

RELIABILITY AND AVAILABILITY OF CLOUD COMPUTING

ERIC
BAUER

RANDEE
ADAMS



 WILEY

 **IEEE**
IEEE PRESS

Table of Contents

COVER

IEEE PRESS

TITLE PAGE

COPYRIGHT PAGE

DEDICATION

FIGURES

TABLES

EQUATIONS

INTRODUCTION

AUDIENCE

ORGANIZATION

ACKNOWLEDGMENTS

I: BASICS

1 CLOUD COMPUTING

1.1 ESSENTIAL CLOUD CHARACTERISTICS

1.2 COMMON CLOUD CHARACTERISTICS

1.3 BUT WHAT, EXACTLY, IS CLOUD COMPUTING?

1.4 SERVICE MODELS

1.5 CLOUD DEPLOYMENT MODELS

1.6 ROLES IN CLOUD COMPUTING

1.7 BENEFITS OF CLOUD COMPUTING

1.8 RISKS OF CLOUD COMPUTING

2 VIRTUALIZATION

2.1 BACKGROUND

2.2 WHAT IS VIRTUALIZATION?

2.3 SERVER VIRTUALIZATION

2.4 VM LIFECYCLE

2.5 RELIABILITY AND AVAILABILITY RISKS OF
VIRTUALIZATION

3 SERVICE RELIABILITY AND SERVICE AVAILABILITY

3.1 ERRORS AND FAILURES

3.2 EIGHT-INGREDIENT FRAMEWORK

3.3 SERVICE AVAILABILITY

3.4 SERVICE RELIABILITY

3.5 SERVICE LATENCY

3.6 REDUNDANCY AND HIGH AVAILABILITY

3.7 HIGH AVAILABILITY AND DISASTER RECOVERY

3.8 STREAMING SERVICES

3.9 RELIABILITY AND AVAILABILITY RISKS OF
CLOUD COMPUTING

II: ANALYSIS

4 ANALYZING CLOUD RELIABILITY AND AVAILABILITY

4.1 EXPECTATIONS FOR SERVICE RELIABILITY AND AVAILABILITY

4.2 RISKS OF ESSENTIAL CLOUD CHARACTERISTICS

4.3 IMPACTS OF COMMON CLOUD CHARACTERISTICS

4.4 RISKS OF SERVICE MODELS

4.5 IT SERVICE MANAGEMENT AND AVAILABILITY RISKS

4.6 OUTAGE RISKS BY PROCESS AREA

4.7 FAILURE DETECTION CONSIDERATIONS

4.8 RISKS OF DEPLOYMENT MODELS

4.9 EXPECTATIONS OF IAAS DATA CENTERS

5 RELIABILITY ANALYSIS OF VIRTUALIZATION

5.1 RELIABILITY ANALYSIS TECHNIQUES

5.2 RELIABILITY ANALYSIS OF VIRTUALIZATION TECHNIQUES

5.3 SOFTWARE FAILURE RATE ANALYSIS

5.4 RECOVERY MODELS

5.5 APPLICATION ARCHITECTURE STRATEGIES

5.6 AVAILABILITY MODELING OF VIRTUALIZED RECOVERY OPTIONS

6 HARDWARE RELIABILITY, VIRTUALIZATION, AND SERVICE

AVAILABILITY

6.1 HARDWARE DOWNTIME EXPECTATIONS

6.2 HARDWARE FAILURES

6.3 HARDWARE FAILURE RATE

6.4 HARDWARE FAILURE DETECTION

6.5 HARDWARE FAILURE CONTAINMENT

6.6 HARDWARE FAILURE MITIGATION

6.7 MITIGATING HARDWARE FAILURES VIA
VIRTUALIZATION

6.8 VIRTUALIZED NETWORKS

6.9 MTTR OF VIRTUALIZED HARDWARE

6.10 DISCUSSION

7 CAPACITY AND ELASTICITY

7.1 SYSTEM LOAD BASICS

7.2 OVERLOAD, SERVICE RELIABILITY, AND
SERVICE AVAILABILITY

7.3 TRADITIONAL CAPACITY PLANNING

7.4 CLOUD AND CAPACITY

7.5 MANAGING ONLINE CAPACITY

7.6 CAPACITY-RELATED SERVICE RISKS

7.7 CAPACITY MANAGEMENT RISKS

