CISCO

The Economics of Cloud Computing

An Overview for Decision Makers

Bill Williams

Foreword by George Reese,
author of *Cloud Application Architectures*

ciscopress.com

The Economics of Cloud Computing

Bill Williams

Cisco Press

800 East 96th Street
Indianapolis, Indiana 46240 USA

The Economics of Cloud Computing

Bill Williams

Published by:
Cisco Press
800 East 96th Street
Indianapolis, IN 46240 USA

Printed in the United States of America

First Printing June 2012

Library of Congress Cataloging-in-Publication Data:

Williams, Bill, 1970-
The economics of cloud computing / Bill Williams.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-58714-306-9 (pbk. : alk. paper) — ISBN 1-58714-306-2 (pbk. : alk. paper)
1. Cloud computing. 2. Information technology—Economic aspects. I. Title.

QA76.585.W55 2012
004.6782—dc23

ISBN-13: 978-1-58714-306-9

ISBN-10: 1-58714-306-2

Warning and Disclaimer

This book is designed to provide information about the economic impact of cloud computing adoption. Every effort has been made to make this book as complete and as accurate as possible, but no warranty or fitness is implied.

The information is provided on an "as is" basis. The author, Cisco Press, and Cisco Systems, Inc. shall have neither liability nor responsibility to any person or entity with respect to any loss or damages arising from the information contained in this book or from the use of the discs or programs that may accompany it.

The opinions expressed in this book belong to the author and are not necessarily those of Cisco Systems, Inc.

Publisher
Paul Boger

Associate Publisher
Dave Dusthimer

Business Operation Manager, Cisco Press
Anand Sundaram

Manager Global Certification
Erik Ullanderson

Executive Editor
Mary Beth Ray

Managing Editor
Sandra Schroeder

Senior Project Editor
Tonya Simpson

Copy Editor
John Edwards

Editorial Assistant
Vanessa Evans

Cover Designer
Sandra Schroeder

Composition
Mark Shirar

Indexer
Cheryl Lenser

Proofreader
Sheri Cain

Trademark Acknowledgments

All terms mentioned in this book that are known to be trademarks or service marks have been appropriately capitalized. Cisco Press or Cisco Systems, Inc., cannot attest to the accuracy of this information. Use of a term in this book should not be regarded as affecting the validity of any trademark or service mark.

Corporate and Government Sales

The publisher offers excellent discounts on this book when ordered in quantity for bulk purchases or special sales, which may include electronic versions and/or custom covers and content particular to your business, training goals, marketing focus, and branding interests. For more information, please contact:

U.S. Corporate and Government Sales
1-800-382-3419
corpsales@pearsontechgroup.com

For sales outside the United States, please contact:
International Sales
international@pearsoned.com

Feedback Information

At Cisco Press, our goal is to create in-depth technical books of the highest quality and value. Each book is crafted with care and precision, undergoing rigorous development that involves the unique expertise of members from the professional technical community.

Readers' feedback is a natural continuation of this process. If you have any comments regarding how we could improve the quality of this book, or otherwise alter it to better suit your needs, you can contact us through email at feedback@ciscopress.com. Please make sure to include the book title and ISBN in your message.

We greatly appreciate your assistance.

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV
Amsterdam, The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at **www.cisco.com/go/offices.**

About the Author

Bill Williams is a 16-year information technology veteran. Fourteen of those years have been with Cisco Systems, where he has held several leadership positions. Currently, Bill is a regional manager for data center and virtualization technologies, covering the service provider market segment. In 2008, 2010, and 2011, Bill lead the top-producing service provider regions in the United States and Canada. In 2010, Bill won the Manager Excellence award.

Bill attended the University of North Carolina at Chapel Hill and holds master's degrees from Harvard Divinity School and the UNC Kenan-Flagler Business School. Bill also holds U.S. Patent 7260590 for a content delivery application.

The Economics of Cloud Computing is Bill's second book for Cisco Press. *The Business Case for Storage Networks* was published in 2004.

Bill lives with his wife and children in Chapel Hill, North Carolina.

Dedication

This book is dedicated to Lia, Isabel, Lee, and Catherine. To the Dream Team: Thank you for making it all worthwhile.

Acknowledgments

First and foremost, I'd like to thank my manager and friend, Curt Reid, for his support and guidance throughout this process. Curt, your continued leadership and thoughtful insights will always remain priceless in my book.

To my team, the hardest-working people in show business, thank you for your tireless dedication to the task at hand.

A special thank-you goes to Toby Ford for his commentary and guidance in thinking through the longer-term impact of cloud computing. The world is waiting for your book, Toby.

