Service Quality of Cloud-Based Applications

Eric Bauer
Randee Adams

IEEE
IEEE PRESS
WILEY

Table of Contents

IEEE Press
445 Hoes Lane
Piscataway, NJ 08854

Technical Reviewers
Kim W. Tracy, *Northeastern Illinois University*
Rocky Heckman, *CISSP, Architect Advisor, Microsoft*

SERVICE QUALITY OF CLOUD-BASED APPLICATIONS

Eric Bauer
Randee Adams

IEEE PRESS

WILEY

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.

Service quality of cloud-based applications / Eric Bauer, Randee Adams.

pages cm

ISBN 978-1-118-76329-2 (cloth)

1. Cloud computing. 2. Application software–Reliability. 3. Quality of service (Computer networks) I. Adams, Randee. II. Title.

QA76.585.B3944 2013

004.67'82–dc23

2013026569

Figures

Tables and Equations

Tables

Equations

1

Introduction

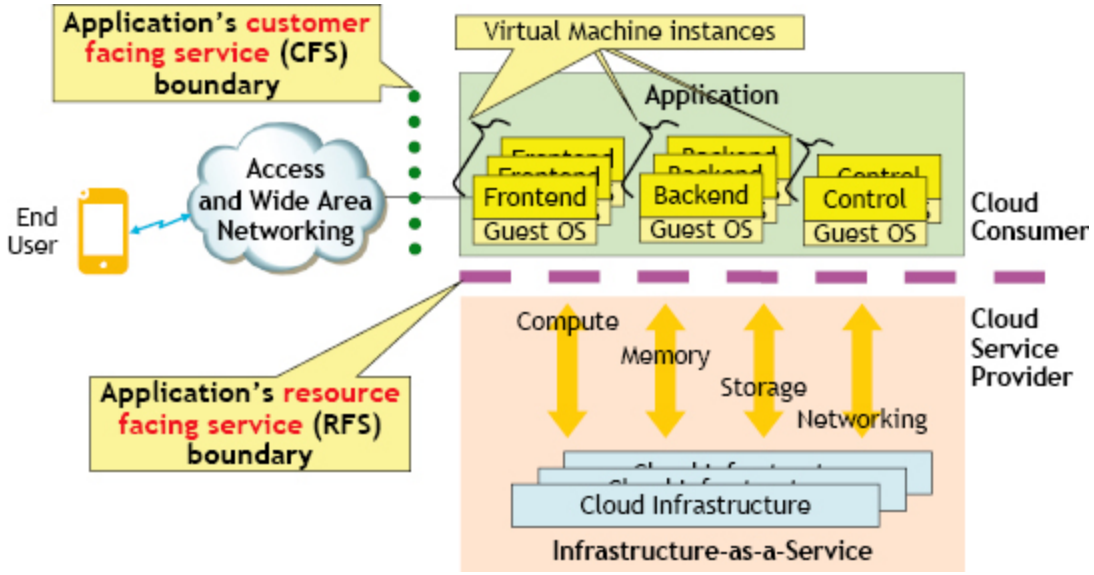
Customers expect that applications and services deployed on cloud computing infrastructure will deliver comparable service quality, reliability, availability, and latency as when deployed on traditional, native hardware configurations. Cloud computing infrastructure introduces a new family of service 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, 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.

Figure 1.1. Sample Cloud-Based Application.



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.