7.8 SECURITY AND SERVICE AVAILABILITY

7.9 ARCHITECTING FOR ELASTIC GROWTH AND
DEGROWTH

8 SERVICE ORCHESTRATION ANALYSIS

8.1 SERVICE ORCHESTRATION DEFINITION

8.2 POLICY-BASED MANAGEMENT

8.3 CLOUD MANAGEMENT

8.4 SERVICE ORCHESTRATION'S ROLE IN RISK MITIGATION

8.5 SUMMARY

9 GEOGRAPHIC DISTRIBUTION, GEOREDUNDANCY, AND DISASTER RECOVERY

9.1 GEOGRAPHIC DISTRIBUTION VERSUS GEOREDUNDANCY

9.2 TRADITIONAL DISASTER RECOVERY

9.3 VIRTUALIZATION AND DISASTER RECOVERY

9.4 CLOUD COMPUTING AND DISASTER RECOVERY

9.5 GEOREDUNDANCY RECOVERY MODELS

9.6 CLOUD AND TRADITIONAL COLLATERAL BENEFITS OF GEOREDUNDANCY

9.7 DISCUSSION

III: RECOMMENDATIONS

10 APPLICATIONS, SOLUTIONS, AND ACCOUNTABILITY

10.1 APPLICATION CONFIGURATION SCENARIOS

10.2 APPLICATION DEPLOYMENT SCENARIO

10.3 SYSTEM DOWNTIME BUDGETS

10.4 END-TO-END SOLUTIONS CONSIDERATIONS

10.5 ATTRIBUTABILITY FOR SERVICE IMPAIRMENTS

10.6 SOLUTION SERVICE MEASUREMENT

10.7 MANAGING RELIABILITY AND SERVICE OF CLOUD COMPUTING

11 RECOMMENDATIONS FOR ARCHITECTING A RELIABLE SYSTEM

11.1 ARCHITECTING FOR VIRTUALIZATION AND CLOUD

11.2 DISASTER RECOVERY

11.3 IT SERVICE MANAGEMENT CONSIDERATIONS

11.4 MANY DISTRIBUTED CLOUDS VERSUS FEWER HUGE CLOUDS

11.5 MINIMIZING HARDWARE-ATTRIBUTED DOWNTIME

11.6 ARCHITECTURAL OPTIMIZATIONS

12 DESIGN FOR RELIABILITY OF VIRTUALIZED APPLICATIONS

12.1 DESIGN FOR RELIABILITY

12.2 TAILORING DFR FOR VIRTUALIZED APPLICATIONS

12.3 RELIABILITY REQUIREMENTS

12.4 QUALITATIVE RELIABILITY ANALYSIS

12.5 QUANTITATIVE RELIABILITY BUDGETING AND MODELING

12.6 ROBUSTNESS TESTING

12.7 STABILITY TESTING

12.8 FIELD PERFORMANCE ANALYSIS

12.9 RELIABILITY ROADMAP

12.10 HARDWARE RELIABILITY

13 DESIGN FOR RELIABILITY OF CLOUD SOLUTIONS

13.1 SOLUTION DESIGN FOR RELIABILITY

13.2 SOLUTION SCOPE AND EXPECTATIONS

13.3 RELIABILITY REQUIREMENTS

13.4 SOLUTION MODELING AND ANALYSIS

13.5 ELEMENT RELIABILITY DILIGENCE

13.6 SOLUTION TESTING AND VALIDATION

13.7 TRACK AND ANALYZE FIELD PERFORMANCE

13.8 OTHER SOLUTION RELIABILITY DILIGENCE

TOPICS

14 SUMMARY

14.1 SERVICE RELIABILITY AND SERVICE AVAILABILITY

14.2 FAILURE ACCOUNTABILITY AND CLOUD COMPUTING

14.3 FACTORING SERVICE DOWNTIME

14.4 SERVICE AVAILABILITY MEASUREMENT POINTS

14.5 CLOUD CAPACITY AND ELASTICITY CONSIDERATIONS

14.6 MAXIMIZING SERVICE AVAILABILITY

14.7 RELIABILITY DILIGENCE

14.8 CONCLUDING REMARKS

ABBREVIATIONS

REFERENCES

[ABOUT THE AUTHORS](#)

