BEYOND REDUNDANCY

How Geographic Redundancy Can Improve Service Availability and Reliability of Computer-based Systems

ERIC BAUER · RANDEE ADAMS · DANIEL EUSTACE

WILEY

IEEE
IEEE PRESS

Table of Contents

Cover

Series page

Title page

Copyright page

Dedication

FIGURES

TABLES

EQUATIONS

PREFACE AND ACKNOWLEDGMENTS

- *AUDIENCE*
- *ORGANIZATION*
- *ACKNOWLEDGMENTS*

PART 1: BASICS

1 SERVICE, RISK, AND BUSINESS CONTINUITY

1.1 SERVICE CRITICALITY AND AVAILABILITY EXPECTATIONS
1.2 THE EIGHT-INGREDIENT MODEL
1.3 CATASTROPHIC FAILURES AND GEOGRAPHIC REDUNDANCY
1.4 GEOGRAPHICALLY SEPARATED RECOVERY SITE
1.5 MANAGING RISK
1.6 BUSINESS CONTINUITY PLANNING
1.7 DISASTER RECOVERY PLANNING
1.8 HUMAN FACTORS
1.9 RECOVERY OBJECTIVES
1.10 DISASTER RECOVERY STRATEGIES

2 SERVICE AVAILABILITY AND SERVICE RELIABILITY

2.1 AVAILABILITY AND RELIABILITY
2.2 MEASURING SERVICE AVAILABILITY
2.3 MEASURING SERVICE RELIABILITY

PART 2: MODELING AND ANALYSIS OF REDUNDANCY

3 UNDERSTANDING REDUNDANCY

3.1 TYPES OF REDUNDANCY
3.2 MODELING AVAILABILITY OF INTERNAL REDUNDANCY
3.3 EVALUATING HIGH-AVAILABILITY MECHANISMS

4 OVERVIEW OF EXTERNAL REDUNDANCY

4.1 GENERIC EXTERNAL REDUNDANCY MODEL
4.2 TECHNICAL DISTINCTIONS BETWEEN GEOREDUNDANCY AND CO-LOCATED REDUNDANCY
4.3 MANUAL GRACEFUL SWITCHOVER AND SWITCHBACK

5 EXTERNAL REDUNDANCY STRATEGY OPTIONS

5.1 REDUNDANCY STRATEGIES
5.2 DATA RECOVERY STRATEGIES
5.3 EXTERNAL RECOVERY STRATEGIES
5.4 MANUALLY CONTROLLED RECOVERY
5.5 SYSTEM-DRIVEN RECOVERY
5.6 CLIENT-INITIATED RECOVERY

6 MODELING SERVICE AVAILABILITY WITH EXTERNAL SYSTEM REDUNDANCY

6.1 THE SIMPLISTIC ANSWER
6.2 FRAMING SERVICE AVAILABILITY OF STANDALONE SYSTEMS
6.3 GENERIC MARKOV AVAILABILITY MODEL OF GEOREDUNDANT RECOVERY
6.4 SOLVING THE GENERIC GEOREDUNDANCY MODEL

6.5 PRACTICAL MODELING OF GEOREDUNDANCY
6.6 ESTIMATING AVAILABILITY BENEFIT FOR PLANNED ACTIVITIES
6.7 ESTIMATING AVAILABILITY BENEFIT FOR DISASTERS

7 UNDERSTANDING RECOVERY TIMING PARAMETERS

7.1 DETECTING IMPLICIT FAILURES
7.2 UNDERSTANDING AND OPTIMIZING RTO

8 CASE STUDY OF CLIENT-INITIATED RECOVERY

8.1 OVERVIEW OF DNS
8.2 MAPPING DNS ONTO PRACTICAL CLIENT-INITIATED RECOVERY MODEL
8.3 ESTIMATING INPUT PARAMETERS
8.4 PREDICTED RESULTS
8.5 DISCUSSION OF PREDICTED RESULTS

9 SOLUTION AND CLUSTER RECOVERY

9.1 UNDERSTANDING SOLUTIONS
9.2 ESTIMATING SOLUTION AVAILABILITY
9.3 CLUSTER VERSUS ELEMENT RECOVERY
9.4 ELEMENT FAILURE AND CLUSTER RECOVERY CASE STUDY
9.5 COMPARING ELEMENT AND CLUSTER RECOVERY

