RELIABILITY AND AVAILABILITY OF CLOUD COMPUTING

ERIC BAUER

RANDEE ADAMS

WILEY

IEEE
IEEE PRESS

RELIABILITY AND AVAILABILITY OF CLOUD COMPUTING

IEEE Press
445 Hoes Lane
Piscataway, NJ 08854

IEEE Press Editorial Board 2012
John Anderson, *Editor in Chief*

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

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
IEEE PRESS
WILEY
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–Reliabillity. 3. Computer software–Quality control. 4. Computer security. I. Adams, Randee. II. Title.
QA76.585.B394 2012
004.6782–dc23
2011052839

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

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

CONTENTS

Figures **xvii**

Tables **xxi**

Equations **xxiii**

Introduction **xxv**

I BASICS **1**

1 CLOUD COMPUTING **3**

1.1 Essential Cloud Characteristics 4

1.1.1 On-Demand Self-Service 4

1.1.2 Broad Network Access 4

1.1.3 Resource Pooling 5

1.1.4 Rapid Elasticity 5

1.1.5 Measured Service 6

1.2 Common Cloud Characteristics 6

1.3 But What, Exactly, Is Cloud Computing? 7

1.3.1 What Is a Data Center? 8

1.3.2 How Does Cloud Computing Differ from Traditional Data Centers? 9

1.4 Service Models 9

1.5 Cloud Deployment Models 11

1.6 Roles in Cloud Computing 12

1.7 Benefits of Cloud Computing 14

1.8 Risks of Cloud Computing 15

2 VIRTUALIZATION 16
2.1 Background 16
2.2 What Is Virtualization? 17
2.2.1 Types of Hypervisors 18
2.2.2 Virtualization and Emulation 19
2.3 Server Virtualization 19
2.3.1 Full Virtualization 20
2.3.2 Paravirtualization 21
2.3.3 OS Virtualization 22
2.3.4 Discussion 22
2.4 VM Lifecycle 23
2.4.1 VM Snapshot 26
2.4.2 Cloning VMs 26
2.4.3 High Availability Mechanisms 28
2.5 Reliability and Availability Risks of Virtualization 28
3 SERVICE RELIABILITY AND SERVICE AVAILABILITY 29
3.1 Errors and Failures 30
3.2 Eight-Ingredient Framework 31
3.3 Service Availability 34
3.3.1 Service Availability Metric 35
3.3.2 MTBF and MTTR 36
3.3.3 Service and Network Element Impact Outages 37
3.3.4 Partial Outages 38
3.3.5 Availability Ratings 40
3.3.6 Outage Attributability 41
3.3.7 Planned or Scheduled Downtime 42
3.4 Service Reliability 43
3.4.1 Service Reliability Metrics 44
3.4.2 Defective Transactions 45
3.5 Service Latency 46
3.6 Redundancy and High Availability 50
3.6.1 Redundancy 51
3.6.2 High Availability 53
3.7 High Availability and Disaster Recovery 56
3.8 Streaming Services 58
3.8.1 Control and Data Planes 58
3.8.2 Service Quality Metrics 59
3.8.3 Isochronal Data 60
3.8.4 Latency Expectations 60
3.8.5 Streaming Quality Impairments 61
3.9 Reliability and Availability Risks of Cloud Computing 62