A huge thank-you goes out to George Reese and Stuart Neumann. George's book, *Cloud Application Architectures*: *Building Applications and Infrastructure in the Cloud*, and Stuart's research at Verdantix on carbon emissions and cloud computing were both instrumental in the thought process behind the book you now hold in your hand. Gentlemen, I cannot thank you enough for your help.

Finally, I must also thank my closest peers and advisors in the industry: Jon Beck, James Christopher, Dominick Delfino, Insa Elliot, Melissa Hinde, Jason Hoffman, Jonathan King, Paul Werner, Ted Stein, Phil Lowden, Dante Malagrino, Frank Palumbo, and Rafi Yahalom. You are all guiding lights in a field filled with stars.

CONTENTS AT A GLANCE

CONTENTS

Foreword

Depending on whom you talk to, cloud computing is either very old or very new. Many cloud computing technologies date back to the 1960s. In fact, it's very hard to point to any single technology and say, "That new thing there is cloud computing." However, cloud adoption—public, private, or otherwise—is a new phenomenon, and the roots of that adoption lie in the economics of cloud computing.

Companies have historically consumed technology as capital expenditure "bursts" combined with fixed operational costs. When you needed a new system, you would finance it separately from your operational budget. The 2000s brought us a one/two punch that challenged that traditional consumption model.

First, the recession in 2001/2002 resulted in a huge downsizing of corporate IT. By the middle of the decade, corporate IT had evolved into a tremendously efficient component of the business. These efficiency gains, however, came at the cost of IT's ability to support strategic business endeavors.

The second punch came in the form of the financial system collapse of 2008. As a result of this economic shock, even the largest companies found it difficult to gain access to affordable capital for new IT projects—or any other capital expenditure, for that matter. Not only did IT now lack the bandwidth to support strategic endeavors, but it also lacked any source of funding to support them.

In 2008 and 2009, the economics of cloud computing were a black-and-white world supporting the simplistic statements, "OPEX good, CAPEX bad" and "public cloud cheap, traditional IT expensive." Q4 2008 and Q1 2009 were parts of an extreme economic situation in which these rules of thumb were more true than not. In fact, I got into cloud computing specifically because capital was so hard to find.

I had a marketing company called Valtira that was working on a new on-demand product offering. The capital expense for this project was insane, and it wasn't clear that the product offering would succeed. We moved into the Amazon cloud in early 2008 (before the crisis hit, but with capital scarce for small companies) to develop this product offering and test it. The advantage of the cloud to us was simple: Without any up-front investment, we could test out a new product offering. If it succeeded, we'd be thrilled to continue spending the money to support its ongoing operations. If it failed, we'd kill it and only be out a few thousand dollars.

In other words, the economics of cloud computing enabled us to take on a strategic project in a weakening economic climate that would never have seen the light of day in a traditional IT setting. That's the true economics of cloud computing.

While it might seem silly from today's economic perspective, the "OPEX good, CAPEX bad" mantra combined with IT's diminished capacity to be a strategic partner in business drove marketers, engineers, salespeople, and HR away from IT into the arms of cloud computing vendors. After these business units tasted the freedom of cloud computing, they have almost always resisted a return to a world in which IT is the gatekeeper to all technology.

Another simplistic idea from the "early days" of cloud computing is that the cloud is cheaper than traditional computing. In many cases, a cloud solution will be cheaper in isolation than a comparable traditional solution. The complex reality is that the agility of cloud computing will result in greater consumption of technology than would occur in a traditional IT infrastructure. The overall costs of the cloud are thus almost always higher—but that can be a good thing!

These simplistic memes about cloud computing economics survive today in spite of the much more complex reality. A strategy based on them is certain to result in unachievable expectations and failed attempts at cloud adoption. Although the comparison of capital expenses versus operational expenses plays a role in this calculus, so many other factors are more important these days. Understanding the true economics of cloud computing is absolutely critical to a mature cloud computing strategy and overall success in the cloud.

— George Reese

Introduction

In my conversations with customers, partners, and peers, one topic seems to bubble to the surface more than any other: How do I financially justify the move to the cloud?

Initially, the notion of a business case for cloud computing seemed almost redundant. It seemed to me that the cost savings associated with cloud computing were self-evident and therefore no further explanation was needed. Based on my conversations with people in the industry—consumers, providers, and manufacturers of IT goods and services—cloud adoption appeared to be a foregone conclusion. Based on the data, cloud implementation was either already well under way or was on the near-term priority list of most IT leaders worldwide.

Yet the reality is otherwise. For many people, the actual journey to the cloud is still fraught with uncertainty and confusion. Spending money on IT services provided *externally*—especially when companies invest millions of dollars a year to implement and operate hardware and software *internally* as part of a long-standing, integrated IT supply chain—crosses a major psychological boundary.