[INDEX](#)

IEEE Press
445 Hoes Lane
Piscataway, NJ 08854
IEEE Press Editorial Board 2012
John Anderson, *Editor in Chief*

Ramesh Abhari	Bernhard M. Haemmerli	Saeid Nahavandi
George W. Arnold	David Jacobson	Tariq Samad
Flavio Canavero	Mary Lanzerotti	George Zobrist
Dmitry Goldgof	Om P. Malik	

Kenneth Moore, *Director of IEEE Book and Information Services (BIS)*

Technical Reviewers

Xuemei Zhang

Principal Member of Technical Staff

Network Design and Performance Analysis

AT&T Labs

Rocky Heckman, CISSP

Architect Advisor

Microsoft

RELIABILITY AND AVAILABILITY OF CLOUD COMPUTING

Eric Bauer
Randee Adams



IEEE PRESS



A JOHN WILEY & SONS, INC., PUBLICATION

cover image: © iStockphoto

cover design: Michael Rutkowski

ITIL® is a Registered Trademark of the Cabinet Office in the United Kingdom and other countries.

Copyright © 2012 by the Institute of Electrical and Electronics Engineers. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.

Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permissions>.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable

for any loss of profit or any other commercial damages,
including but not limited to special, incidental,
consequential, or other damages.

For general information on our other products and services
or for technical support, please contact our Customer Care
Department within the United States at (800) 762-2974,
outside the United States at (317) 572-3993 or fax (317)
572-4002.

Wiley also publishes its books in a variety of electronic
formats. Some content that appears in print may not be
available in electronic formats. For more information about
Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Bauer, Eric.

Reliability and availability of cloud computing / Eric Bauer,
Randee Adams.

p. cm.

ISBN 978-1-118-17701-3 (hardback)

1. Cloud computing. 2. Computer software-
Reliability. 3. Computer software-Quality
control. 4. Computer security. I. Adams, Randee. II. Title.

QA76.585.B394 2012

004.6782-dc23

2011052839

*To our families and friends for their continued
encouragement and support.*

FIGURES

Figure 1.1	Service Models
Figure 1.2	OpenCrowd's Cloud Taxonomy
Figure 1.3	Roles in Cloud Computing
Figure 2.1	Virtualizing Resources
Figure 2.2	Type 1 and Type 2 Hypervisors
Figure 2.3	Full Virtualization
Figure 2.4	Paravirtualization
Figure 2.5	Operating System Virtualization
Figure 2.6	Virtualized Machine Lifecycle State Transitions
Figure 3.1	Fault Activation and Failures
Figure 3.2	Minimum Chargeable Service Disruption
Figure 3.3	Eight-Ingredient ("8i") Framework
Figure 3.4	Eight-Ingredient plus Data plus Disaster (8i + 2d) Model
Figure 3.5	MTBF and MTTR
Figure 3.6	Service and Network Element Impact Outages of Redundant Systems
Figure 3.7	Sample DSL Solution
Figure 3.8	Transaction Latency Distribution for Sample Service
Figure 3.9	Requirements Overlaid on Service Latency Distribution for Sample Solution
Figure 3.10	Maximum Acceptable Service Latency
Figure 3.11	Downtime of Simplex Systems
Figure 3.12	Downtime of Redundant Systems
Figure 3.13	Simplified View of High Availability
Figure 3.14	High Availability Example
Figure 3.15	Disaster Recovery Objectives
Figure 3.16	ITU-T G.114 Bearer Delay Guideline
Figure 4.1	TL 9000 Outage Attributability Overlaid on Augmented 8i + 2d Framework
Figure 4.2	Outage Responsibilities Overlaid on Cloud 8i + 2d Framework
Figure 4.3	ITIL Service Management Visualization
Figure 4.4	IT Service Management Activities to Minimize Service Availability Risk