II ANALYSIS 63

4 ANALYZING CLOUD RELIABILITY AND AVAILABILITY 65

4.1 Expectations for Service Reliability and Availability 65

4.2 Risks of Essential Cloud Characteristics 66

4.2.1 On-Demand Self-Service 66

4.2.2 Broad Network Access 67

4.2.3 Resource Pooling 67

4.2.4 Rapid Elasticity 67

4.2.5 Measured Service 69

4.3 Impacts of Common Cloud Characteristics 70

4.3.1 Virtualization 70

4.3.2 Geographic Distribution 70

4.3.3 Resilient Computing 71

4.3.4 Advanced Security 71

4.3.5 Massive Scale 71

4.3.6 Homogeneity 71

4.4 Risks of Service Models 72

4.4.1 Traditional Accountability 72

4.4.2 Cloud-Based Application Accountability 73

4.5 IT Service Management and Availability Risks 74

4.5.1 ITIL Overview 74

4.5.2 Service Strategy 75

4.5.3 Service Design 76

4.5.4 Service Transition 77

4.5.5 Service Operation 77

4.5.6 Continual Service Improvement 78

4.5.7 IT Service Management Summary 79

4.5.8 Risks of Service Orchestration 79

4.5.9 IT Service Management Risks 80

4.6 Outage Risks by Process Area 80

4.6.1 Validating Outage Attributability 82

4.7 Failure Detection Considerations 83

4.7.1 Hardware Failures 83

4.7.2 Programming Errors 85

4.7.3 Data Inconsistency and Errors 85

4.7.4 Redundancy Errors 86

4.7.5 System Power Failures 86

4.7.6 Network Errors 86

4.7.7 Application Protocol Errors 86

4.8 Risks of Deployment Models 87

4.9 Expectations of IaaS Data Centers 87

5 RELIABILITY ANALYSIS OF VIRTUALIZATION 90

5.1 Reliability Analysis Techniques 90

5.1.1 Reliability Block Diagrams 90
5.1.2 Single Point of Failure Analysis 92
5.1.3 Failure Mode Effects Analysis 92

5.2 Reliability Analysis of Virtualization Techniques 95

5.2.1 Analysis of Full Virtualization 95
5.2.2 Analysis of OS Virtualization 95
5.2.3 Analysis of Paravirtualization 96
5.2.4 Analysis of VM Coresidency 97
5.2.5 Discussion 99

5.3 Software Failure Rate Analysis 100

5.3.1 Virtualization and Software Failure Rate 100
5.3.2 Hypervisor Failure Rate 101
5.3.3 Miscellaneous Software Risks of Virtualization and Cloud 101

5.4 Recovery Models 101

5.4.1 Traditional Recovery Options 101
5.4.2 Virtualized Recovery Options 102
5.4.3 Discussion 107

5.5 Application Architecture Strategies 108

5.5.1 On-Demand Single-User Model 108
5.5.2 Single-User Daemon Model 109
5.5.3 Multiuser Server Model 109
5.5.4 Consolidated Server Model 109

5.6 Availability Modeling of Virtualized Recovery Options 110

5.6.1 Availability of Virtualized Simplex Architecture 111
5.6.2 Availability of Virtualized Redundant Architecture 111
5.6.3 Critical Failure Rate 112
5.6.4 Failure Coverage 113
5.6.5 Failure Detection Latency 113
5.6.6 Switchover Latency 113
5.6.7 Switchover Success Probability 114
5.6.8 Modeling and “Fast Failure” 114
5.6.9 Comparison of Native and Virtualized Deployments 115

6 HARDWARE RELIABILITY, VIRTUALIZATION, AND SERVICE AVAILABILITY 116

6.1 Hardware Downtime Expectations 116

6.2 Hardware Failures 117

6.3 Hardware Failure Rate 119

6.4 Hardware Failure Detection 121
6.5 Hardware Failure Containment 122
6.6 Hardware Failure Mitigation 122
6.7 Mitigating Hardware Failures via Virtualization 124
6.7.1 Virtual CPU 124
6.7.2 Virtual Memory 125
6.7.3 Virtual Storage 126
6.8 Virtualized Networks 127
6.8.1 Virtual Network Interface Cards 127
6.8.2 Virtual Local Area Networks 128
6.8.3 Virtual IP Addresses 129
6.8.4 Virtual Private Networks 129
6.9 MTTR of Virtualized Hardware 129
6.10 Discussion 131