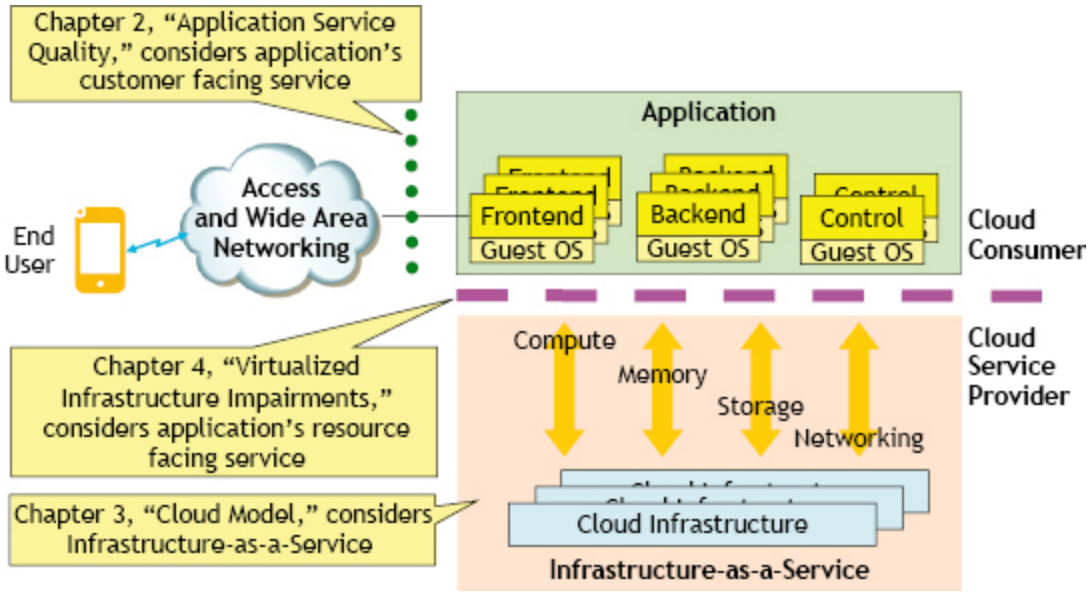
I

Context

Figure 2.0 frames the context of this book: cloud-based applications rely on virtualized compute, memory, storage, and networking resources to provide information services to end users via access and wide area networks. The application's primary quality focus is on the user service delivered across the application's customer facing service boundary (dotted line in Figure 2.0).

- *Chapter 2, "Application Service Quality,"* focuses on application service delivered across that boundary. The application itself relies on virtualized computer, memory, storage, and networking delivered by the cloud service provider to execute application software.
- *Chapter 3, "Cloud Model,"* frames the context of the cloud service that supports this virtualized infrastructure.
- *Chapter 4, "Virtualized Infrastructure Impairments,"* focuses on the service impairments presented to application components across the application's resource facing service boundary.

Figure 2.0. Organization of Part I: Context.


Chapter 2, "Application Service Quality," considers application's customer facing service
Application
Access and Wide Area Networking
End User
Frontend
Guest OS
Backend
Guest OS
Control
Guest OS
Cloud Consumer
Chapter 4, "Virtualized Infrastructure Impairments," considers application's resource facing service
Compute
Memory
Storage
Networking
Cloud Service Provider
Chapter 3, "Cloud Model," considers Infrastructure-as-a-Service
Cloud Infrastructure
Infrastructure-as-a-Service

2

Application Service Quality

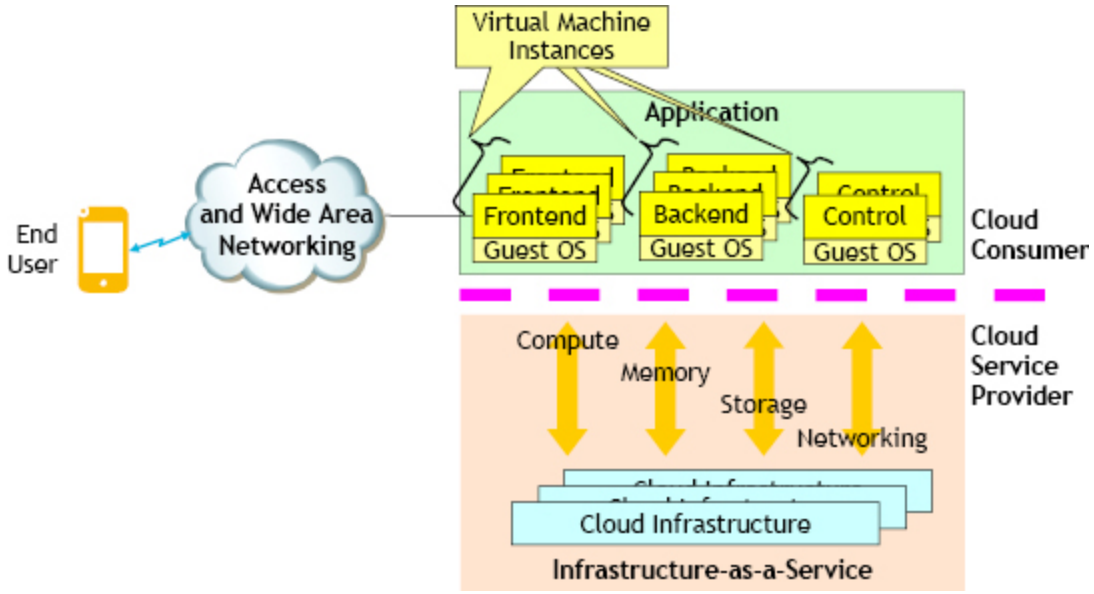
This section considers the service offered by applications to end users and the metrics used to characterize the quality of that service. A handful of common service quality metrics that characterize application service quality are detailed. These user service key quality indicators (KQIs) are considered in depth in Part II: Analysis.

