of highly extensive infrastructure that offers pools of IT resources that can be leased using a pay-for-use model whereby only the actual usage of the IT resources is billable. When compared to equivalent on-premise environments, clouds provide the potential for reduced initial investments and operational costs proportional to measured usage.
- The inherent ability of a cloud to scale IT resources enables organizations to accommodate unpredictable usage fluctuations without being limited by pre-defined thresholds that may turn away usage requests from customers. Conversely, the ability of a cloud to decrease required scaling is a feature that relates directly to the proportional costs benefit.
- By leveraging cloud environments to make IT resources highly available and reliable, organizations are able to increase quality-of-service guarantees to customers and further reduce or avoid potential loss of business resulting from unanticipated runtime failures.

## 3.4 Risks and Challenges

Several of the most critical cloud computing challenges pertaining mostly to cloud consumers that use IT resources located in public clouds are presented and examined.

### Increased Security Vulnerabilities

The moving of business data to the cloud means that the responsibility over data security becomes shared with the cloud provider. The remote usage of IT resources requires an expansion of trust boundaries by the cloud consumer to include the external cloud. It can be difficult to establish a security architecture that spans such a trust boundary without introducing vulnerabilities, unless cloud consumers and cloud providers happen to support the same or compatible security frameworks—which is unlikely with public clouds.

Another consequence of overlapping trust boundaries relates to the cloud provider's privileged access to cloud consumer data. The extent to which the data is secure is now limited to the security controls and policies applied by both the cloud consumer and cloud provider. Furthermore, there can be overlapping trust boundaries from different cloud consumers due to the fact that cloud-based IT resources are commonly shared.

The overlapping of trust boundaries and the increased exposure of data can provide malicious cloud consumers (human and automated) with greater opportunities to attack IT resources and steal or damage business data. Figure 3.9 illustrates a scenario whereby two organizations accessing the same cloud service are required to extend their respective trust boundaries to the cloud, resulting in overlapping trust boundaries. It can be challenging for the cloud provider to offer security mechanisms that accommodate the security requirements of both cloud service consumers.

Overlapping trust boundaries is a security threat that is discussed in more detail in Chapter 6.

### Reduced Operational Governance Control

Cloud consumers are usually allotted a level of governance control that is lower than that over on-premise IT resources. This can introduce risks associated with how the cloud provider operates its cloud, as well as the external connections that are required for communication between the cloud and the cloud consumer.

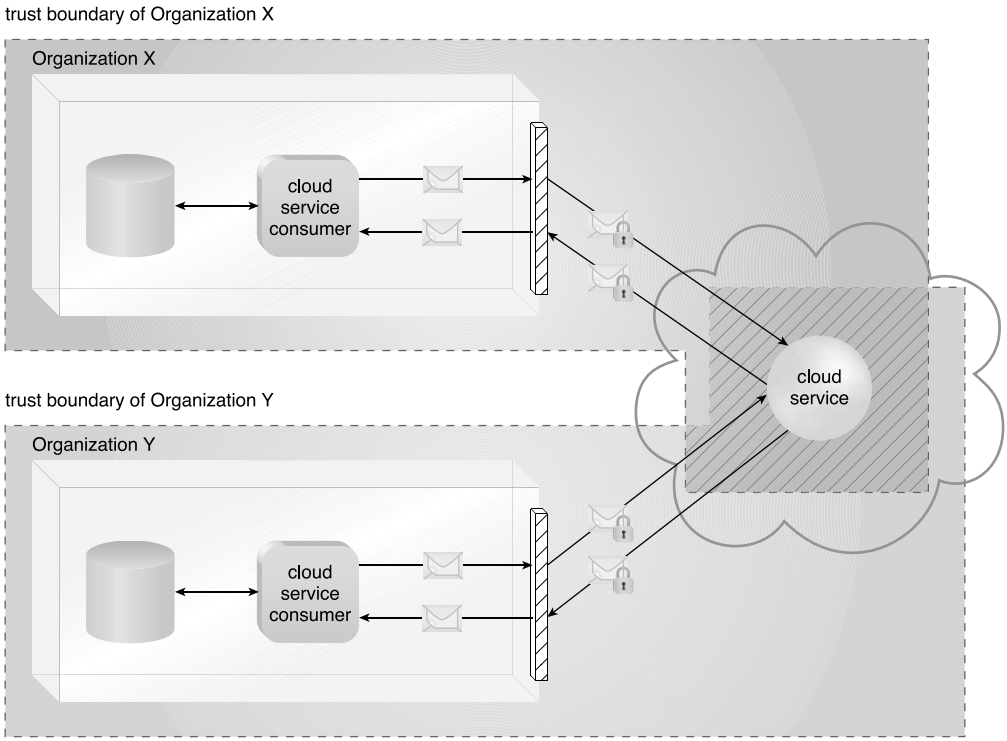



**Figure 3.9**
The shaded area with diagonal lines indicates the overlap of two organizations' trust boundaries.

Consider the following examples:

- An unreliable cloud provider may not maintain the guarantees it makes in the SLAs that were published for its cloud services. This can jeopardize the quality of the cloud consumer solutions that rely on these cloud services.
- Longer geographic distances between the cloud consumer and cloud provider can require additional network hops that introduce fluctuating latency and potential bandwidth constraints.

The latter scenario is illustrated in Figure 3.10.