7 CAPACITY AND ELASTICITY 132

7.1 System Load Basics 132
7.1.1 Extraordinary Event Considerations 134
7.1.2 Slashdot Effect 134
7.2 Overload, Service Reliability, and Service Availability 135
7.3 Traditional Capacity Planning 136
7.4 Cloud and Capacity 137
7.4.1 Nominal Cloud Capacity Model 138
7.4.2 Elasticity Expectations 141
7.5 Managing Online Capacity 144
7.5.1 Capacity Planning Assumptions of Cloud Computing 145
7.6 Capacity-Related Service Risks 147
7.6.1 Elasticity and Elasticity Failure 147
7.6.2 Partial Capacity Failure 149
7.6.3 Service Latency Risk 150
7.6.4 Capacity Impairments and Service Reliability 152
7.7 Capacity Management Risks 153
7.7.1 Brittle Application Architecture 154
7.7.2 Faulty or Inadequate Monitoring Data 155
7.7.3 Faulty Capacity Decisions 155
7.7.4 Unreliable Capacity Growth 155
7.7.5 Unreliable Capacity Degrowth 156
7.7.6 Inadequate Slew Rate 156
7.7.7 Tardy Capacity Management Decisions 156
7.7.8 Resource Stock Out Not Covered 157

7.7.9 Cloud Burst Fails 157
7.7.10 Policy Constraints 157
7.8 Security and Service Availability 157
7.8.1 Security Risk to Service Availability 157
7.8.2 Denial of Service Attacks 159
7.8.3 Defending against DoS Attacks 160
7.8.4 Quantifying Service Availability Impact of Security Attacks 161
7.8.5 Recommendations 162
7.9 Architecting for Elastic Growth and Degrowth 162

8 SERVICE ORCHESTRATION ANALYSIS 164
8.1 Service Orchestration Definition 164
8.2 Policy-Based Management 166
8.2.1 The Role of SLRs 167
8.2.2 Service Reliability and Availability Measurements 168
8.3 Cloud Management 168
8.3.1 Role of Rapid Elasticity in Cloud Management 169
8.3.2 Role of Cloud Bursting in Cloud Management 169
8.4 Service Orchestration's Role in Risk Mitigation 169
8.4.1 Latency 170
8.4.2 Reliability 170
8.4.3 Regulatory 171
8.4.4 Security 171
8.5 Summary 172

9 GEOGRAPHIC DISTRIBUTION, GEOREDUNDANCY, AND DISASTER RECOVERY 174
9.1 Geographic Distribution versus Georedundancy 175
9.2 Traditional Disaster Recovery 175
9.3 Virtualization and Disaster Recovery 177
9.4 Cloud Computing and Disaster Recovery 178
9.5 Georedundancy Recovery Models 180
9.6 Cloud and Traditional Collateral Benefits of Georedundancy 180
9.6.1 Reduced Planned Downtime 180
9.6.2 Mitigate Catastrophic Network Element Failures 181
9.6.3 Mitigate Extended Uncovered and Duplex Failure Outages 181
9.7 Discussion 182

III RECOMMENDATIONS 183

10 APPLICATIONS, SOLUTIONS, AND ACCOUNTABILITY 185

10.1 Application Configuration Scenarios 185
10.2 Application Deployment Scenario 187
10.3 System Downtime Budgets 188
10.3.1 Traditional System Downtime Budget 189
10.3.2 Virtualized Application Downtime Budget 189
10.3.3 IaaS Hardware Downtime Expectations 191
10.3.4 Cloud-Based Application Downtime Budget 193
10.3.5 Summary 195
10.4 End-to-End Solutions Considerations 197
10.4.1 What is an End-to-End Solution? 197
10.4.2 Consumer-Specific Architectures 198
10.4.3 Data Center Redundancy 199
10.5 Attributability for Service Impairments 201
10.6 Solution Service Measurement 204
10.6.1 Service Availability Measurement Points 204
10.7 Managing Reliability and Service of Cloud Computing 207