This psychological hurdle, coupled with all the various political implications of "build versus buy" decisions, makes the financial justification of cloud adoption all the more imperative.

Goals and Methods

The most important goal of this book is to help you understand—from an economic standpoint—both the short-term and long-term impacts of cloud computing.

We are in the middle of a major technological and sociological revolution, one that will take years to fully unfold. Evidence of this revolution is everywhere and nowhere all at once. For example, we can now access millions of titles of streamed content from multiple devices in our homes, including tablet computers and smartphones. At the same time, however, the servers that process and distribute this data are quickly becoming invisible. Server virtualization, the primary technical driver for cloud computing, has essentially dissolved the concept of a physical server. In the last 40 years, servers have very literally morphed from massive "big iron" mainframes to nothing more than central processing units (CPU) and memory driven by the network.

Economics—"the dismal science"—is a broad topic touching nearly every aspect of human society. It would be supremely arrogant (if not impossible) to do a thorough economic analysis of how cloud computing will change the world as we know it in an executive-level overview designed for the mainstream reader.

There are a number of pure scientists—professional economists, researchers, and educators (like Federico Etro)—who are far more qualified and proficient at this type of analysis and explication. Etro's work (alongside several others listed in Appendix A) is recommended for readers interested in going two or three (or even *N*) layers beneath the surface.

If you know nothing about cloud computing or finance and you walk away at the end of this book with a fundamental understanding of cloud service and deployment models, of basic financial metrics, and how to apply these concepts together in a business case methodology, I will consider my primary objectives met.

If, on the other hand, you have more than a cursory understanding of cloud computing and the impact the cloud has on IT budgeting and finance, and if you are steeped in both ITIL and capital-budgeting methodologies, feel free to fast-forward. Feel free to fast-forward and imagine how we, as a networked, interconnected global society, can best leverage the extreme economies of scale associated with cloud computing. Imagine how—as the adoption of cloud computing accelerates over the coming years—we can best utilize the power of ubiquitous (and nearly free) computing. If you participate in this thought experiment and share in the ongoing dialogue concerning "the cloud economy," I will consider this effort a success overall.

Who Should Read This Book

This book is meant to serve as a primer on the financial and economic impacts of cloud computing. As such, anyone responsible for making decisions regarding IT solutions and platforms can find value here.

Individuals who work in IT procurement, legal, and finance—persons whose roles are already being impacted by the shift to cloud computing—might be interested in understanding more clearly how the technological revolution that is cloud computing fits in a broader social and historical context.

Finally, people who consider themselves well-versed in the nomenclature and business of cloud computing—people who live, eat, sleep, and breathe the cloud—can be challenged to think more deeply about the potential social and global benefits of cheap and ubiquitous computing.

While my primary concern is to enable good decision-making with respect to adopting cloud platforms, it is my hope that the economic surplus that stems from cloud computing can and will be put to extraordinary use.

How This Book Is Organized

This book is designed to be read straight through, ideally in one sitting. Accordingly, it is concise—only four chapters—and organized in such a manner as to enable you to put the information straight to work.

The core of the book (Chapters 1 through 4) covers the following material:

- **Chapter 1, "What Is Cloud Computing?—The Journey to Cloud":** This chapter defines cloud computing service and deployment models and outlines many common characteristics of clouds. Additionally, this chapter introduces two concepts—the *IT supply chain* and the *value chain*—that can be used to baseline IT costs and justify the investment in cloud computing technologies.
- **Chapter 2, "Metrics That Matter—What You Need to Know":** This chapter introduces concepts essential to the financial analysis and justification of IT solutions. Critical business value measurements are broken into two categories: *indirect metrics* and *direct metrics*. Total cost of ownership (TCO), time to market, opportunity costs, churn rate, productivity, and others are introduced as *indirect metrics*. Payback method, net present value (NPV), return on investment (ROI), return on equity (ROE), and economic value added (EVA) are covered as *direct metrics*.
- **Chapter 3, "Sample Case Studies—Applied Metrics":** This chapter applies the *indirect* and *direct metrics* from Chapter 2 to the implementation of cloud computing solutions and platforms at a fictional startup in the pharmaceutical industry. Software as a Service (SaaS), Infrastructure as a Service (IaaS), and Platform as a Service (PaaS) examples are discussed.
- **Chapter 4, "The Cloud Economy—The Human-Economic Impact of Cloud Computing":** This chapter covers technological revolutions and paradigm changes as related to human development. Analysis in this chapter pertains to cloud computing as both an economic enabler (for established and emerging economies alike) and as a driver for global sustainability.

