1

INTRODUCTION

COPYRIGHTED MATERIAL

Customers expect that applications and services deployed on cloud computing infrastructure will deliver comparable service quality, reliability, availability, and latency as when deployed on traditional, native hardware configurations. Cloud computing infrastructure introduces a new family of service impairment risks based on the virtualized compute, memory, storage, and networking resources that an Infrastructure-as-a-Service (IaaS) provider delivers to hosted application instances. As a result, application developers and cloud consumers must mitigate these impairments to assure that application service delivered to end users is not unacceptably impacted. This book methodically analyzes the impacts of cloud infrastructure impairments on application service delivered to end users, as well as the opportunities for improvement afforded by cloud. The book also recommends architectures, policies, and other techniques to maximize the likelihood of delivering comparable or better service to end users when applications are deployed to cloud.

1.1 APPROACH

Cloud-based application software executes within a set of virtual machine instances, and each individual virtual machine instance relies on virtualized compute, memory,

Service Quality of Cloud-Based Applications, First Edition. Eric Bauer and Randee Adams.
© 2014 The Institute of Electrical and Electronics Engineers, Inc. Published 2014 by John Wiley & Sons, Inc.

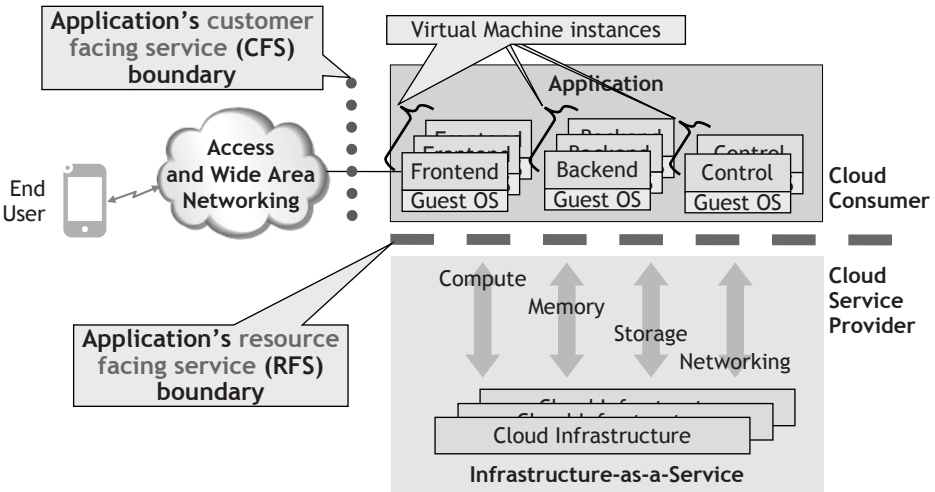

Figure 1.1. Sample Cloud-Based Application.

storage, and networking service delivered by the underlying cloud infrastructure. As shown in Figure 1.1, the application presents **customer facing service** toward end users across the dotted service boundary, and consumes virtualized resources offered by the Infrastructure-as-a-Service provider across the dashed **resource facing service** boundary. The application's service quality experienced by the end users is primarily a function of the application's architecture and software quality, as well as the service quality of the virtualized infrastructure offered by the IaaS across the resource facing service boundary, and the access and wide area networking that connects the end user to the application instance. This book considers both the new impairments and opportunities of virtualized resources offered to applications deployed on cloud and how user service quality experienced by end users can be maximized. By ignoring service impairments of the end user's device, and access and wide area network, one can narrowly consider how application service quality differs when a particular application is hosted on cloud infrastructure compared with when it is natively deployed on traditional hardware.

The key technical difference for application software between native deployment and cloud deployment is that native deployments offer the application's (guest) operating system direct access to the physical compute, memory, storage, and network resources, while cloud deployment inserts a layer of hypervisor or virtual machine management software between the guest operating system and the physical hardware. This layer of hypervisor or virtual machine management software enables sophisticated resource sharing, technical features, and operational policies. However, the hypervisor or virtual machine management layer does not deliver perfect hardware emulation to the guest operating system and application software, and these imperfections can adversely impact application service delivered to end users. While Figure 1.1 illustrates application deployment to a single data center, real world applications are often deployed

to multiple data centers to improve user service quality by shortening transport latency to end users, to support business continuity and disaster recovery, and for other business reasons. Application service quality for deployment across multiple data centers is also considered in this book.

This book considers how application architectures, configurations, validation, and operational policies should evolve so that the acceptable application service quality can be delivered to end users even when application software is deployed on cloud infrastructure. This book approaches application service quality from the end users perspective while considering standards and recommendations from NIST, TM Forum, QuEST Forum, ODCA, ISO, ITIL, and so on.

1.2 TARGET AUDIENCE

This book provides application architects, developers, and testers with guidance on architecting and engineering applications that meet their customers' and end users' service reliability, availability, quality, and latency expectations. Product managers, program managers, and project managers will also gain deeper insights into the service quality risks and mitigations that must be addressed to assure that an application deployed onto cloud infrastructure consistently meets or exceeds customers' expectations for user service quality.