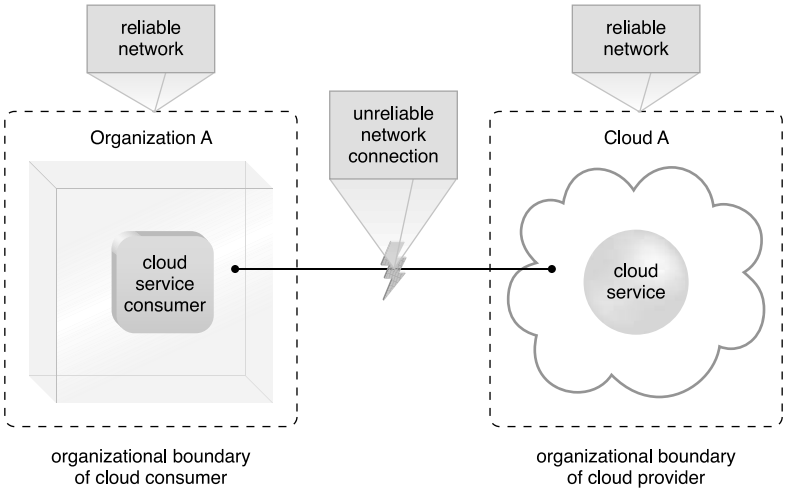



**Figure 3.10**
An unreliable network connection compromises the quality of communication between cloud consumer and cloud provider environments.

Legal contracts, when combined with SLAs, technology inspections, and monitoring, can mitigate governance risks and issues. A cloud governance system is established through SLAs, given the "as-a-service" nature of cloud computing. A cloud consumer must keep track of the actual service level being offered and the other warranties that are made by the cloud provider.

Note that different cloud delivery models offer varying degrees of operational control granted to cloud consumers, as further explained in Chapter 4.

### Limited Portability Between Cloud Providers

Due to a lack of established industry standards within the cloud computing industry, public clouds are commonly proprietary to various extents. For cloud consumers that have custom-built solutions with dependencies on these proprietary environments, it can be challenging to move from one cloud provider to another.

Portability is a measure used to determine the impact of moving cloud consumer IT resources and data between clouds (Figure 3.11).

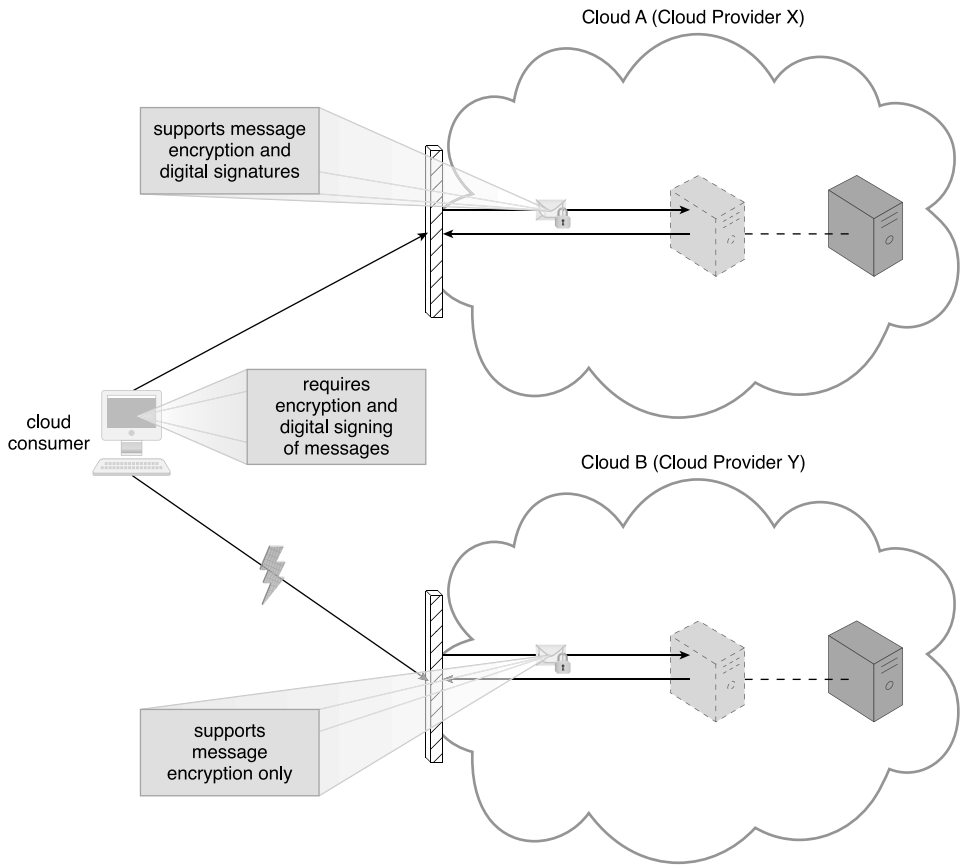



**Figure 3.11**

A cloud consumer's application has a decreased level of portability when assessing a potential migration from Cloud A to Cloud B, because the cloud provider of Cloud B does not support the same security technologies as Cloud A.

### Multi-Regional Compliance and Legal Issues

Third-party cloud providers will frequently establish data centers in affordable or convenient geographical locations. Cloud consumers will often not be aware of the physical location of their IT resources and data when hosted by public clouds. For some organizations, this can pose serious legal concerns pertaining to industry or government regulations that specify data privacy and storage policies. For example, some UK laws require personal data belonging to UK citizens to be kept within the United Kingdom.

Another potential legal issue pertains to the accessibility and disclosure of data. Countries have laws that require some types of data to be disclosed to certain government

agencies or to the subject of the data. For example, a European cloud consumer's data that is located in the U.S. can be more easily accessed by government agencies (due to the U.S. Patriot Act) when compared to data located in many European Union countries.

Most regulatory frameworks recognize that cloud consumer organizations are ultimately responsible for the security, integrity, and storage of their own data, even when it is held by an external cloud provider.