9.6 MODELING CLUSTER RECOVERY

PART 3: RECOMMENDATIONS

10 GEOREDUNDANCY STRATEGY

10.1 WHY SUPPORT MULTIPLE SITES?
10.2 RECOVERY REALMS
10.3 RECOVERY STRATEGIES
10.4 LIMP-ALONG ARCHITECTURES
10.5 SITE REDUNDANCY OPTIONS
10.6 VIRTUALIZATION, CLOUD COMPUTING, AND STANDBY SITES
10.7 RECOMMENDED DESIGN METHODOLOGY

11 MAXIMIZING SERVICE AVAILABILITY VIA GEOREDUNDANCY

11.1 THEORETICALLY OPTIMAL EXTERNAL REDUNDANCY
11.2 PRACTICALLY OPTIMAL RECOVERY STRATEGIES
11.3 OTHER CONSIDERATIONS

12 GEOREDUNDANCY REQUIREMENTS

12.1 INTERNAL REDUNDANCY REQUIREMENTS
12.2 EXTERNAL REDUNDANCY REQUIREMENTS

12.3 MANUALLY CONTROLLED REDUNDANCY REQUIREMENTS
12.4 AUTOMATIC EXTERNAL RECOVERY REQUIREMENTS
12.5 OPERATIONAL REQUIREMENTS

13 GEOREDUNDANCY TESTING
13.1 GEOREDUNDANCY TESTING STRATEGY
13.2 TEST CASES FOR EXTERNAL REDUNDANCY
13.3 VERIFYING GEOREDUNDANCY REQUIREMENTS
13.4 SUMMARY

14 SOLUTION GEOREDUNDANCY CASE STUDY
14.1 THE HYPOTHETICAL SOLUTION
14.2 STANDALONE SOLUTION ANALYSIS
14.3 GEOREDUNDANT SOLUTION ANALYSIS
14.4 AVAILABILITY OF THE GEOREDUNDANT SOLUTION
14.5 REQUIREMENTS OF HYPOTHETICAL SOLUTION
14.6 TESTING OF HYPOTHETICAL SOLUTION

SUMMARY

APPENDIX: MARKOV MODELING OF SERVICE AVAILABILITY

ACRONYMS

REFERENCES

ABOUT THE AUTHORS

Index

IEEE Press
445 Hoes Lane
Piscataway, NJ 08854

Technical Reviewers

Xuemei Zhang

Network Design and Performance Analysis Division, AT&T Labs

Kim W. Tracy

Northeastern Illinois University

BEYOND REDUNDANCY

How Geographic Redundancy Can Improve Service Availability and Reliability of Computer-Based Systems

Eric Bauer
Randee Adams
Daniel Eustace

IEEE PRESS

A JOHN WILEY & SONS, INC., PUBLICATION

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

Published simultaneously in Canada.

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.

Beyond redundancy : how geographic redundancy can improve service availability and reliability of computer-based systems / Eric Bauer, Randee Adams, Daniel Eustace.

p. cm.

ISBN 978-1-118-03829-1 (hardback)

1. Computer input-output equipment–Reliability. 2. Computer networks–Reliability. 3. Redundancy (Engineering) I. Adams, Randee. II. Eustace, Daniel. III. Title.

TK7887.5.B395 2011

004.6–dc22

2011008324

oBook ISBN: 978-1-118-10491-0

ePDF ISBN: 978-1-118-10492-7

ePub ISBN: 978-1-118-10493-4

To our families for their encouragement and support:
Eric's wife Sandy and children Lauren and Mark
Randee's husband Scott and son Ryan
Dan's wife Helen and daughters Christie and Chelsea

FIGURES

Figure 1.1. The Eight-Ingredient Model
Figure 1.2. Canonical Disaster Recovery Scenario
Figure 1.3. Recovery Time Objective and Recovery Point Objective
Figure 2.1. Canonical Service Impact Timeline
Figure 2.2. Reliability, Availability, and Quality of Service
Figure 2.3. Outage Downtime in Standalone NE Deployment
Figure 2.4. Outage Downtime with Redundant NE Deployment
Figure 3.1. Types of System Redundancy
Figure 3.2. State Transition Diagram of Simplex System
Figure 3.3. Availability Improvement Strategies for a Simplex System
Figure 3.4. Reliability Block Diagram of Redundant Pair
Figure 3.5. Service Availability of Active–Standby Redundant Pair
Figure 3.6. Sample Reliability Block Diagram
Figure 3.7. Sample Standalone Redundant System
Figure 3.8. Active–Active Markov Availability Model
Figure 3.9. Active–Active Markov Availability Model with Formulas
Figure 3.10. Active–Standby Markov Availability Model
Figure 3.11. Simplex Model with Mathematical Formulas
Figure 3.12. Outage Duration
Figure 3.13. Outage Duration for Disasters
Figure 4.1. Generic High Availability Model
Figure 4.2. Stable Service Delivery Path Across Generic Model