Figure 4.5	8i + 2d Attributability by Process or Best Practice Areas
Figure 4.6	Traditional Error Vectors
Figure 4.7	IaaS Provider Responsibilities for Traditional Error Vectors
Figure 4.8	Software Supplier (and SaaS) Responsibilities for Traditional Error Vectors
Figure 5.1	Sample Reliability Block Diagram
Figure 5.2	Traversal of Sample Reliability Block Diagram
Figure 5.3	Nominal System Reliability Block Diagram
Figure 5.4	Reliability Block Diagram of Full virtualization
Figure 5.5	Reliability Block Diagram of OS Virtualization
Figure 5.6	Reliability Block Diagram of Paravirtualization
Figure 5.7	Reliability Block Diagram of Coresident Application Deployment
Figure 5.8	Canonical Virtualization RBD
Figure 5.9	Latency of Traditional Recovery Options
Figure 5.10	Traditional Active-Standby Redundancy via Active VM Virtualization
Figure 5.11	Reboot of a Virtual Machine
Figure 5.12	Reset of a Virtual Machine
Figure 5.13	Redundancy via Paused VM Virtualization
Figure 5.14	Redundancy via Suspended VM Virtualization
Figure 5.15	Nominal Recovery Latency of Virtualized and Traditional Options
Figure 5.16	Server Consolidation Using Virtualization
Figure 5.17	Simplified Simplex State Diagram
Figure 5.18	Downtime Drivers for Redundancy Pairs
Figure 6.1	Hardware Failure Rate Questions
Figure 6.2	Application Reliability Block Diagram with Virtual Devices
Figure 6.3	Virtual CPU
Figure 6.4	Virtual NIC
Figure 7.1	Sample Application Resource Utilization by Time of Day
Figure 7.2	Example of Extraordinary Event Traffic Spike
Figure 7.3	The Slashdot Effect: Traffic Load Over Time (in Hours)
Figure 7.4	Offered Load, Service Reliability, and Service Availability of a Traditional System
Figure 7.5	Visualizing VM Growth Scenarios
Figure 7.6	Nominal Capacity Model
Figure 7.7	Implementation Architecture of Compute Capacity Model

Figure 7.8	Orderly Reconfiguration of the Capacity Model
Figure 7.9	Slew Rate of Square Wave Amplification
Figure 7.10	Slew Rate of Rapid Elasticity
Figure 7.11	Elasticity Timeline by ODCA SLA Level
Figure 7.12	Capacity Management Process
Figure 7.13	Successful Cloud Elasticity
Figure 7.14	Elasticity Failure Model
Figure 7.15	Virtualized Application Instance Failure Model
Figure 7.16	Canonical Capacity Management Failure Scenarios
Figure 7.17	ITU X.805 Security Dimensions, Planes, and Layers
Figure 7.18	Leveraging Security and Network Infrastructure to Mitigate Overload Risk
Figure 8.1	Service Orchestration
Figure 8.2	Example of Cloud Bursting
Figure 10.1	Canonical Single Data Center Application Deployment Architecture
Figure 10.2	RBD of Sample Application on Blade-Based Server Hardware
Figure 10.3	RBD of Sample Application on IaaS Platform
Figure 10.4	Sample End-to-End Solution
Figure 10.5	Sample Distributed Cloud Architecture
Figure 10.6	Sample Recovery Scenario in Distributed Cloud Architecture
Figure 10.7	Simplified Responsibilities for a Canonical Cloud Application
Figure 10.8	Recommended Cloud-Related Service Availability Measurement Points
Figure 10.9	Canonical Example of MP 1 and MP 2
Figure 10.10	End-to-End Service Availability Key Quality Indicators
Figure 11.1	Virtual Machine Live Migration
Figure 11.2	Active-Standby Markov Model
Figure 11.3	Pie Chart of Canonical Hardware Downtime Prediction
Figure 11.4	RBD for the Hypothetical Web Server Application
Figure 11.5	Horizontal Growth of Hypothetical Application
Figure 11.6	Outgrowth of Hypothetical Application
Figure 11.7	Aggressive Protocol Retry Strategy
Figure 11.8	Data Replication of Hypothetical Application
Figure 11.9	Disaster Recovery of Hypothetical Application
Figure 11.10	Optimal Availability Architecture of Hypothetical Application
Figure 12.1	Traditional Design for Reliability Process