**SUMMARY OF KEY POINTS**

- Cloud environments can introduce distinct security challenges, some of which pertain to overlapping trust boundaries imposed by a cloud provider sharing IT resources with multiple cloud consumers.
- A cloud consumer's operational governance can be limited within cloud environments due to the control exercised by a cloud provider over its platforms.
- The portability of cloud-based IT resources can be inhibited by dependencies upon proprietary characteristics imposed by a cloud.
- The geographical location of data and IT resources can be out of a cloud consumer's control when hosted by a third-party cloud provider. This can introduce various legal and regulatory compliance concerns.

*This page intentionally left blank*

# Index


## A

**acceptable use policy (cloud provisioning contract), 452-453**
**active-active failover system (specialized mechanism), 191**
**active-passive failover system (specialized mechanism), 194**
**Advanced Research Projects Agency Network (ARPANET), 26**
**Advanced Telecom Networks (ATN) case study. *See* case study examples**
**agent**
- deployment, 310
- discovery, 310
- monitoring, 155
- polling, 155
- resource, 155
- service, 111
  - *malicious, 123*
- threat, 121-124

**anonymous attacker, 122**
**application**
- configuration baseline, 314
- layer protocol, 85
- multitenant, 106-108
- package, 313
- packager, 313
- subscription duration metric, 390-391
- usage, 370
- Web, 104-106

**architectures**
- bare-metal provisioning, 309-312
- cloud balancing, 299-301, 321-322
- cloud bursting, 271, 277-279
- cross-storage device vertical tiering, 332-337
- direct I/O access, 324-326
- direct LUN access, 326-327
- dynamic data normalization, 329-330
- dynamic failure detection and recovery, 306-309
- dynamic scalability, 262-264
- elastic disk provisioning, 272-274
- elastic network capacity, 330-332
- elastic resource capacity, 265-267
- hypervisor clustering, 282-287
- intra-storage device vertical data tiering, 337-339
- load balanced virtual server instances, 288-291
- load balanced virtual switches, 340-341
- multipath resource access, 342-343
- non-disruptive service relocation, 293-29

persistent virtual network configuration, 344-346
rapid provisioning, 312-315
redundant physical connection for virtual servers, 347-349
redundant storage, 275-277
resource pooling, 257-262
resource reservation, 301-305
service load balancing, 268-270
storage maintenance window, 350-356
storage workload management, 315-321
workload distribution, 256-257
zero downtime, 298-299
**ARPANET (Advanced Research Projects Agency Network), 26**
**"as-a-service" usage model, 42**
**asymmetric encryption (security mechanism), 231-232**
**asymmetric distribution, 176**
**ATN (Advanced Telecom Networks) case study. *See* case study examples**
**attack. *See* threat**
**attacker. *See* threat agent**
**audit monitor mechanism (specialized), 189-190**
**authentication**
IAM (identity and access management), 243
weak, 127
**authenticity (characteristic), 119**
**authorization**
IAM (identity and access management), 243
insufficient, 127
**automated scaling listener mechanism (specialized), 170-172**
**automation (data center), 91**
**autonomic computing, 444-445**
**availability (characteristic), 119**
data center, 92
IT resource, 43-44
**availability rate metric, 405-406**

## B

**bare-metal provisioning architecture, 309-312**
**billing management system mechanism (management), 225-227**
**boundary**
logical network perimeter, 58
organizational, 56
overlapping trust, 57
**broadband networks, 80-89**
**business case**
mapping to SLA, 413
template, 462-464
**business cost metrics, 380-387**
**business drivers, cloud computing, 28-30**

## C

**CA (certificate authority), 240**
**capacity planning, 28-29**
**capacity watchdog system, 289**
**carrier and external networks interconnection, 95**
**case study examples**
ATN (Advanced Telecom Networks), 14
*background, 14-16*
*business cost metrics, 382-387*
*cloud bursting architecture, 277-279*
*cloud security, 135*
*conclusion, 422*
*hashing, 235*
*IAM (identity and access management), 244*
*load balancer, 177*

*ready-made environment, 167-168*
*SSO (single sign-on), 246*
*state management database, 211-212*
DTGOV, 14
*automated scaling listener, 172*
*background, 16-19*
*billing management system, 227*
*cloud delivery model, 375-377*
*cloud storage device, 152-154*
*cloud usage monitor, 157-159*
*conclusion, 422-424*
*digital signature, 238*
*failover system, 196-198*
*hardened virtual server images, 252*
*hypervisor, 201*
*logical network perimeter, 142*
*service technologies, 113-115*
*pay-per-use monitor, 187*
*PKI (public key infrastructure), 242*
*pricing models, 396-401*
*remote administration system, 219*
*resource cluster, 206-207*
*resource management system, 221-222*
*resource replication, 162*
*resource segmentation, 249*
*SLA management system, 224*
*SLA monitor, 180-183*
*SLA template, 416-418*
*virtual server, 145-147*
Innovartus Technologies Inc., 14
*audit monitor, 189-190*
*background, 20-21*
*cloud balancing architecture, 321-322*
*conclusion, 424-425*
*encryption, 233*
*multi-device broker, 209*
*service quality metrics, 412-413*
**certificate authority (CA), 240**
**characteristics. *See* cloud characteristics**
**cipher, 230**
**ciphertext, 230**
**cloud architectures. *See* architectures**
**cloud characteristic, 58-63**
elasticity, 61
measured usage, 61
multitenancy, 59
resource pooling, 59
resiliency, 61
ubiquitous access, 59
mapped to cloud computing mechanisms, 434-435
supported by autonomic computing, 445
supported by service grids, 448
**cloud auditor (role), 56**
**cloud balancing architecture, 299-301**
Innovartus case study, 321-322
**cloud-based IT resource, 34**
usage cost metrics, 387-391
versus on-premise IT resource, 86-88
versus on-premise IT resource in private clouds, 76
**cloud-based security group mechanism (security), 247-249**
**cloud broker (role), 56**
**cloud bursting architecture, 271-272**
ATN case study, 277-279
**cloud carrier (role), 56**
selection, 89
**cloud computing, 27-28**
business drivers, 28-30
history, 26-27
mechanisms, mapped to cloud characteristics, 434-435
risks and challenges, 45-49
technology innovations, 31-33
terminology, 33-40