Figure 4.3. Degenerate Generic Model Without Element “C”
Figure 4.4. Failure Scenario in Generic Model
Figure 4.5. Recovery Scenario in Generic Model
Figure 4.6. Generic High Availability Model with Load Sharing
Figure 4.7. Georedundancy Using DNS SRV Records
Figure 5.1. Client-Initiated Recovery Scenario
Figure 5.2. Typical Client Processing Logic for Standalone Server
Figure 5.3. Generic Client-Initiated Recovery Logic with Redundant Servers
Figure 5.4. Session States Seen by a SIP Client “A” in Client-Initiated Recovery
Figure 5.5. Session States Seen by “A” in Client-Initiated Recovery Without Registration
Figure 5.6. Browser Query to Web Server
Figure 6.1. Generic High Availability Model
Figure 6.2. Sample Unavailability Contribution for Active–Standby Redundancy
Figure 6.3. Simplified Standalone High-Availability Downtime Model
Figure 6.4. General Georedundant Manual Recovery Markov Transition Diagram
Figure 6.5. System-Driven Georedundant Recovery Markov Transition Diagram
Figure 6.6. Client-Initiated Georedundant Recovery Markov Transition Diagram
Figure 6.7. Client-Initiated and System-Driven Georedundant Recovery
Figure 6.8. Overlaying Generic Georedundancy Model Onto Simplistic Model
Figure 6.9. Outage Durations for Sample System with internal Redundancy

Figure 6.10. Simplified Client-Initiated Recovery Markov Model
Figure 6.11. Modified Client-Initiated Recovery Model
Figure 6.12. Estimating Service Unavailability During Manual Disaster Recovery
Figure 7.1. Standard Protocol Timeout
Figure 7.2. Adaptive Protocol Timeout
Figure 7.3. Timeout with Multiple Parallel Requests
Figure 7.4. Client/Server Keepalive
Figure 7.5. Manually Controlled Recovery Timeline
Figure 7.6. System-Driven Recovery Timeline
Figure 7.7. Client-Initiated Recovery Timeline
Figure 8.1. Generic Model of DNS
Figure 8.2. Practical Client-Initiated Model for DNS
Figure 8.3. Modeling Normal DNS Operation
Figure 8.4. Modeling Server Failure
Figure 8.5. Modeling Timeout Failure
Figure 8.6. Modeling Abnormal Server Failure
Figure 8.7. Modeling Multiple Server Failures
Figure 8.8. Simplified Client-Initiated Recovery Model with Formulas
Figure 8.9. Critical Failure Rate Parameter
Figure 8.10. Failure Rate as a Function of MTTR
Figure 8.11. F_{EXPLICIT} Parameter
Figure 8.12. C_{CLIENT} Parameter
Figure 8.13. μ_{TIMEOUT} Parameter
Figure 8.14. $\mu_{\text{CLIENTSFD}}$ Parameter
Figure 8.15. μ_{CLIENT} Parameter
Figure 8.16. $A_{\text{CLUSTER}-1}$ Parameter
Figure 8.17. F_{CLIENT} Parameter
Figure 8.18. μ_{GRECOVER} and $\mu_{\text{MIGRATION}}$ Parameters

Figure 8.19. μ_{DUPLEX} Parameter
Figure 8.20. Sensitivity of Critical Failures per Year
Figure 8.21. Sensitivity of C_{CLIENT}
Figure 8.22. Sensitivity of $\mu_{GRECOVERY}$
Figure 8.23. Sensitivity of μ_{DUPLEX}
Figure 8.24. Sensitivity of μ_{CLIENT}
Figure 9.1. External interfaces to Hypothetical Solution
Figure 9.2. Network Diagram of Hypothetical Solution
Figure 9.3. External interfaces to Hypothetical Solution
Figure 9.4. Reliability Block Diagram for End User Service of Hypothetical Solution
Figure 9.5. Reliability Block Diagram for User Provisioning Service of Hypothetical Solution
Figure 9.6. Downtime Expectation for End User Service of Sample Solution
Figure 9.7. Generic Redundant Cluster Model
Figure 9.8. Element Recovery upon Element Failure
Figure 9.9. Cluster Recovery upon Element Failure
Figure 9.10. Reliability Block Diagram for Georedundant Redundant Hypothetical Solution
Figure 9.11. Client Recovery for a Frontend Server Failure
Figure 9.12. Element Recovery of Security Server in Sample Solution
Figure 9.13. Cluster Recovery of Database Server in Sample Solution
Figure 9.14. Modeling Super Element Recovery
Figure 9.15. Client-Initiated Super Element Cluster Recovery Model
Figure 9.16. Active–Active States That Can Be Mitigated by Client-Initiated Recovery
Figure 9.17. Active–Active State Transitions That Can Activate Client-Initiated Recovery
Figure 10.1. Georedundancy with a Standby Site