11 RECOMMENDATIONS FOR ARCHITECTING A RELIABLE SYSTEM 209

11.1 Architecting for Virtualization and Cloud 209
11.1.1 Mapping Software into VMs 210
11.1.2 Service Load Distribution 210
11.1.3 Data Management 211
11.1.4 Software Redundancy and High Availability Mechanisms 212
11.1.5 Rapid Elasticity 214
11.1.6 Overload Control 214
11.1.7 Coresidency 215
11.1.8 Multitenancy 215
11.1.9 Isochronal Applications 216
11.2 Disaster Recovery 216
11.3 IT Service Management Considerations 217
11.3.1 Software Upgrade and Patch 217
11.3.2 Service Transition Activity Effect Analysis 218
11.3.3 Mitigating Service Transition Activity Effects via VM Migration 219
11.3.4 Testing Service Transition Activities 221

11.3.5 Minimizing Procedural Errors 221
11.3.6 Service Orchestration Considerations 223
11.4 Many Distributed Clouds versus Fewer Huge Clouds 224
11.5 Minimizing Hardware-Attributed Downtime 225
11.5.1 Hardware Downtime in Traditional High Availability Configurations 226
11.6 Architectural Optimizations 231
11.6.1 Reliability and Availability Criteria 232
11.6.2 Optimizing Accessibility 233
11.6.3 Optimizing High Availability, Retainability, Reliability, and Quality 235
11.6.4 Optimizing Disaster Recovery 235
11.6.5 Operational Considerations 236
11.6.6 Case Study 236
11.6.7 Theoretically Optimal Application Architecture 241

12 DESIGN FOR RELIABILITY OF VIRTUALIZED APPLICATIONS 244
12.1 Design for Reliability 244
12.2 Tailoring DfR for Virtualized Applications 246
12.2.1 Hardware Independence Usage Scenario 246
12.2.2 Server Consolidation Usage Scenario 247
12.2.3 Multitenant Usage Scenario 248
12.2.4 Virtual Appliance Usage Scenario 248
12.2.5 Cloud Deployment Usage Scenario 248
12.3 Reliability Requirements 248
12.3.1 General Availability Requirements 249
12.3.2 Service Reliability and Latency Requirements 250
12.3.3 Overload Requirements 251
12.3.4 Online Capacity Growth and Degrowth 253
12.3.5 (Virtualization) Live Migration Requirements 253
12.3.6 System Transition Activity Requirements 254
12.3.7 Georedundancy and Service Continuity Requirements 255
12.4 Qualitative Reliability Analysis 256
12.4.1 SPOF Analysis for Virtualized Applications 256
12.4.2 Failure Mode Effects Analysis for Virtualized Applications 258
12.4.3 Capacity Growth and Degrowth Analysis 258
12.5 Quantitative Reliability Budgeting and Modeling 259
12.5.1 Availability (Downtime) Modeling 259
12.5.2 Converging Downtime Budgets and Targets 260
12.5.3 Managing Maintenance Budget Allocation 260

12.6 Robustness Testing 260
12.6.1 Baseline Robustness Testing 261
12.6.2 Advanced Topic: Can Virtualization Enable Better Robustness Testing? 265
12.7 Stability Testing 267
12.8 Field Performance Analysis 268
12.9 Reliability Roadmap 269
12.10 Hardware Reliability 270

13 DESIGN FOR RELIABILITY OF CLOUD SOLUTIONS 271
13.1 Solution Design for Reliability 271
13.2 Solution Scope and Expectations 273
13.3 Reliability Requirements 275
13.3.1 Solution Availability Requirements 275
13.3.2 Solution Reliability Requirements 276
13.3.3 Disaster Recovery Requirements 277
13.3.4 Elasticity Requirements 277
13.3.5 Specifying Configuration Parameters 278
13.4 Solution Modeling and Analysis 279
13.4.1 Reliability Block Diagram of Cloud Data Center Deployment 279
13.4.2 Solution Failure Mode Effects Analysis 280
13.4.3 Solution Service Transition Activity Effects Analysis 280
13.4.4 Cloud Data Center Service Availability (MP 2) Analysis 280
13.4.5 Aggregate Service Availability (MP 3) Modeling 281
13.4.6 Recovery Point Objective Analysis 285
13.5 Element Reliability Diligence 285
13.6 Solution Testing and Validation 285
13.6.1 Robustness Testing 286
13.6.2 Service Reliability Testing 286
13.6.3 Georedundancy Testing 286
13.6.4 Elasticity and Orchestration Testing 287
13.6.5 Stability Testing 287
13.6.6 In Service Testing 288
13.7 Track and Analyze Field Performance 288
13.7.1 Cloud Service Measurements 289
13.7.2 Solution Reliability Roadmapping 291
13.8 Other Solution Reliability Diligence Topics 292
13.8.1 Service-Level Agreements 292
13.8.2 Cloud Service Provider Selection 293
13.8.3 Written Reliability Plan 293