**cloud consumer (role), 36, 40, 52**
perspective in cloud delivery models, 370-375
**cloud delivery models, 63-73**
cloud consumer perspective, 370-375
cloud provider perspective, 360-370
combining, 69-73
comparing, 67-69
IaaS (Infrastructure-as-a-Service), 64
PaaS (Platform-as-a-Service), 65-66
SaaS (Software-as-a-Service), 66-67
**cloud deployment models, 73-78**
community, 74
hybrid, 77-78
inter, 78
private, 75-76
public, 73-74
virtual private, 78
**cloud-enabling technologies, 32**
**cloud mechanisms. *See* mechanisms**
**cloud provider (role), 36, 52**
perspective in cloud delivery models, 360-370
portability, 47
selection, 89, 458-459
**cloud provisioning contract, 450-457**
**cloud resource administrator (role), 54-56**
**Cloud Security Alliance (CSA), 429**
**cloud service, 38-39**
lifecycle phases, 391-392
**cloud service consumer (role), 40**
**cloud service owner (role), 53**
**cloud service usage cost metrics, 390-391**
**cloud storage device mechanism (infrastructure), 149-154**
in bare-metal provisioning architecture, 310
in multipath resource access architecture, 343
in storage maintenance window architecture, 350-356
usage cost metrics, 390
**cloud storage gateway, 209**
**cloud usage monitor mechanism (infrastructure), 155-159**
in cross-storage device vertical tiering architecture, 337
in direct I/O access architecture, 326
in direct LUN access architecture, 327
in dynamic scaling architecture, 264
in elastic disk provisioning architecture, 274
in elastic network capacity architecture, 331
in elastic resource capacity architecture, 265
in load balanced virtual switches architecture, 341
in non-disruptive service relocation architecture, 297
in resource pooling architecture, 260
in resource reservation architecture, 305
in service load balancing architecture, 268
in storage workload management architecture, 321
in workload distribution architecture, 257
in zero downtime architecture, 299
**clustering, 31-33**
**cluster**
database, 203
HA (high-availability), 205
large dataset, 204
load balanced, 205
resource, 203-207
server, 203
**community cloud, 74**

**completion time metric, 409**
**computational grid, 31**
**computer room (data center), 439**
**computing hardware, 93**
**confidentiality (characteristic), 118, 232**
**connectionless packet switching (datagram networks), 83**
**content-aware distribution, 176**
**cost(s)**
archiving, 396
integration, 381
locked-in, 381-382
management of, 391-396
of capital, 381
on-going, 380-381
proportional, 41-43, 61
reduction, 29-30
sunk, 381
up-front, 380
**CPU pool, 258**
**credential management, 243**
**cross-storage device vertical tiering architecture, 332-337**
**cryptography, 230-233**
**CSA (Cloud Security Alliance), 429**

## D

**database**
cluster, 203
state management, 210-212
storage interface, 151-152
**data block, 151**
**data center, 90-96**
automation, 91
component redundancy, 442
availability, 92
environmental controls, 440-441
facilities, 92-93, 437-442
hardware, 93-96
*computing, 93*
*network, 95-96*
*storage, 93-94*
persistence, 367
remote operation and management, 92
security awareness, 92
standardization and modularity, 90
technical and business considerations, 96
**data normalization, 152**
**data storage, 363, 151**
non-relational (NoSQL), 152
relational, 151
**datagram networks (connectionless packet switching), 83**
**dedicated cloud (virtual private cloud), 78**
**delivery models, 63-73**
**denial of service (DoS), 126**
**deployment agent, 310**
**deployment component, 310**
**deployment data store, 314**
**deployment models, 73-78**
**design constraints, REST, 111**
**design patterns, Web site, 10**
**digital signature mechanism (security), 236-238**
in PKI (public key infrastructure), 240-242
**direct I/O access architecture, 324-326**
**direct LUN access architecture, 326-327**
**discovery agent, 310**
**discovery section, 310**
**Distributed Management Task Force (DMTF), 429**
**DoS (denial of service), 126**
**DTGOV case study. *See* case study examples**
**dynamic data normalization architecture, 329-330**

**dynamic failure detection and recovery architecture, 306-309**
**dynamic horizontal scaling, 262**
**dynamic relocation, 262**
**dynamic scalability architecture, 262-264**
**dynamic vertical scaling, 262**

## E

**eavesdropping, traffic, 124**
**EDM (equipment distribution area) (data center), 439**
**Elastic Compute Cloud (EC2) services, 27**
**elastic disk provisioning architecture, 272-274**
**elasticity (cloud characteristic), 61**
mapped to cloud computing mechanisms, 435
supported by autonomic computing, 445
**elastic network capacity architecture, 330-332**
**elastic resource capacity architecture, 265-267**
**electrical power interconnections (data center), 440**
**electrical room (data center), 438**
**encryption mechanism (security), 230-233**
asymmetric, 231-232
symmetric, 231
**enterprise service bus (ESB) platform, 112**
**environmental controls (data center), 440-441**
**equipment distribution area (EDM) (data center), 439**
**errata, Web site, 9**
**ESB (enterprise service bus) platform, 112**
**European Telecommunications Standards Institute (ETSI), 431**
**event triggers, 364, 367**