Figure 10.2. Spare Site Georedundancy After a Failure
Figure 10.3. Georedundancy with $N + K$ Load Sharing
Figure 10.4. $N + K$ Load Sharing Georedundancy After a Failure
Figure 10.5. $N + K$ Load Sharing with 1 + 1 Redundant Elements
Figure 10.6. $N + K$ Load Sharing with 1 + 1 Redundancy After a Failure
Figure 14.1. Hypothetical Solution Architecture
Figure 14.2. Reliability Block Diagram for End User Service of Hypothetical Solution
Figure 14.3. Reliability Block Diagram for User Provisioning Service of Sample Solution
Figure 14.4. Estimating Service Availability Experienced by End Users
Figure S1. Generic High Availability Model
Figure S2. Visualization of Sample Generic Modeling Results
Figure A1. Andrey Markov (1856–1922).
Figure A2. Simple Availability Transition Diagram

TABLES

Table 1.1. FAA’s Availability Expectations by Service Thread Criticality

Table 2.1. Service Availability and Downtime Ratings

Table 3.1. Sample Input Parameters for Active–Active Model

Table 3.2. Probability of Time Spent in Each Active–Active State

Table 3.3. Sample Additional Input Parameters for Active–Standby Model

Table 3.4. Probability of Time Spent in Each Active–Standby State

Table 3.5. Probability of Time Spent in Each Simplex State

Table 3.6. Nominal Downtime Predictions for Different Redundancy Schemes

Table 6.1. Mitigation of Service Downtime by Georedundancy Recovery Category

Table 6.2. Nominal Modeling Parameters and Values

Table 6.3. Nominal Modeling Results Summary

Table 7.1. Nominal RTO Times for Each External Redundancy Strategy

Table 8.1. DNS RCODE Return Code Values

Table 8.2. Deducing Server Status from DNS RCODE

Table 8.3. Modeling Input Parameters

Table 8.4. DNS RCODEs by C_{CLIENT}

Table 8.5. DNS Client-Initiated Recovery Modeling Parameters

Table 8.6. Predicted Service Availability for DNS

Table 9.1. Comparison of Client-Initiated Recovery Parameter Values

Table 9.2. Input Parameters for Standalone Active–Active Model
Table 9.3. Time Spent in Each Standalone Active–Active State
Table 9.4. Sample Cluster Client-Initiated Recovery Modeling Parameters
Table 9.5. Client-Initiated Recovery Prediction
Table 9.6. Solution Downtime Prediction
Table 14.1. Service Criticality by External Solution Interface
Table 14.2. Availability Expectations for Hypothetical Solution Elements
Table 14.3. Redundancy Schemes for the Solution Elements

EQUATIONS

Equation 2.1. Service Availability Equation
Equation 2.2. Defects per Million Equation
Equation 2.3. Converting Service Reliability to DPM
Equation 3.1. Availability of Simplex System Formula
Equation 3.2. Availability of Active–Standby System Formula
Equation 6.1. Simplistic Redundancy Availability Formula
Equation 6.2. Maximum Feasible Availability Benefit of System-Driven Recovery
Equation 7.1. General Equation for RTO
Equation 8.1. Availability as a Function of MTBF and MTTR
Equation 8.2. MTBF as a Function of Availability and MTTR
Equation 8.3. Failure Rate as a Function of Availability and MTTR
Equation 9.1. Estimating $\lambda_{SUPERELEMENT}$