14 SUMMARY 296
14.1 Service Reliability and Service Availability 297
14.2 Failure Accountability and Cloud Computing 299
14.3 Factoring Service Downtime 301
14.4 Service Availability Measurement Points 303
14.5 Cloud Capacity and Elasticity Considerations 306
14.6 Maximizing Service Availability 306
14.6.1 Reducing Product Attributable Downtime 307
14.6.2 Reducing Data Center Attributable Downtime 307
14.6.3 Reducing IT Service Management Downtime 307
14.6.4 Reducing Disaster Recovery Downtime 308
14.6.5 Optimal Cloud Service Availability 308
14.7 Reliability Diligence 309
14.8 Concluding Remarks 310

Abbreviations 311
References 314
About the Authors 318
Index 319

FIGURES

Figure 1.1 Service Models 10
Figure 1.2 OpenCrowd's Cloud Taxonomy 11
Figure 1.3 Roles in Cloud Computing 13
Figure 2.1 Virtualizing Resources 18
Figure 2.2 Type 1 and Type 2 Hypervisors 19
Figure 2.3 Full Virtualization 21
Figure 2.4 Paravirtualization 21
Figure 2.5 Operating System Virtualization 22
Figure 2.6 Virtualized Machine Lifecycle State Transitions 25
Figure 3.1 Fault Activation and Failures 30
Figure 3.2 Minimum Chargeable Service Disruption 31
Figure 3.3 Eight-Ingredient ("8i") Framework 32
Figure 3.4 Eight-Ingredient plus Data plus Disaster (8i + 2d) Model 34
Figure 3.5 MTBF and MTTR 36
Figure 3.6 Service and Network Element Impact Outages of Redundant Systems 38
Figure 3.7 Sample DSL Solution 39
Figure 3.8 Transaction Latency Distribution for Sample Service 47
Figure 3.9 Requirements Overlaid on Service Latency Distribution for Sample Solution 48
Figure 3.10 Maximum Acceptable Service Latency 49
Figure 3.11 Downtime of Simplex Systems 50
Figure 3.12 Downtime of Redundant Systems 51
Figure 3.13 Simplified View of High Availability 54
Figure 3.14 High Availability Example 55
Figure 3.15 Disaster Recovery Objectives 57
Figure 3.16 ITU-T G.114 Bearer Delay Guideline 61
Figure 4.1 TL 9000 Outage Attributability Overlaid on Augmented 8i + 2d Framework 72
Figure 4.2 Outage Responsibilities Overlaid on Cloud 8i + 2d Framework 73