The supplemental materials include

- **Appendix A, "References":** Included here are books, articles, and papers that are either cited in this manuscript or were consulted during my research.
- **Appendix B, "Decision Maker's Checklist":** Included here are items to consider when choosing to purchase and implement cloud solutions.
- **Glossary:** Commonly used terms and phrases related to cloud computing are defined herein.

1

What Is Cloud Computing?—The Journey to Cloud

This chapter begins with a definition of cloud computing before providing an in-depth look at the following topics:

- *Cloud Service Models*
- *Cloud Deployment Models*

In this chapter, we also compare IT and application delivery processes to manufacturing supply chains. The introduction of Michael Porter's concept of the value chain will be helpful in understanding the IT cost center. Both the supply chain analogy and the value chain concept are used in future chapters to establish a baseline for cost analysis for IT deliverables. Understanding the IT supply chain will in turn simplify the process of cost justification for cloud-computing adoption.

It is often joked that if you ask five people to define cloud computing, you will get ten different definitions. Generally speaking, we seem to want to overcomplicate cloud computing and what the cloud means in real life. While in some cases, there can be complex technologies involved behind the scenes, there is nothing inherently complex about cloud computing.

In fact, the technology behind cloud computing is by and large the easy part. Frankly, the hardest part of cloud computing is the people. The politics of migrating from legacy platforms to the cloud is inherently complicated because the adoption of cloud computing affects the way many people—not just IT professionals—do their jobs. Over time, cloud computing might drastically change some roles so that they are no longer recognizable from their current form, or even potentially eliminate some jobs entirely. Thus, the human-economic implications of adopting and migrating to cloud computing platforms and processes should not be taken lightly.

There are also, of course, countless benefits stemming from the adoption of cloud computing, both in the short term and the longer term. Many benefits of cloud computing in the corporate arena are purely financial, while other network externalities relating to cloud computing will have much broader positive effects. The ubiquity of free or inexpensive computing accessed through the cloud is already impacting both communications in First World and established economies, and research and development, agriculture, and banking in Third World and emerging economies.

Therefore, it is important for decision makers to understand the impact of cloud computing both from a financial and from a sociological standpoint. This understanding begins with a clear definition of cloud computing.

Cloud Computing Defined

Cloud computing is not one single technology, nor is it one single architecture. Cloud computing is essentially the next phase of innovation and adoption of a platform for computing, networking, and storage technologies designed to provide rapid time to market and drastic cost reductions. (We talk more about adoption and innovation cycles in the scope of economic development in Chapter 4, "The Cloud Economy—The Human-Economic Impact of Cloud Computing.")

There have been both incremental and exponential advances made in computing, networking, and storage over the last several years, but only recently have these advancements—coupled with the financial drivers related to economic retraction and recession—reached a tipping point, creating a major market shift toward cloud adoption.

The business workflows (the rules and processes behind business functions like accounts payable and accounts receivable) in use in corporations today are fairly commonplace. With the exception of relatively recent changes required to support regulatory compliance—Sarbanes-Oxley (SOX), Payment Card Industry Data Security Standard (PCI DSS), or the Health Insurance Portability and Accountability Act (HIPAA), for example—most software functions required to pay bills, make payroll, process purchase orders, and so on have remained largely unchanged for many years.

Similarly, the underlying technologies of cloud computing have been in use in some form or another for decades. Virtualization, for example—arguably the biggest technology driver behind cloud computing—is almost 40 years old. Virtualization—the logical abstraction of hardware through a layer of software—has been in use since the mainframe era.[1] Just as server and storage vendors have been using different types of virtualization for nearly four decades, virtualization has become equally commonplace in the corporate network: It would be almost impossible to find a LAN today that does not use VLAN functionality.

In the same way that memory and network virtualization have standardized over time, server virtualization solutions—such as those offered by Microsoft, VMware, Parallels, and Xen—and the virtual machine, or VM, have become the fundamental building blocks of the cloud.

Over the last few decades, the concept of a computer and its role in corporate and academic environments have changed very little, while the physical, tangible reality of the computer has changed greatly: Processing power has more than doubled every two years while the physical footprint of a computer has dramatically decreased (think mainframe versus handheld).[2]

Moore's Law aside, at its most basic level, the CPU takes I/O and writes it to RAM and/or to a hard drive. This simple function allows applications to create, process, and save mission-critical data. Radically increased speed and performance, however, means that this function can be performed faster than ever before and at massive scale. Additionally, new innovations and enhancements to these existing technology paradigms (hypervisor-bypass and Cisco Extended Memory Technology, for example) are changing our concepts of what a computer is and does. (Where should massive amounts of data reside during processing? What functions should the network interface card perform?) This material and functional evolution, coupled with economic and business drivers, are spurring a dramatic market shift toward the cloud and the anticipated creation and growth of many new markets.