Figure 12.2	Mapping Virtual Machines across Hypervisors
Figure 12.3	A Virtualized Server Failure Scenario
Figure 12.4	Robustness Testing Vectors for Virtualized Applications
Figure 12.5	System Design for Reliability as a Deming Cycle
Figure 13.1	Solution Design for Reliability
Figure 13.2	Sample Solution Scope and KQI Expectations
Figure 13.3	Sample Cloud Data Center RBD
Figure 13.4	Estimating MP 2
Figure 13.5	Modeling Cloud-Based Solution with Client-Initiated Recovery Model
Figure 13.6	Client-Initiated Recovery Model
Figure 14.1	Failure Impact Duration and High Availability Goals
Figure 14.2	Eight-Ingredient Plus Data Plus Disaster (8i + 2d) Model
Figure 14.3	Traditional Outage Attributability
Figure 14.4	Sample Outage Accountability Model for Cloud Computing
Figure 14.5	Outage Responsibilities of Cloud by Process
Figure 14.6	Measurement Pointss (MPs) 1, 2, 3, and 4
Figure 14.7	Design for Reliability of Cloud-Based Solutions

TABLES

Table 2.1	Comparison of Server Virtualization Technologies
Table 2.2	Virtual Machine Lifecycle Transitions
Table 3.1	Service Availability and Downtime Ratings
Table 3.2	Mean Opinion Scores
Table 4.1	ODCA's Data Center Classification
Table 4.2	ODCA's Data Center Service Availability Expectations by Classification
Table 5.1	Example Failure Mode Effects Analysis
Table 5.2	Failure Mode Effect Analysis Figure for Coresident Applications
Table 5.3	Comparison of Nominal Software Availability Parameters
Table 6.1	Example of Hardware Availability as a Function of MTTR/MTTRS
Table 7.1	ODCA IaaS Elasticity Objectives
Table 9.1	ODCA IaaS Recoverability Objectives
Table 10.1	Sample Traditional Five 9's Downtime Budget
Table 10.2	Sample Basic Virtualized Five 9's Downtime Budget
Table 10.3	Canonical Application-Attributable Cloud-Based Five 9's Downtime Budget
Table 10.4	Evolution of Sample Downtime Budgets
Table 11.1	Example Service Transition Activity Failure Mode Effect Analysis
Table 11.2	Canonical Hardware Downtime Prediction
Table 11.3	Summary of Hardware Downtime Mitigation Techniques for Cloud Computing
Table 12.1	Sample Service Latency and Reliability Requirements at MP 2
Table 13.1	Sample Solution Latency and Reliability Requirements
Table 13.2	Modeling Input Parameters
Table 14.1	Evolution of Sample Downtime Budgets

EQUATIONS

Equation 3.1	Basic Availability Formula
Equation 3.2	Practical System Availability Formula
Equation 3.3	Standard Availability Formula
Equation 3.4	Estimation of System Availability from MTBF and MTTR
Equation 3.5	Recommended Service Availability Formula
Equation 3.6	Sample Partial Outage Calculation
Equation 3.7	Service Reliability Formula
Equation 3.8	DPM Formula
Equation 3.9	Converting DPM to Service Reliability
Equation 3.10	Converting Service Reliability to DPM
Equation 3.11	Sample DPM Calculation
Equation 6.1	Availability as a Function of MTBF/MTTR
Equation 11.1	Maximum Theoretical Availability across Redundant Elements
Equation 11.2	Maximum Theoretical Service Availability

INTRODUCTION

Cloud computing is a new paradigm for delivering information services to end users, offering distinct advantages over traditional IS/IT deployment models, including being more economical and offering a shorter time to market. Cloud computing is defined by a handful of essential characteristics: on-demand self service, broad network access, resource pooling, rapid elasticity, and measured service. Cloud providers offer a variety of service models, including infrastructure as a service, platform as a service, and software as a service; and cloud deployment options include private cloud, community cloud, public cloud and hybrid clouds. End users naturally expect services offered via cloud computing to deliver at least the same service reliability and service availability as traditional service implementation models. This book analyzes the risks to cloud-based application deployments achieving the same service reliability and availability as traditional deployments, as well as opportunities to improve service reliability and availability via cloud deployment. We consider the service reliability and service availability risks from the fundamental definition of cloud computing—the essential characteristics—rather than focusing on any particular virtualization hypervisor software or cloud service offering. Thus, the insights of this higher level analysis and the recommendations should apply to all cloud service offerings and application deployments. This book also offers recommendations on architecture, testing, and engineering diligence to assure that cloud deployed applications meet users' expectations for service reliability and service availability.