Figure 4.3 ITIL Service Management Visualization 75
Figure 4.4 IT Service Management Activities to Minimize Service Availability Risk 79
Figure 4.5 8i + 2d Attributability by Process or Best Practice Areas 81
Figure 4.6 Traditional Error Vectors 84
Figure 4.7 IaaS Provider Responsibilities for Traditional Error Vectors 84
Figure 4.8 Software Supplier (and SaaS) Responsibilities for Traditional Error Vectors 85
Figure 5.1 Sample Reliability Block Diagram 91
Figure 5.2 Traversal of Sample Reliability Block Diagram 91
Figure 5.3 Nominal System Reliability Block Diagram 92
Figure 5.4 Reliability Block Diagram of Full virtualization 95
Figure 5.5 Reliability Block Diagram of OS Virtualization 96
Figure 5.6 Reliability Block Diagram of Paravirtualization 97
Figure 5.7 Reliability Block Diagram of Coresident Application Deployment 98
Figure 5.8 Canonical Virtualization RBD 99
Figure 5.9 Latency of Traditional Recovery Options 103
Figure 5.10 Traditional Active-Standby Redundancy via Active VM Virtualization 104
Figure 5.11 Reboot of a Virtual Machine 104
Figure 5.12 Reset of a Virtual Machine 105
Figure 5.13 Redundancy via Paused VM Virtualization 106
Figure 5.14 Redundancy via Suspended VM Virtualization 106
Figure 5.15 Nominal Recovery Latency of Virtualized and Traditional Options 107
Figure 5.16 Server Consolidation Using Virtualization 110
Figure 5.17 Simplified Simplex State Diagram 111
Figure 5.18 Downtime Drivers for Redundancy Pairs 112
Figure 6.1 Hardware Failure Rate Questions 120
Figure 6.2 Application Reliability Block Diagram with Virtual Devices 124
Figure 6.3 Virtual CPU 125
Figure 6.4 Virtual NIC 128
Figure 7.1 Sample Application Resource Utilization by Time of Day 133
Figure 7.2 Example of Extraordinary Event Traffic Spike 134
Figure 7.3 The Slashdot Effect: Traffic Load Over Time (in Hours) 134
Figure 7.4 Offered Load, Service Reliability, and Service Availability of a Traditional System 135

Figure 7.5 Visualizing VM Growth Scenarios 138
Figure 7.6 Nominal Capacity Model 139
Figure 7.7 Implementation Architecture of Compute Capacity Model 139
Figure 7.8 Orderly Reconfiguration of the Capacity Model 140
Figure 7.9 Slew Rate of Square Wave Amplification 141
Figure 7.10 Slew Rate of Rapid Elasticity 142
Figure 7.11 Elasticity Timeline by ODCA SLA Level 143
Figure 7.12 Capacity Management Process 144
Figure 7.13 Successful Cloud Elasticity 148
Figure 7.14 Elasticity Failure Model 148
Figure 7.15 Virtualized Application Instance Failure Model 150
Figure 7.16 Canonical Capacity Management Failure Scenarios 154
Figure 7.17 ITU X.805 Security Dimensions, Planes, and Layers 158
Figure 7.18 Leveraging Security and Network Infrastructure to Mitigate Overload Risk 161
Figure 8.1 Service Orchestration 167
Figure 8.2 Example of Cloud Bursting 170
Figure 10.1 Canonical Single Data Center Application Deployment Architecture 188
Figure 10.2 RBD of Sample Application on Blade-Based Server Hardware 192
Figure 10.3 RBD of Sample Application on IaaS Platform 192
Figure 10.4 Sample End-to-End Solution 197
Figure 10.5 Sample Distributed Cloud Architecture 199
Figure 10.6 Sample Recovery Scenario in Distributed Cloud Architecture 200
Figure 10.7 Simplified Responsibilities for a Canonical Cloud Application 203
Figure 10.8 Recommended Cloud-Related Service Availability Measurement Points 205
Figure 10.9 Canonical Example of MP 1 and MP 2 206
Figure 10.10 End-to-End Service Availability Key Quality Indicators 207
Figure 11.1 Virtual Machine Live Migration 219
Figure 11.2 Active–Standby Markov Model 227
Figure 11.3 Pie Chart of Canonical Hardware Downtime Prediction 228
Figure 11.4 RBD for the Hypothetical Web Server Application 237
Figure 11.5 Horizontal Growth of Hypothetical Application 238