1. "The Virtualization Reality: Are hypervisors the new foundation for system software?" Simon Crosby, Xensource and David Brown, Sun Microsystems. Accessed January 2012, http://queue.acm.org/detail.cfm?id=1189289.

2. "Variations of Moore's Law have been applied to improvement over time in disk drive capacity, display resolution, and network bandwidth. In these and many other cases of digital improvement, doubling happens both quickly and reliably." Brynjolfsson, Erik; McAfee, Andrew (2011-10-17). *Race Against The Machine: How the Digital Revolution is Accelerating Innovation, Driving Productivity, and Irreversibly Transforming Employment and the Economy (Kindle Locations 286-289)*. Digital Frontier Press. Kindle Edition.

While it is fair to say that what is truly new about the cloud is the use of innovative and interrelated technologies to solve complex business problems in novel ways, that is not the whole story. Perhaps what is most promising about cloud computing, aside from the breadth of solutions currently available and the functionality and scalability of new and emerging platforms, is the massive potential for future products and solutions developed in and for the cloud. The untapped potential of the cloud and the externalities stemming from consumer and corporate adoption of cloud computing can create significant benefits for both developed and underdeveloped economies.

With a basic understanding of the technology and market drivers behind cloud computing, it is appropriate to move forward with a deeper discussion of what cloud computing means in real life. To do this, we turn to the National Institute of Standards and Technology (NIST).

NIST Definition of Cloud Computing

For the record, here is the definition of cloud computing offered by the National Institute of Standards and Technology (NIST):

> Cloud computing is a model for enabling 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.[3]

This definition is considered the gold standard of definitions for cloud computing, and if we unpack it, we can see why. First, note that cloud computing is a usage model and not a technology. There are multiple different flavors of cloud computing, each with its own distinctive traits and advantages. Using this definition, *cloud computing* is an umbrella term highlighting the similarities and differences in each deployment model while avoiding being prescriptive about the particular technologies required to implement or support a certain platform.

Second, we can see that cloud computing is based on a pool of network, compute, storage, and application resources. Here, we have the first premise for the business value analysis and metrics we use in later chapters. Typically speaking, a total cost of ownership (TCO) analysis starts with tallying the costs of each of the combined elements necessary in a solution. Just like the TCO of automobile ownership includes the cost of gas and maintenance, the TCO of a computing solution includes the cost of software licenses, upgrades, and expansions, as well as power consumption. Just as we will analyze the TCO of the computing status quo (that is, the legacy or noncloud model), treating all the resources in the data center as a pool will enable us to more

3. National Institute of Standards and Technology, "NIST Definition of Cloud Computing," www.nist.gov/itl/cloud/upload/cloud-def-v15.pdf, accessed December 2011.

accurately quantify the business value of cloud computing as a solution at each stage of implementation.

Finally, we see that the fundamental benefits of cloud computing are provisioning speed and ease of use. Here is the next premise on which we will base the business value analysis for choosing cloud computing platforms: time to market (TTM) and reduction of operational expenditures (OPEX).

OPEX reductions related to provisioning costs—the costs associated with moves, adds, changes (MAC) necessary to provide and support a computing solution—coupled with reducing the time to implement (TTI) a platform are the principal cost benefits of cloud computing. The former is a measure of reducing ongoing expenses, while the latter is a measure of how quickly we can generate the benefits related to implementing a solution.

Whether it is a revenue-generating application, as in the case of a service provider monitoring network performance, or whether it is a business-critical platform supporting, say, accounts receivable, the measurements used to quantify the associated benefits are essentially the same.

Characteristics of Clouds

The NIST definition also highlights five essential characteristics of cloud computing:

- Broad network access
- On-demand self-service
- Resource pooling
- Measured service
- Rapid elasticity[4]

Let's step through these concepts individually.

First, we cover *broad network access*. Access to resources in the cloud is available over multiple device types. This not only includes the most common devices (laptops, workstations, and so on) but also mobile phones, thin clients, and the like. Contrast broad network access with access to compute and network resources during the mainframe era. Compute resources 40 years ago were scarce and costly. To conserve those resources, usage was limited based on priority and criticality of workloads. Similarly, network resources were also scarce. IP-based networks were not in prevalent usage four decades ago; consequently, access to ubiquitous high-bandwidth, low-latency networks did not exist. Over time, costs associated with the

4. National Institute of Standards and Technology, "NIST Definition of Cloud Computing," www.nist.gov/itl/cloud/upload/cloud-def-v15.pdf, accessed December 2011.

network (like costs associated with computing and storage) have decreased because of manufacturing scalability, commoditization of associated technologies, and competition in the marketplace. As network bandwidth has increased, network access and scalability have also increased accordingly. Broad network access can and should be seen both as a trait of cloud computing and as an enabler.