PREFACE AND ACKNOWLEDGMENTS

The best practice for mitigating the risk of site destruction, denial, or unavailability causing disastrous loss of critical services is to deploy redundant systems in a geographically separated location; this practice is called geographic redundancy or georedundancy. Enterprises deploying a geographically redundant system may spend significantly more than when deploying a standalone configuration up front, and will have higher ongoing operating expenses to maintain the geographically separated redundant recovery site and system. While the business continuity benefits of georedundancy are easy to understand, the feasible and likely service availability benefits of georedundancy are not generally well understood. This book considers the high-level question of what service availability improvement is feasible and likely with georedundancy. The emphasis is on system availability of IP-based applications. WAN availability is briefly mentioned where applicable, but is not factored into any of the modeling. The service availability benefit is characterized both for product attributable failures, as well as for nonproduct attributable failures, such as site disasters. Human factors are also taken into consideration as they relate to procedural downtime. Furthermore, this book considers architectural and operational topics, such as: whether it is better to only do a georedundancy failover for a failed element or for the entire cluster of elements that contains the failed element; whether georedundancy can/should be used to reduce planned downtime for activities such as hardware growth and software upgrade; what availability-related georedundancy requirements should apply to each network element and to clusters of elements; and what network element- and cluster-level

testing is appropriate to assure expected service availability benefits of georedundancy.

This book considers the range of IP-based information and communication technology (ICT) systems that are typically deployed in enterprise data centers and telecom central offices. The term “enterprise” is used to refer to the service provider or enterprise operating the system, “supplier” is used to refer to the organization that develops and tests the system, and “user” is used for the human or system that uses the system. In some cases, “enterprise,” “supplier,” and “user” may all be part of the same larger organization (e.g., system that is developed, tested and operated by the IT department of a larger, and used by employees of the organization), but often two or all three of these parties are in different organizations.

The term network element refers to a system device, entity, or node including all relevant hardware and/or software components deployed at one location providing a particular primary function; an instance of a domain name system (DNS) server is a network element. A system is “*a collection of components organized to accomplish a specific function or set of functions*” (*IEEE Standard Glossary*, 1991); a pool of DNS servers is an example of system. A solution is an integrated suite of network elements that can provide multiple primary functions; a customer care center that may include functionality, such as call handling facilities, web servers, and billing servers, is an example of a solution. With improvements in technology and hardware capacity, the distinction between these terms often blurs, since a single server could perform all of the functionality required of the solution and might be considered a network element. The more general term “external redundancy” is used to encompass both traditional geographic redundancy in which redundant system instances are physically separated to minimize the risk of a single catastrophic event impacting

both instances, as well as the situation in which redundant system instances are physically co-located. While physically co-located systems do not mitigate the risk of catastrophic site failure, they can mitigate the risk of system failures. External redundancy is contrasted with internal redundancy in which the redundancy is confined to a single element instance. For example, a RAID array is a common example of internal redundancy because the software running on the element or the RAID hardware assures that disk failures are detected and mitigated without disrupting user service. If each element requires a dedicated RAID array and an enterprise chooses to deploy a pair of elements for redundancy, then those elements could either be co-located in a single facility or installed in separate, presumably geographically distant, facilities. Both co-located and geographically separated configurations are considered “externally redundant,” as the redundancy encompasses multiple element instances. Elements can be deployed with no redundancy, internal redundancy, external redundancy, or hybrid arrangements. This book discusses internal redundancy but focuses on external redundancy arrangements.

AUDIENCE

This book is written for network architects and designers, maintenance and operations engineers, and decision makers in IT organizations at enterprises who are considering or have deployed georedundant systems. This book is also written for system architects, system engineers, developers, testers, and others (including technical sales and support staff) involved in the development of systems supporting external redundancy and solutions considering system redundancy. This book is also written for reliability engineers and others who model service availability of

systems that include external redundancy, including georedundancy.

ORGANIZATION

The book is organized to enable different audiences to easily access the information they are most interested in. Part 1, “Basics,” gives background on georedundancy and service availability, and is suitable for all readers. Part 2, “Modeling and Analysis of Redundancy,” gives technical and mathematical details of service availability modeling of georedundant configurations, and thus is most suitable for reliability engineers and others with deeper mathematical interest in the topic. Part 3 ‘Recommendations’ offers specific recommendations on architecture, design, specification, testing, and analysis of georedundant configurations. The recommendations section ends with Chapter 15 which offers a summary of the material. Most readers will focus on Parts 1 and 3; reliability engineers will focus on Parts 2 and 3; and readers looking for a high-level summary can focus on Chapter 15, “Summary.”