Figure 11.6 Outgrowth of Hypothetical Application 239
Figure 11.7 Aggressive Protocol Retry Strategy 239
Figure 11.8 Data Replication of Hypothetical Application 240
Figure 11.9 Disaster Recovery of Hypothetical Application 241
Figure 11.10 Optimal Availability Architecture of Hypothetical Application 242
Figure 12.1 Traditional Design for Reliability Process 245
Figure 12.2 Mapping Virtual Machines across Hypervisors 257
Figure 12.3 A Virtualized Server Failure Scenario 257
Figure 12.4 Robustness Testing Vectors for Virtualized Applications 265
Figure 12.5 System Design for Reliability as a Deming Cycle 268
Figure 13.1 Solution Design for Reliability 272
Figure 13.2 Sample Solution Scope and KQI Expectations 274
Figure 13.3 Sample Cloud Data Center RBD 279
Figure 13.4 Estimating MP 2 281
Figure 13.5 Modeling Cloud-Based Solution with Client-Initiated Recovery Model 283
Figure 13.6 Client-Initiated Recovery Model 283
Figure 14.1 Failure Impact Duration and High Availability Goals 298
Figure 14.2 Eight-Ingredient Plus Data Plus Disaster (8i+2d) Model 299
Figure 14.3 Traditional Outage Attributability 300
Figure 14.4 Sample Outage Accountability Model for Cloud Computing 301
Figure 14.5 Outage Responsibilities of Cloud by Process 302
Figure 14.6 Measurement Pointss (MPs) 1, 2, 3, and 4 305
Figure 14.7 Design for Reliability of Cloud-Based Solutions 310

TABLES

Table 2.1 Comparison of Server Virtualization Technologies 24
Table 2.2 Virtual Machine Lifecycle Transitions 27
Table 3.1 Service Availability and Downtime Ratings 40
Table 3.2 Mean Opinion Scores 59
Table 4.1 ODCA's Data Center Classification 88
Table 4.2 ODCA's Data Center Service Availability Expectations by Classification 89
Table 5.1 Example Failure Mode Effects Analysis 94
Table 5.2 Failure Mode Effect Analysis Figure for Coresident Applications 99
Table 5.3 Comparison of Nominal Software Availability Parameters 114
Table 6.1 Example of Hardware Availability as a Function of MTTR/MTTRS 130
Table 7.1 ODCA IaaS Elasticity Objectives 143
Table 9.1 ODCA IaaS Recoverability Objectives 179
Table 10.1 Sample Traditional Five 9's Downtime Budget 190
Table 10.2 Sample Basic Virtualized Five 9's Downtime Budget 191
Table 10.3 Canonical Application-Attributable Cloud-Based Five 9's Downtime Budget 195
Table 10.4 Evolution of Sample Downtime Budgets 196
Table 11.1 Example Service Transition Activity Failure Mode Effect Analysis 218
Table 11.2 Canonical Hardware Downtime Prediction 227
Table 11.3 Summary of Hardware Downtime Mitigation Techniques for Cloud Computing 231
Table 12.1 Sample Service Latency and Reliability Requirements at MP 2 250
Table 13.1 Sample Solution Latency and Reliability Requirements 276
Table 13.2 Modeling Input Parameters 284
Table 14.1 Evolution of Sample Downtime Budgets 304

EQUATIONS

Equation 3.1	Basic Availability Formula	35
Equation 3.2	Practical System Availability Formula	35
Equation 3.3	Standard Availability Formula	35
Equation 3.4	Estimation of System Availability from MTBF and MTTR	36
Equation 3.5	Recommended Service Availability Formula	38
Equation 3.6	Sample Partial Outage Calculation	39
Equation 3.7	Service Reliability Formula	44
Equation 3.8	DPM Formula	44
Equation 3.9	Converting DPM to Service Reliability	44
Equation 3.10	Converting Service Reliability to DPM	44
Equation 3.11	Sample DPM Calculation	45
Equation 6.1	Availability as a Function of MTBF/MTTR	130
Equation 11.1	Maximum Theoretical Availability across Redundant Elements	241
Equation 11.2	Maximum Theoretical Service Availability	242

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.
- C*hapter 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