On-demand self-service is a key—some say the primary—characteristic of the cloud. Think of IT as a complex supply chain with the application and the end user at the tail end of the chain. In noncloud environments, the ability to self-provision resources fundamentally disrupts most (if not all) of the legacy processes of corporate IT. This includes workflow related to procurement and provisioning of storage, servers, network nodes, software licenses, and so on.

Historically, capacity planning has been performed in "silos" or in isolated organizational structures with little or no communication between decision makers and stakeholders. In noncloud or legacy environments, when the end user can self-provision without interacting with the provider, the downstream result is usually extreme inefficiency and waste.

Note

In his classic *Competitive Advantage: Creating and Sustaining Superior Performance*, Michael Porter outlined the concept of the *value chain*. Porter's work highlights how firms can increase their competitive advantage by understanding and optimizing the support and operational functions related to bringing products to market.

In short, Porter breaks down the functional components of the firm into fundamental building blocks: *primary* and *support activities.* Primary activities include inbound and outbound logistics, operations, service, and sales and marketing. Support activities include processes like procurement and human resources. Within primary and support activities, there are *direct*, *indirect*, and *quality assurance* activities that directly create value, indirectly contribute to value creation, or ensure the quality of other processes.[5] Each of these are areas that are touched or will be touched by the adoption of cloud computing.

Porter analyzes economies and diseconomies of scale related to value chain activities, indicating that economies of scale increase with both operating efficiencies and capacity utilization.[6] Analysis of the IT supply chain and the use of simple cost-accounting methodologies will show that adoption of cloud computing can positively influence operational efficiency and capacity utilization, and thereby increase economies of scale.

5. Michael E. Porter, *Competitive Advantage: Creating and Sustaining Superior Performance*, The Free Press, New York, 1985, pp. 41–44.

6. Ibid, p. 70.

Self-provisioning in noncloud environments causes legacy processes and functions—such as capacity planning, network management (providing quality of service [QoS]), and security (management of firewalls and access control lists [ACL])—to grind to a halt or even break down completely. The well-documented "bullwhip effect" in supply chain management—when incomplete or inaccurate information results in high variability in production costs—applies not only to manufacturing environments but also to the provisioning of IT resources in noncloud environments.[7]

Cloud-based architectures, however, are designed and built with self-provisioning in mind. This premise implies the use of fairly sophisticated software frameworks and portals to manage provisioning and back-office functions. Historically, the lack of commercial off-the-shelf (COTS) software purpose-built for cloud automation led many companies to build their own frameworks to support these processes. While many companies do still use homegrown portals, adoption of COTS software packages designed to manage and automate enterprise workloads has increased as major ISVs and startups alike find ways to differentiate their solutions.

Resource pooling is a fundamental premise of scalability in the cloud. Without pooled computing, networks, and storage, a service provider must provision across multiple silos (discrete, independent resources with few or no interconnections.) Multitenant environments, where multiple customers share adjacent resources in the cloud with their peers, are the basis of public cloud infrastructures. With multitenancy, there is an inherent increase in operational expenditures, which can be mitigated by certain hardware configurations and software solutions, such as application and server profiles.

Imagine a telephone network that is not multitenant. This is extremely difficult to do: It would imply dedicated circuits from end to end, all the way from the provider to each and every consumer. Now imagine the expense: not only the exorbitant capital costs of the dedicated hardware but also the operating expenses associated with maintenance. Simple troubleshooting processes would require an operator to authenticate into multiple thousands of systems just to verify access. If a broader system issue affected more than one network, the mean time to recovery (MTTR) would be significant. Without resource pooling and multitenancy, the economics of cloud computing do not make financial sense.

Measured service implies that usage of these pooled resources is monitored and reported to the consumer, providing visibility into rates of consumption and associated costs. Accurate measurement of resource consumption, for the purposes of

7. The bullwhip effect and supply chain management have been widely studied and documented. "The Bullwhip Effect in Supply Chains," by Hau L. Lee, V. Padmanabhan, and Seungjin Whang, is a classic in this field. MIT Sloan Management Review, http://sloanreview.mit.edu/the-magazine/1997-spring/3837/the-bullwhip-effect-in-supply-chains/, accessed December 2011.

chargeback (or merely for cross-departmental reporting and planning), has long been a wish-list item for IT stakeholders. Building and supporting a system capable of such granular reporting, however, has always been a tall order.

As computing resources moved from the command-and-control world of the mainframe (where measurement and reporting software was built in to the system) to the controlled chaos of open systems and client-server platforms (where measurement and reporting were bolted on as an afterthought, if at all), visibility into costs and consumption has become increasingly limited. Frequently enough, IT teams have built systems to monitor the usage of one element (the CPU, for example) while using COTS software for another element (perhaps storage).