Part 1—Basics, contains the following chapters:

- *“Service, Risk, and Business Continuity”* reviews risk management, business continuity and disaster recovery in the context of service availability of critical systems.
- *“Service Availability and Service Reliability”* reviews the concepts of service availability and service reliability, including how these key metrics are measured in the field.

Part 2—Modeling and Analysis of Redundancy contains the following chapters:

- *“Understanding Redundancy”* factors redundancy into three broad categories: simplex (no redundancy), internal system redundancy, and

external system redundancy (including co-located and geographically separated configurations). The fundamentals of high-availability mechanisms and modeling of availability improvements from internal redundancy are covered. Criteria for evaluating high-availability mechanisms are also given.

- *"Overview of External Redundancy"* reviews the key techniques and mechanisms that support failure detection and recovery that enable internal and external redundancy. This chapter also reviews the technical differences between local (co-located) and geographically separated redundancy.
- *"External Redundancy Strategy Options"* reviews the three fundamental system-level external redundancy strategies that are used today: manually controlled, system-driven, and client-initiated recovery. Case studies are given to illustrate how these techniques can be integrated to achieve highly available and reliable systems.
- *"Modeling Service Availability with External System Redundancy"* presents mathematical modeling of the service availability benefit of the three external redundancy strategies. First, a generic model that roughly covers all external redundancy strategies is presented to highlight the differences between the recovery strategies; then more practical strategy specific models are presented and analyzed.
- *"Understanding Recovery Timing Parameters"* details how key recovery-related timing parameters used in the mathematical modeling of the previous chapter should be set to optimize the recovery time for the various external redundancy strategies.

- *“Case Study of Client-Initiated Recovery”* uses a domain name system (DNS) cluster as an example of client-initiated recovery to illustrate the concepts and models discussed earlier in this section.
- *“Solution and Cluster Recovery”* considers how clusters of network elements organized into solutions delivering sophisticated services to enterprises and their customers can be recovered together, and discusses the potential benefits of cluster recovery compared to recovery of individual elements.

Part 3—Recommendations contains the following chapters

- *“Georedundancy Strategy”* reviews considerations when engineering the number of sites to deploy a solution across to assure acceptable quality service is highly available to users.
- *“Maximizing Service Availability via Georedundancy”* reviews the architectural characteristics that can maximize the service availability benefit of external system redundancy.
- *“Georedundancy Requirements”* lists sample redundancy requirements for enterprise IT organizations to consider when specifying critical services.
- *“Georedundancy Testing”* discusses how the verifiable requirements of the “Georedundancy Requirements” chapter should be tested across the integration, system validation, deployment/installation, and operational lifecycle phases.
- *“Solution Georedundancy Case Study”* discusses analysis, architecture, design, specification, and testing of a hypothetical solution.

- *“Summary”* reviews the feasible improvements in service availability that can be practically achieved by properly configuring solutions and redundant systems.

Since many readers will not be familiar with the principles of Markov modeling of service availability used in this book, a basic overview of Markov modeling of service availability is included as an appendix.

ACKNOWLEDGMENTS

The authors acknowledge Chuck Salisbury for his diligent early work to understand the service availability benefits of georedundancy. The authors are also grateful for Ron Santos’ expert input on DNS. Bill Baker provided extensive comments and shared his valuable insights on this subject. Doug Kimber provided detail and thoughtful review, and the technical reviewers provided excellent feedback that led us to improve the content and flow of the book. Anil Macwan provided guidance on procedural reliability considerations. Ted Lach and Chun Chan provided expert input on several subtle reliability items. Michael Liem provided valuable feedback.

Eric Bauer

Randee Adams

Daniel Eustace

PART 1: BASICS

1

SERVICE, RISK, AND BUSINESS CONTINUITY

Enterprises implement computer-based systems to provide various information services to customers, staff, and other systems. By definition, unavailability of services deemed “critical” to an enterprise poses a significant risk to the enterprise customers or stakeholders. Prolonged unavailability of a critical system—or the information held on that system—can be a business disaster. For example, without access to logistics, inventory, order entry, or other critical systems, an enterprise may struggle to operate; a prolonged outage can cause substantial harm to the business, and a very long duration outage or loss of critical data can cause a business to fail.

This chapter introduces service criticality and the linkage to service availability expectations. Georedundancy and risk management in the context of critical computer-based services is covered, along with business continuity planning, recovery objectives, and strategies.

1.1 SERVICE CRITICALITY AND AVAILABILITY EXPECTATIONS