1.3 ORGANIZATION

The work is organized into three parts: context, analysis, and recommendations. **Part I: Context** frames the context of service quality of cloud-based applications via the following:

- *"Application Service Quality"* (Chapter 2). Defines the application service metrics that will be used throughout this work: service availability, service latency, service reliability, service accessibility, service retainability, service throughput, and timestamp accuracy.
- *"Cloud Model"* (Chapter 3). Explains how application deployment on cloud infrastructure differs from traditional application deployment from both a technical and an operational point of view, as well as what new opportunities are presented by rapid elasticity and massive resource pools.
- *"Virtualized Infrastructure Impairments"* (Chapter 4). Explains the infrastructure service impairments that applications running in virtual machines on cloud infrastructure must mitigate to assure acceptable quality of service to end users. The application service impacts of the impairments defined in this chapter will be rigorously considered in Part II: Analysis.

Part II: Analysis methodically considers how application service defined in Chapter 2, "Application Service Quality," is impacted by the infrastructure impairments

enumerated in Chapter 4, "Virtualized Infrastructure Impairments," across the following topics:

- *"Application Redundancy and Cloud Computing"* (Chapter 5). Reviews fundamental redundancy architectures (simplex, sequential redundancy, concurrent redundancy, and hybrid concurrent redundancy) and considers their ability to mitigate application service quality impact when confronted with virtualized infrastructure impairments.
- *"Load Distribution and Balancing"* (Chapter 6). Methodically analyzes work load distribution and balancing for applications.
- *"Failure Containment"* (Chapter 7). Considers how virtualization and cloud help shape failure containment strategies for applications.
- *"Capacity Management"* (Chapter 8). Methodically analyzes application service risks related to rapid elasticity and online capacity growth and degrowth.
- *"Release Management"* (Chapter 9). Considers how virtualization and cloud can be leveraged to support release management actions.
- *"End-to-End Considerations"* (Chapter 10). Explains how application service quality impairments accumulate across the end-to-end service delivery path. The chapter also considers service quality implications of deploying applications to smaller cloud data centers that are closer to end users versus deploying to larger, regional cloud data centers that are farther from end users. Disaster recovery and georedundancy are also discussed.

Part III: Recommendations covers the following:

- *"Accountabilities for Service Quality"* (Chapter 11). Explains how cloud deployment profoundly changes traditional accountabilities for service quality and offers guidance for framing accountabilities across the cloud service delivery chain. The chapter also uses the service gap model to review how to connect specification, architecture, implementation, validation, deployment, and monitoring of applications to assure that expectations are met. Service level agreements are also considered.
- *"Service Availability Measurement"* (Chapter 12). Explains how traditional application service availability measurements can be applied to cloud-based application deployments, thereby enabling efficient side-by-side comparisons of service availability performance.
- *"Application Service Quality Requirements"* (Chapter 13). Reviews high level service quality requirements for applications deployed to cloud.
- *"Virtualized Infrastructure Measurement and Management"* (Chapter 14). Reviews strategies for quantitatively measuring virtualized infrastructure impairments on production systems, along with strategies to mitigate the application service quality risks of unacceptable infrastructure performance.

- *"Analysis of Cloud-Based Applications"* (Chapter 15). Presents a suite of analysis techniques to rigorously assess the service quality risks and mitigations of a target application architecture.
- *"Testing Considerations"* (Chapter 16). Considers testing of cloud-based applications to assure that service quality expectations are likely to be met consistently despite inevitable virtualized infrastructure impairments.
- *"Connecting the Dots"* (Chapter 17). Discusses how to apply the recommendations of Part III to both existing and new applications to mitigate the service quality risks introduced in Part I: Basics and analyzed in Part II: Analysis.

As many readers are likely to study sections based on the technical needs of their business and their professional interest rather than strictly following this work's running order, cross-references are included throughout the work so readers can, say, dive into detailed Part II analysis sections, and follow cross-references back into Part I for basic definitions and follow references forward to Part III for recommendations. A detailed index is included to help readers quickly locate material.

ACKNOWLEDGMENTS

The authors acknowledge the consistent support of Dan Johnson, Annie Lequesne, Sam Samuel, and Lawrence Cowsar that enabled us to complete this work. Expert technical feedback was provided by Mark Clougherty, Roger Maitland, Rich Sohn, John Haller, Dan Eustace, Geeta Chauhan, Karsten Oberle, Kristof Boeynaems, Tony Imperato, and Chuck Salisbury. Data and practical insights were shared by Karen Woest, Srujal Shah, Pete Fales, and many others. Bob Brownlie offered keen insights into service measurements and accountabilities. Expert review and insight on release management for virtualized applications was provided by Bruce Collier. The work benefited greatly from insightful review feedback from Mark Cameron. Iraj Saniee, Katherine Guo, Indra Widjaja, Davide Cherubini, and Karsten Oberle offered keen and substantial insights. The authors gratefully acknowledge the external reviewers who took time to provide through review and thoughtful feedback that materially improved this book: Tim Coote, Steve Woodward, Herbert Ristock, Kim Tracy, and Xuemei Zhang.

The authors welcome feedback on this book; readers may e-mail us at Eric .Bauer@alcatel-lucent.com and Randee.Adams@alcatel-lucent.com.