Tying the two systems together, however, across a large enterprise often becomes a full-time effort. If chargeback is actually implemented, it becomes imperative to drop everything else when the COTS vendor releases a patch or an upgrade; otherwise, access to reporting data is lost. Assuming that usage accounting and reporting are handled accordingly, billing then becomes yet another internal IT function requiring management and full-time equivalent (FTE) resources. Measured service, in terms of the cloud, takes the majority of the above effort out of the equation, thereby dramatically reducing the associated operational expense.

The final trait highlighted in the NIST definition of cloud computing is *rapid elasticity*. Elastic resources are critical to reducing costs and decreasing time to market (TTM). Indeed, the notion of elastic computing in the IT supply chain is so desirable that Amazon even named its cloud platform Elastic Compute Cloud (EC2). As I demonstrate in later chapters, the majority of the costs associated with deploying applications stems from provisioning (moves, adds, and changes, or MAC) in the IT supply chain. Therefore, simplifying the provisioning process can generate significant cost reductions and enable faster revenue generation.

Think of the workflow and business processes related to the provisioning of a simple application. Whether the application is for external customers or for internal employees, the provisioning processes are often similar (if not identical.) The costs associated with a delayed customer release, however, can be significantly higher. The opportunity costs of a delayed customer-facing application in a highly competitive market can be exorbitant, particularly in terms of customer acquisition and retention. In short, the stakes are much higher with respect to bringing revenue-generating applications to market. We look at different methods of measuring the impact of time-to-market in Chapter2, “Metrics That Matter—What You Need to Know.”

For a simple application (either internal or external) the typical workflow will look something like the following. Disk storage requirements are gathered prompting the storage workflow—logical unit number (LUN) provisioning and masking, file system creation, and so on. A database is created and disks are allocated. Users are created on the server and the associated database, and privileges are assigned based on roles and responsibilities. Server and application access is granted on the network based on ACLs and IP address assignments.

At each step of this process functional owners (network, storage, and server administrators) have the opportunity to preprovision resources in advance of upcoming requests. Unfortunately, there is also the opportunity for functional owners to overprovision to limit the frequency of requests and to mitigate delays in the supply chain.

Overprovisioning in any one function, however, can also lead to deprivation and delays in the next function, thereby igniting the aforementioned bullwhip effect.[8] The costs associated with the bullwhip effect in a typical IT supply chain can be significant. Waste associated with poor resource utilization can easily cost multiple millions of dollars a year in a medium to large enterprise. Delays in deprovisioning unused or unneeded resources also add to this waste factor, increasing poor utilization rates. Imagine the expense of a hotel with no capability to book rooms. That unlikely scenario occurs frequently in IT when projects are cancelled or discontinued. Legacy funding models assume allocated capital expenditures (CAPEX) are constantly in use, always generating a return. The reality is otherwise: The capability to quickly decommission and reassign hardware outside the cloud does not exist, so costly resources can remain idle much of their useful lives.

In a cloud-based architecture, resources can be provisioned so quickly as to appear unlimited to the consumer. If there is one single hallmark trait of the cloud, it is likely this one: the ability to flatten the IT supply chain to provision applications in a matter of minutes instead of days or weeks.

Of these essential characteristics, the fifth—rapid elasticity, or the ability to quickly provision and deprovision—is perhaps the most critical in terms of cost savings relative to legacy architectures.

The NIST definition also includes the notion of *service* and *deployment* models. For a more complete picture of what is meant by the term *cloud computing*, it is necessary to spend a few minutes with these concepts.

Cloud Service Models

- Software as a Service (SaaS)
- Platform as a Service (PaaS)
- Infrastructure as a Service (IaaS)

8. An in-depth analysis of the bullwhip effect in manufacturing, wholesale, and retail can be found at http://opim.wharton.upenn.edu/~cachon/pdf/bwv2.pdf.

Cachon, Randall, and Schmidt: "In Search of the Bullwhip Effect," *Manufacturing & Service Operations Management 9(4)*, pp. 457–479. INFORMS, accessed January 2012.

Software as a Service

Software as a Service (SaaS) is the cloud service model with which most individuals are familiar, even if they do not consider themselves cloud-savvy. Google's Gmail, for example, is one of the most widely known and commonly used SaaS platforms existing today.

SaaS, simply put, is the ability to use a software package on someone else's infrastructure. Gmail differs from typical corporate email platforms like Microsoft Exchange in that the hardware and the software supporting the mail service do not live on corporate-owned, IT-managed servers—the infrastructure supporting Gmail belongs to Google. The ability to use email without implementing expensive hardware and complex software on-site offers great flexibility (and cost reductions) to even small- and medium-sized businesses.