## F

**failover system mechanism (specialized), 191-198**
active-active, 191
active-passive, 194
in dynamic failure detection architecture, 309
in redundant physical connection for virtual servers architecture, 349
in zero downtime architecture, 298-299
**failure conditions, 364, 367**
**fast data replication mechanisms, 94**
**figures (conventions), 9**
**flawed implementations (IT security), 131**

## G—H

**gateway**
cloud storage, 209
mobile device, 209
XML, 209
**grid computing, 31-33, 445-448**

**HA (high-availability), 406**
cluster, 205
**hard disk arrays, 94**
**hardened virtual server image mechanism (security), 251-252**
**hardware**
computing, 93
independence, 98
network, 95-96
obsolescence, 96
storage, 93-94
**hardware-based virtualization, 101**
**hashing mechanism (security), 234-235**
**HDM (horizontal distribution area) (data center), 439**
**heartbeats, 282**
**high-availability (HA). *See* (HA) high availability**

**history, cloud computing, 26-27**
**horizontal distribution area (HDM) (data center), 439**
**horizontal scaling, 37-38**
**hosted cloud (virtual private cloud), 78**
**host operating system, 99**
**host (physical server), 36**
**hot-swappable hard disks, 94**
**HTML, 104**
**HTTP (Hypertext Transfer Protocol), 104**
**HTTPS, 232**
**hybrid cloud, 77-78**
**hypermedia, 104**
**Hypertext Transfer Protocol (HTTP), 104**
**hypervisor mechanism (specialized), 97-98, 101, 200-201**
in bare-metal provisioning architecture, 310
in dynamic scaling architecture, 264
in elastic network capacity architecture, 331
in hypervisor clustering architecture, 282
in load balanced virtual switches architecture, 341
in multipath resource access architecture, 343
in persistent virtual network configuration architecture, 346
in redundant physical connection for virtual servers architecture, 349
in resource pooling architecture, 260
in resource reservation architecture, 305
in workload distribution architecture, 257
in zero downtime architecture, 299
**hypervisor clustering architecture, 282-287**

## I

**IaaS (Infrastructure-as-a-Service), 64-65**
cloud provider perspective of, 360-364
cloud consumer perspective of, 370-373
in combination with PaaS, 69-70
in combination with PaaS and SaaS, 72
in comparison with SaaS and PaaS, 67-69
pricing models, 394
**IAM (identity and access management) mechanism (security), 243-244**
**identity and access management (IAM) mechanism (security), 243-244**
**inbound network usage cost metric, 387-388**
**infrastructure redundancy summary, data center, 442**
**Innovartus Technologies Inc. case study. *See* case study examples**
**instance starting time metric, 409**
**insufficient authorization, 127**
**integration costs, 381**
**integrity (IT security), 119**
**intelligent automation engine, 265**
**inter-cloud, 78**
**International Service Technology Symposium conference series, 10**
**Internet**
architecture, 80-89
service provider (ISP), 80-83
versus cloud, 33-34
**internetworks (Internet), 80**
**intra-cloud WAN usage metric, 388**
**intra-storage device vertical data tiering architecture, 337-339**
**I/O**
caching, 94
data transferred metric, 390
**ISP (Internet service provider), 80-83**

**IT resource, 34-36**
cloud-based versus on-premise, 86-88
cloud-based versus on-premise, costs, 380-387
provisioning considerations
*of IaaS environments, 372-373*
*of PaaS environments, 373-374*
virtualization, 97-103
versus Web resource, 103

## J–K–L

**lag strategy (capacity planning), 29**
**LAN fabric, 95**
**large dataset cluster, 204**
**lead strategy (capacity planning), 29**
**Liberty Alliance, 432**
**live VM migration, 283**
**load balanced cluster, 205**
**load balanced virtual server instances architecture, 288-291**
**load balanced virtual switches architecture, 340-341**
**load balancer mechanism (specialized), 176-177**
in load balanced virtual server instances architecture, 290
in load balanced virtual switches architecture, 341
in service load balancing architecture, 268
in storage workload management architecture, 321-322
in workload distribution architecture, 257
**locked-in costs, 381**
**logical network perimeter mechanism (infrastructure), 58, 140-142**
in bare-metal provisioning architecture, 310
in direct I/O access architecture, 326
in elastic network capacity architecture, 332
in hypervisor clustering architecture, 288
in load balanced virtual server instances architecture, 291
in load balanced virtual switches architecture, 341
in multipath resource access architecture, 343
in persistent virtual network configuration architecture, 346
in redundant physical connection for virtual servers architecture, 349
in resource pooling architecture, 261
in resource reservation architecture, 305
in storage workload management architecture, 321
in workload distribution architecture, 257
in zero downtime architecture, 299
**logical unit number (LUN), 275**
**loose coupling, 445**
**LUN (logical unit number), 275**
in direct LUN access architecture, 326-327
migration, 315

## M

**main distribution area (MDA), 439**
**malicious insider, 123**
**malicious intermediary threat, 124-125**
**malicious service agent, 123**
**malicious tenant, 123**
**management loader, 310**
**markup languages, 104**
**match strategy (capacity planning), 29**
**MDA (main distribution area), 439**
**mean-time between failures (MTBF) metric, 407**