2.1 Simple Application Model

Figure 2.1 illustrates a simple cloud-based application with a pool of frontend components distributing work across a pool of backend components. The suite of frontend and backend components is managed by a pair of control components that provide management visibility and control for the entire application instance. Each of the application's components, along with their supporting guest operating systems, execute in distinct virtual machine instances served by the cloud service provider. The Distributed Management Task Force (DMTF) defines ***virtual machine*** as:

> the complete environment that supports the execution of guest software. A virtual machine is a full encapsulation of the virtual hardware, virtual disks, and the metadata associated with it. Virtual machines allow multiplexing of the underlying physical machine through a software layer called a hypervisor. [DSP0243]

Figure 2.1. Simple Cloud-Based Application.



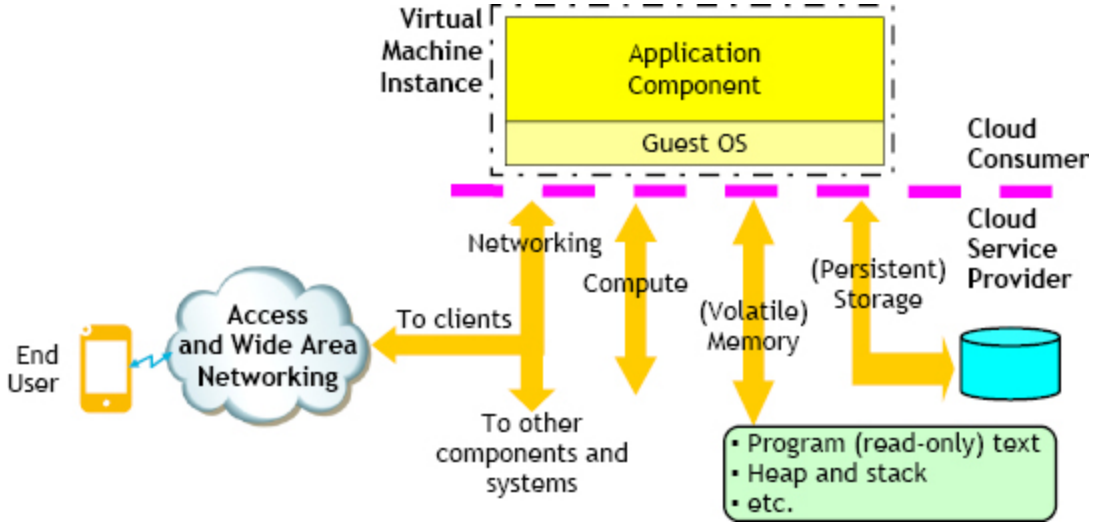
For simplicity, this simple model ignores systems that directly support the application, such as security appliances that protect the application from external attack, domain name servers, and so on.

Figure 2.2 shows a single application component deployed in a virtual machine on cloud infrastructure. The application software and its underlying operating system—referred to as a *guest* OS—run within a virtual machine instance that emulates a dedicated physical server. The cloud service provider's infrastructure delivers the following resource services to the application's guest OS instance:

- *Networking.* Application software is networked to other application components, application clients, and other systems.
- *Compute.* Application programs ultimately execute on a physical processor.
- *(Volatile) Memory.* Applications execute programs out of memory, using heap memory, stack storage, shared memory, and main memory to maintain dynamic data, such as application state
- *(Persistent) Storage.* Applications maintain program executables, configuration, and application data on

persistent storage in files and file systems.

Figure 2.2. Simple Virtual Machine Service Model.



2.2 Service Boundaries

It is useful to define boundaries that demark applications and service offerings to better understand the dependencies, interactions, roles, and responsibilities of each element in overall user service delivery. This work will focus on the two high-level application service boundaries shown in Figure 2.3:

- Application's **customer facing service** (CFS) boundary (dotted line in Figure 2.3), which demarks the edge of the application instance that faces users. User service reliability, such as call completion rate, and service latency, such as call setup, are well-known service quality measurements of telecommunications customer facing service.
- Application's **resource facing service** (RFS) boundary (dashed line in Figure 2.3), which demarks the boundary between the application's guest OS instances executing in virtual machine instances and the virtual compute, memory, storage, and networking provided by the cloud