Virtualization technology enables enterprises to move their existing applications from traditional deployment scenarios in which applications are installed directly on native hardware to more evolved scenarios that include hardware independence and server consolidation. Use of virtualization technology is a common characteristic of cloud computing that enables cloud service providers to better manage usage of their resource pools by multiple cloud consumers. This book also considers the reliability and availability risks along this evolutionary path to guide enterprises planning the evolution of their application to virtualization and on to full cloud computing enablement over several releases.

AUDIENCE

The book is intended for IS/IT system and solution architects, developers, and engineers, as well as technical sales, product management, and quality management professionals.

ORGANIZATION

The book is organized into three parts: *Part I, "Basics," Part II, "Analysis,"* and *Part III—,"Recommendations."* Part I, "Basics," defines key terms and concepts of cloud computing, virtualization, service reliability, and service availability. Part I contains three chapters:

- *Chapter 1, "Cloud Computing."* This book uses the cloud terminology and taxonomy defined by the U.S. National Institute of Standards and Technology. This chapter defines cloud computing and reviews the essential and common characteristics of cloud computing. Standard service and deployment models of cloud computing are

reviewed, as well as roles of key cloud-related actors. Key benefits and risks of cloud computing are summarized.

- *Chapter 2, “Virtualization.”* Virtualization is a common characteristic of cloud computing. This chapter reviews virtualization technology, offers architectural models for virtualization that will be analyzed, and compares and contrasts “virtualized” applications to “native” applications.
- *Chapter 3, “Service Reliability and Service Availability.”* This chapter defines service reliability and availability concepts, reviews how those metrics are measured in traditional deployments, and how they apply to virtualized and cloud based deployments. As the telecommunications industry has very precise standards for quantification of service availability and service reliability measurements, concepts and terminology from the telecom industry will be presented in this chapter and used in Part II, “Analysis,” and Part III, “Recommendations.”

Part II, “Analysis,” methodically analyzes the service reliability and availability risks inherent in application deployments on cloud computing and virtualization technology based on the essential and common characteristics given in Part I.

- *Chapter 4, “Analyzing Cloud Reliability and Availability.”* Considers the service reliability and service availability risks that are inherent to the essential and common characteristics, service model, and deployment model of cloud computing. This includes implications of service transition activities, elasticity, and service orchestration. Identified risks are analyzed in detail in subsequent chapters in Part II.
- *Chapter 5, “Reliability Analysis of Virtualization.”* Analyzes full virtualization, OS

virtualization, paravirtualization, and server virtualization and coresidency using standard reliability analysis methodologies. This chapter also analyzes the software reliability risks of virtualization and cloud computing.

- *Chapter 6, “Hardware Reliability, Virtualization, and Service Availability.”* This chapter considers how hardware reliability risks and responsibilities shift as applications migrate to virtualized and cloud-based hardware platforms, and how hardware attributed service downtime is determined.
- *Chapter 7, “Capacity and Elasticity.”* The essential cloud characteristic of rapid elasticity enables cloud consumers to dispense with the business risk of locking-in resources weeks or months ahead of demand. Rapid elasticity does, however, introduce new risks to service quality, reliability, and availability that must be carefully managed.
- *Chapter 8, “Service Orchestration Analysis.”* Service orchestration automates various aspects of IT service management, especially activities associated with capacity management. This chapter reviews policy-based management in the context of cloud computing and considers the associated risks to service reliability and service availability.
- *Chapter 9, “Geographic Distribution, Georedundancy, and Disaster Recovery.”* Geographic distribution of application instances is a common characteristic of cloud computing and a best practice for disaster recovery. This chapter considers the service availability implications of georedundancy on applications deployed in clouds.

Part III, “Recommendations,” considers techniques to maximize service reliability and service availability of applications deployed on clouds, as well as the design for

reliability diligence to assure that virtualized applications and cloud based solutions meet or exceed the service reliability and availability of traditional deployments.