**mean-time system recovery (MTSR) metric, 412**
**mean-time to switchover (MTSO) metric, 411**
**measured usage (cloud characteristic), 61**
mapped to cloud computing mechanisms, 435
**mechanical room (data center), 438**
**mechanisms**
specialized, 170-212
*audit monitor, 189-190*
*automated scaling listener, 170-172*
*failover system, 191-199*
*hypervisor, 200-202*
*load balancer, 176-178*
*multi-device broker, 209-209*
*pay-per-use monitor, 184-188*
*resource cluster, 203-207*
*SLA monitor, 178-184*
*state management database, 210-212*
infrastructure, 140-186
*cloud storage device, 149-154*
*cloud usage monitor, 155-160*
*logical network perimeter, 140-143*
*ready-made environment, 166-168*
*resource replication, 161-165*
*virtual server, 144-147*
management, 214-227
*billing management system, 225-227*
*remote administration system, 214-219*
*resource management system, 219-222*
*SLA management system, 222-224*
security, 230-252
*cloud-based security groups, 247-250*
*digital signature, 236-239*
*encryption, 230-233*
*hardened virtual server images, 251-252*
*hashing, 234-235*
*identity and access management (IAM), 243-244*
*public key infrastructure (PKI), 240-242*
*single sign-on (SSO), 244-246*
**message digest, 234**
**metrics**
application subscription duration, 390-391
availability rate, 405-406
business cost, 380-387
completion time, 409
inbound network usage cost, 387-388
instance starting time, 409
intra-cloud WAN usage, 388
I/O data transferred, 390
mean-time between failures (MTBF), 407
mean-time system recovery (MTSR), 412
mean-time to switchover (MTSO), 411
network capacity, 408
network usage cost, 387-388
number of nominated users, 391
number of transactions users, 391
on-demand storage space allocation , 390
on-demand virtual machine instance allocation, 389
outage duration, 406
outbound network usage, 388
reserved virtual machine instance allocation, 389
response time, 409
server capacity, 408
service performance, 407-409

service quality, 404-413
service reliability, 407
service resiliency, 411-412
service scalability, 409-410
storage device capacity, 408
usage cost, 387-391
Web application capacity, 408-409
**middleware platforms, 112**
enterprise service bus (ESB), 112
orchestration, 112
**middleware, service, 112**
**migration**
LUN, 315
virtual server, 293-297
live VM, 283
**mobile device gateway, 209**
**model**
"as-a-service" usage, 42
delivery, 63-73, 375-377
deployment, 73-78, 370-375
pricing, 393-394, 396-401
**monitoring agent, 155**
**monitor**
audit, 189-190
cloud usage, 155-159
pay-per-use, 184-187
SLA, 178-183
**MTBF (mean-time between failures) metric, 407**
**MTSO (mean-time to switchover) metric, 411**
**MTSR (mean-time system recovery) metric, 412**
**multi-device broker mechanism (specialized), 208-209**
**multipath resource access architecture, 342-343**
**multitenancy, 59-61**
and resource pooling, 59-61
mapped to cloud computing mechanisms, 434
supported by service grids, 448
versus virtualization, 108
**multitenant application, 106-108**

## N

**NAS (network-attached storage), 94**
gateway, 95
**National Institute of Standards and Technology (NIST), 428**
**network-attached storage (NAS), 94**
**network capacity**
in elastic network capacity architecture, 330-332
metric, 408
**network hardware, 95-96**
**network pool, 258**
**network storage interface, 150-151**
**network traffic, 363**
**network usage, 367**
**network usage cost metrics, 387-388**
**NIST (National Institute of Standards and Technology), 428**
**non-disruptive service relocation architecture, 293-297**
**non-relational (NoSQL) data storage, 152**
**normalization, data, 152**
**NoSQL (non-relational) data storage, 152**
**notification service, 11**
**number of nominated users metric, 391**
**number of transactions users metric, 391**

## O

**OASIS (Organization for the Advancement of Structured Information Standards), 430**
**object storage interface, 151**
**OCC (Open Cloud Consortium), 431**
**office area (data center), 438**
**OGF (Open Grid Forum), 432**
**on-demand storage space allocation metric, 390**

**on-demand usage (cloud characteristic), 59, 434**
**on-demand virtual machine instance allocation metric, 389**
**on-going cost, 380-381**
**on-premise IT resource, 36**
versus cloud-based IT resource, 380-387
*in private cloud, 76*
**Open Cloud Consortium (OCC), 431**
**Open Grid Forum (OGF), 432**
**The Open Group, 430**
**operating system-based virtualization, 99-101**
**operating system baseline, 313**
**operations center (data center), 438**
**orchestration platform, 112**
**organizational agility, 30**
**organizational boundary, 56**
**Organization for the Advancement of Structured Information Standards (OASIS), 430**
**outage duration metric, 406**
**outbound network usage metric, 388**
**overlapping trust boundaries, 129-130**

## P

**PaaS (Platform-as-a-Service), 65-66**
cloud consumer perspective, 373-374
cloud provider perspective, 364-367
combination with IaaS, 69-70
combination with IaaS and SaaS, 72
comparison with IaaS and SaaS, 67-69
pricing models, 394
**pay-per-use monitor mechanism (specialized), 184-187**
in cross-storage device vertical tiering architecture, 337
in direct I/O access architecture, 326
in direct LUN access architecture, 327
in dynamic scaling architecture, 264
in elastic network capacity architecture, 332
in elastic resource capacity architecture, 265
in non-disruptive service relocation architecture, 297
in resource pooling architecture, 261
**performance overhead (virtualization), 102**
**persistent virtual network configuration architecture, 344-346**
**physical host, 36**
**physical network, 84**
**physical RAM pool, 258**
**physical server pool, 258**
**PKI (public key infrastructure) mechanism (security), 240-242**
**plaintext, 230**
**polling agent, 157**
**pool (resource), 258-259**
CPU, 258
network, 258
physical RAM, 258
physical server, 258
storage, 258
virtual server, 258
**portability**
cloud provider, 47
virtualization solution, 102
requirements, 464
**portal**
self-service, 215
usage and administration, 215
**power distribution system (data center), 441**
**power engine-generator, 441**
**power usage effectiveness (PUE), 441**
**pricing and billing (cloud provisioning contract), 457**
**pricing models, 393-394**
DTGOV case study, 396-401
**primary rooms (data center), 438-439**

**private cloud, 75-76**
**proportional costs, 41-42**
**public cloud, 73-74**
**public key cryptography, 231**
**public key identification, 240**
**public key infrastructure (PKI) mechanism (security), 240-242**
**PUE (power usage effectiveness), 441**

## Q-R

**quality of service (QoS), 404-413.** *See also* **SLA**

**rapid provisioning architecture, 312-315**
**ready-made environment mechanism (infrastructure), 166-168**
instances, 367
**recovery point objective (RPO), 457**
**recovery time objective (RTO), 457**
**reduction, cost, 29-30**
**redundant physical connection for virtual servers architecture, 347-349**
**redundant storage architecture, 275-277**
**relational data storage, 151**
**reliability rate metric, 407**
**remote administration system mechanism (management), 214-219**
in resource pooling architecture, 261
**remote operation and management (data center), 92**
**renewal (cloud provisioning contract), 456**
**reserved virtual machine instance allocation metric, 389**
**resiliency (cloud characteristic), 59, 61**
mapped to cloud computing mechanisms, 435
supported by autonomic computing, 445
supported by service grids, 448
**resilient watchdog system, 306**
**resource agent, 155**
**resource cluster mechanism (specialized), 203-207**
in service load balancing architecture, 268
in workload distribution architecture, 257
in zero downtime architecture, 299
**resource constraints, 301**
**resource management system mechanism (management), 219-222, 262**
**resource pool, 257-259**
**resource pooling (multitenancy), 59-61**
mapped to cloud computing mechanisms, 434
supported by service grids, 448
**resource pooling architecture, 257-262**
**resource replication mechanism (infrastructure), 161-162**
in bare-metal provisioning architecture, 312
in direct I/O access architecture, 326
in direct LUN access architecture, 327
in elastic disk provisioning architecture, 274
in elastic network capacity architecture, 332
in elastic resource capacity architecture, 265
in hypervisor clustering architecture, 288
in load balanced virtual server instances architecture, 291
in load balanced virtual switches architecture, 341
in multipath resource access architecture, 343
in non-disruptive service relocation architecture, 297

in persistent virtual network configuration architecture, 346
in redundant physical connection for virtual servers architecture, 349
in resource pooling architecture, 262
in resource reservation architecture, 305
in service load balancing architecture, 268
in storage maintenance window architecture, 356
in workload distribution architecture, 257
in zero downtime architecture, 299
**resource reservation architecture, 301-305**
**resource, Web, 103**
versus IT resource, 103
**resources, Web site, 9**
**response time metric, 409**
**REST service, 110**
**REST design constraints, 111**
**rights and responsibilities (cloud provisioning contract), 455-456**
**risk (IT security), 120**
**risk assessment, 133**
**risk control, 134**
**risk management, 133-134**
**risk treatment, 134**
**roles, 52-56**
cloud auditor, 56
cloud broker, 56
cloud carrier, 56
cloud consumer, 52-53
cloud provider, 52
cloud resource administrator, 54
cloud service owner, 53-54
**router-based interconnectivity, 83-85**
**RPO (recovery point objective), 457**
**RTO (recovery time objective), 457**

## S

**SaaS (Software-as-a-Service), 66-67**
cloud consumer perspective, 374-375
cloud provider perspective, 367-370
combination with IaaS and PaaS, 72
comparison with PaaS and IaaS, 67-69
pricing models, 394
**SAN (storage area network), 94**
**SAN fabric, 95**
**scalability**
cloud-based IT resource, 42-43
supported by multitenant applications, 107
supported by service grids, 448
**scaling, 37-38**
dynamic horizontal, 62
dynamic vertical, 62
horizontal, 37
vertical, 37-38
**secret key cryptography, 231**
**secure sockets layer (SSL), 232**
**security**
ATN case study, 135
controls, 120
mechanisms, 121
terminology, 118-121
**security policy, 121**
in cloud provisioning contracts, 453-455
disparity, 132
**self-adaptation, 444**
**self-configuration, 444**
**self-healing, 444**
**self-optimization, 444**
**self-protection, 444**
**self-service portal, 215**
**sequence logger, 313**
**sequence manager, 313**

**server**
capacity metric, 408
cluster, 203
consolidation, 98
images, 313
scalability (horizontal) metric, 410
scalability (vertical) metric, 410
templates, 312
usage, 389
virtual (physical host), 36
virtualization, 97, 144-147
**service, 108-112**
agent, 111
middleware, 112
REST, 110
Web, 109
Web-based, 108
**service agent, 111**
malicious, 123
**service availability metrics, 405-406**
**service grids, 446-448**
**service-level agreement.** ***See*** **SLA**
**service load balancing architecture, 268-270**
**service performance metrics, 407-409**
**service quality metrics, 404-413**
**service reliability metrics, 407**
**service resiliency metrics, 411-412**
**service scalability metrics, 409-410**
***Service Technology Magazine*****, 10**
**service usage (acceptable use) policy (cloud provisioning contract), 452-453**
**Simple Object Access Protocol (SOAP), 109**
**single sign-on (SSO) mechanism (security), 244-246**
**SLA management system mechanism (management), 222-224**
in bare-metal provisioning architecture, 312
in dynamic failure detection architecture, 309
in non-disruptive service relocation architecture, 297
**SLA monitor mechanism (specialized), 178-183**
in dynamic failure detection architecture, 309
in non-disruptive service relocation architecture, 297
**SLA (service-level agreement), 39, 404**
in cloud provisioning contract, 456-457
DTGOV case study, 416-418
guidelines, 413-415
**snapshotting, 94, 361**
**SNIA (Storage Networking Industry Association), 430**
**SOAP-based Web service, 109**
**SOAP, 109**
**Software-as-a-Service.** ***See*** **SaaS (Software-as-a-Service)**
**software, virtualization (hypervisor), 97-98, 101, 200-201**
**specifications (cloud provisioning contract), 456-457**
**SSL (secure sockets layer), 232**
**SSO (single sign-on) mechanism (security), 244-246**
**state management database mechanism (specialized), 210-212**
**storage**
hardware, 93-94
replication, 276
virtualization, 94, 97
**storage device, 149-154**
capacity metric, 408
levels, 149
usage, 390