- *Chapter 10, “Applications, Solutions and Accountability.”* This chapter considers how virtualized applications fit into service solutions, and explains how application service downtime budgets change as applications move to the cloud. This chapter also proposes four measurement points for service availability, and discusses how accountability for impairments in each of those measurement points is attributed.
- *Chapter 11, “Recommendations for Architecting a Reliable System.”* This chapter covers architectures and techniques to maximize service availability and service reliability via virtualization and cloud deployment. A simple case study is given to illustrate key architectural points.
- *Chapter 12, “Design for Reliability of Virtualized Applications.”* This chapter reviews how design for reliability diligence for virtualized applications differs from reliability diligence for traditional applications.
- *Chapter 13, “Design for Reliability of Cloud Solutions.”* This chapter reviews how design for reliability diligence for cloud deployments differs from reliability diligence for traditional solutions.
- *Chapter 14, “Summary.”* This gives an executive summary of the analysis, insights, and recommendations on assuring that reliability and availability of cloud-based solutions meet or exceed the performance of traditional deployment.

ACKNOWLEDGMENTS

The authors were greatly assisted by many deeply knowledgeable and insightful engineers at Alcatel-Lucent, especially: Mark Clougherty, Herbert Ristock, Shawa Tam, Rich Sohn, Bernard Bretherton, John Haller, Dan Johnson, Srujal Shah, Alan McBride, Lyle Kipp, and Ted East. Joe Tieu, Bill Baker, and Thomas Voith carefully reviewed the early manuscript and provided keen review feedback. Abhaya Asthana, Kasper Reinink, Roger Maitland, and Mark Cameron provided valuable input. Gary McElvany raised the initial architectural questions that ultimately led to this work. This work would not have been possible without the strong management support of Tina Hinch, Werner Heissenhuber, Annie Lequesne, Vickie Owens-Rinn, and Dor Skuler.

Cloud computing is an exciting, evolving technology with many avenues to explore. Readers with comments or corrections on topics covered in this book, or topics for a future edition of this book, are invited to send email to the authors (Eric.Bauer@Alcatel-Lucent.com, Randee.Adams@Alcatel-Lucent.com, or pressbooks@ieee.org).

Eric Bauer
Randee Adams

I

BASICS

1

CLOUD COMPUTING

The U.S. National Institute of Standards and Technology (NIST) defines cloud computing as follows:

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

[NIST-800-145].

This definition frames cloud computing as a “utility” (or a “pay as you go”) consumption model for computing services, similar to the utility model deployed for electricity, water, and telecommunication service. Once a user is connected to the computing (or telecommunications, electricity, or water utility) cloud, they can consume as much service as they would like whenever they would like (within reasonable limits), and are billed for the resources consumed. Because the resources delivering the service can be shared (and hence amortized) across a broad pool of users, resource utilization and operational efficiency can be higher than they would be for dedicated resources for each individual user, and thus the price of the service to the consumer may well be lower from a cloud/utility provider compared with the alternative of deploying and operating private resources to provide the same service. Overall, these characteristics facilitate outsourcing production and delivery of these crucial “utility” services. For example, how

many individuals or enterprises prefer to generate all of their own electricity rather than purchasing it from a commercial electric power supplier?

This chapter reviews the essential characteristics of cloud computing, as well as several common characteristics of cloud computing, considers how cloud data centers differ from traditional data centers, and discusses the cloud service and cloud deployment models. The terminologies for the various roles in cloud computing that will be used throughout the book are defined. The chapter concludes by reviewing the benefits of cloud computing.

1.1 ESSENTIAL CLOUD CHARACTERISTICS

Per [NIST-800-145], there are five essential functional characteristics of cloud computing:

- 1.** on-demand self service;
- 2.** broad network access;
- 3.** resource pooling;
- 4.** rapid elasticity; and
- 5.** measured service.

Each of these is considered individually.

1.1.1 On-Demand Self-Service

Per [NIST-800-145], the essential cloud characteristic of “on-demand self-service” means “a consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service’s provider.” Modern telecommunications networks offer on-demand self service: one has direct dialing access to any other telephone