**storage area network (SAN), 94**
**storage interface, 150-151**
database, 151-152
object, 151
network, 150
**storage maintenance window architecture, 350-356**
**Storage Networking Industry Association (SNIA), 430**
**storage pool, 258**
**storage room (data center), 438**
**storage workload management architecture, 315-321**
**sunk costs, 381**
**symbols (conventions), 9**
**symmetric encryption mechanism (security), 231**

## T

**telecommunications entrance (data center), 438**
**Telecommunications Industry Association (TIA), 431**
**tenant application functional module, 370**
**tenant subscription period, 370**
**termination (cloud provisioning contract), 456**
**terms of service (cloud provisioning contract), 452-456**
**threat, 120**
DoS (denial of service), 126
insufficient authorization, 127
malicious intermediary, 124-125
overlapping trust boundaries, 129-130
traffic eavesdropping, 124
virtualization attack, 127-129
**threat agent, 121-124**
anonymous attacker, 122
malicious insider, 123
malicious service, 123
trusted attacker, 123
**TIA (Telecommunications Industry Association), 431**
**TIA-942 Telecommunications Infrastructure Standard for Data Centers, 438**
**TLS (transport layer security), 232**
**traffic eavesdropping, 124**
**transport layer protocol, 84**
**transport layer security (TLS), 232**
**trust boundary, 57**
overlapping, 45, 129-130
**trusted attacker, 123**

## U

**ubiquitous access (cloud characteristic), 59**
mapped to cloud computing mechanisms, 434
**uniform resource locator (URL), 104**
**uninterruptible power source (UPS), 441**
**Universal Description, Discovery, and Integration (UDDI), 109**
**updates, Web site, 9**
**up-front costs, 380**
**UPS (uninterruptible power source), 441**
**URL (uniform resource locator), 104**
**usage and administration portal, 215**
**usage cost metrics, 387-391**
cloud service, 390-391
cloud storage device, 390
inbound network, 387-388
network, 387-388
server, 389
**user management, 243**
**utility computing, 2, 26**

## V

**vertical scaling, 37-38**
**VIM (virtual infrastructure manager), 219**
**virtual firewall, 141**
**virtual infrastructure manager (VIM), 219**
**virtual machine (VM), 97**
**virtual machine manager (VMM), 98**
**virtual machine monitor (VMM), 98**
**virtual network, 141**
**virtual private cloud, 78**
**virtual server mechanism (infrastructure), 144-147**
- images, hardened, 251-252
- in elastic network capacity architecture, 332
- in load balanced virtual server instances architecture, 288-291
- in load balanced virtual switches architecture, 341
- in non-disruptive service relocation architecture, 293-297
- in multipath resource access architecture, 343
- in persistent virtual network configuration architecture, 344-346
- in redundant physical connection for virtual servers architecture, 347-349
- in zero downtime architecture, 298-299
- lifecycles, 363

**virtual server pool, 258**
**virtual switch**
- in elastic network capacity architecture, 331
- in load balanced virtual switches architecture, 340-341
- in persistent virtual network configuration architecture, 344-346
- in redundant physical connection for virtual servers architecture, 347-349

**virtualization, 32, 90, 97-103**
- attack, 127-129
- hardware-based, 101
- operating system-based, 99-101
- management, 102
- software (hypervisor), 97-98, 101, 200-201
- storage, 94
- versus multitenancy, 108

**VIM (virtual infrastructure manager), 219**
**VM (virtual machine), 97**
**VMM (virtual machine manager), 98**
**volume cloning, 94**
**vulnerability (IT security), 120. *See also* threat**

## W

**weak authentication, 127**
**Web application, 104-106**
**Web application capacity metric, 408-409**
**Web-based**
- resource, 372
- service, 108

**Web resource, 103**
**Web Service Description Language (WSDL), 109**
**Web service, 109**
**SOAP-based, 109**
**Web sites**
- errata, 9
- resources, 9
- updates, 9
- www.cloudpatterns.org, 10
- www.cloudschool.com, 11
- www.cloudsecurityalliance.org, 429
- www.dmtf.org, 429
- www.nist.gov, 428
- www.oasis-open.org, 430
- www.ogf.org, 432

www.opencloudconsortium.org, 431
www.opengroup.org, 430
www.projectliberty.org, 432
www.serviceorientation.com, 11
www.servicetechbooks.com, 9, 11, 109, 111, 364
www.servicetechmag.com, 10
www.servicetechspecs.com, 10, 105
www.servicetechsymposium.com, 10
www.snia.org, 430
www.tiaonline.org, 432
www.whatiscloud.com, 10
www.whatisrest.com, 10, 111
**Web technology, 103-106**
**Web-tier load balancing, 95**
**workload distribution architecture, 256-257**
**workload prioritization, 176**
**WSDL (Web Service Description Language), 109**

## X-Z

**XML, 104, 109**
**XML gateway, 209**
**XML Schema Definition Language, 109**

**zero downtime architecture, 298-299**