Customer relationship management (CRM) SaaS packages such as Salesforce.com also have significant adoption rates in corporate environments for exactly the same reasons. The increased adoption rate of SaaS in corporate IT stems from SaaS platforms' ability to provide all the benefits of a complex software package while mitigating (if not eliminating entirely) the challenges seen with legacy software environments.[9]

We look at a specific example in Chapter 3, "Sample Case Studies—Applied Metrics," but consider the following: SaaS models enable customers to use vendors' software without the CAPEX associated with the hardware required to run the platform, and without the OPEX associated with managing that hardware. Significant OPEX reductions are also related to the elimination of ongoing maintenance and support. For example, using a SaaS model, when a new release of the software is available, it can simply be pushed out "over the wire," removing the need for complex upgrades, which normally would require hours of FTE time to test and implement.

Infrastructure as a Service

Infrastructure as a Service (IaaS) can almost be seen as the inverse of Software as a Service. With an IaaS model, the service provider delivers the necessary hardware resources (network, compute, storage) required to run a customer's applications.

9. The costs associated with ERP implementations have been researched and documented heavily. Of particular note are the implications for developing countries. See Huang, Z. and Palvia, P. "ERP Implementation Issues in Advanced and Developing Countries." *Business Process Management Journal*. Vol 7, No 3, 2001, pp. 276–284. See also "Why ERP may not be Suitable for Organisations in Developing Countries in Asia," by Rajapakse, Jayanatha, and Seddon, Peter B.

Service providers who have built their businesses on colocation services are typically inclined to offer IaaS cloud service models. Colocation service providers (such as Terremark's NAP of the Americas, Switch and Data, and Level 3, as well as many others) have significant investments in networking infrastructure designed to provide high-bandwidth connectivity for services such as video, voice, and peering.[10]

IaaS service models allow customers to take advantage of these massively scalable networks and data centers at a fraction of the cost associated with building and managing their own infrastructures.

Platform as a Service

Finally, Platform as a Service (PaaS) is best described as a development environment hosted on third-party infrastructure to facilitate rapid design, testing, and deployment of new applications. PaaS environments are often used as application "sandboxes," where developers are free to create (and in a sense improvise) in an environment where the cost of consuming resources is greatly reduced.

Google App Engine, VMware's SpringSource, and Amazon's Amazon Web Services (AWS) are common examples of PaaS offerings. PaaS service models offer customers the ability to quickly build, test, and release software products—with often complex requirements for add-on services—using infrastructure that is purpose-built for application development. Adopting PaaS service models thereby eliminates the need for costly infrastructure buildup and teardown typically seen in most corporate development environments.

Given the increased demand for new smartphone applications, it should come as no surprise that of the three cloud computing service models, PaaS currently has the highest growth rate.[11]

Cloud Deployment Models

To close out our discussion of what cloud computing is and is not, we should review one more element highlighted in the NIST definition of cloud computing: deployment models.

---

10. The Colocation Service Provider Directory, www.colocationprovider.org/whatiscolocation.htm, accessed December 2011.

11. 7Economy Global Economy Library, "Cloud Computing: PaaS: Application Development and Deployment Platform in the Cloud," http://7economy.com/archives/6857, accessed December 2011.

Our gold standard of cloud computing definitions calls out the following deployment models:

- Private cloud
- Community cloud
- Public cloud
- Hybrid cloud

Let us briefly walk through each of these models.

Private Cloud

Using the notion of "siloed infrastructures," many corporate IT environments today could be considered private clouds in that they are designed and built by and for a single customer to support specific functions critical for the success of a single line of business.

In today's parlance, however, a private cloud might or might not be hosted on the customer's premises. Correspondingly, a customer implementing his own private cloud on-premise might not achieve the financial benefits of a private cloud offered by a service provider that has built a highly scalable cloud solution. An in-depth analysis of costs associated with legacy platforms should highlight the differences between today's private clouds and yesterday's legacy silos.

It should also go without saying that legacy silos are not true private clouds because they do not embody the five essential characteristics we outlined earlier.

Community Cloud

In a community cloud model, more than one group with common and specific needs shares the cloud infrastructure. This can include environments such as a U.S. federal agency cloud with stringent security requirements, or a health and medical cloud with regulatory and policy requirements for privacy matters. There is no mandate for the infrastructure to be either on-site or off-site to qualify as a community cloud.

Public Cloud

The public cloud deployment model is what is most often thought of as a cloud, in that it is multitenant capable and is shared by a number of customers/consumers who likely have nothing in common. Amazon, Apple, Microsoft, and Google, to name but a few, all offer public cloud services.

Hybrid Cloud

A hybrid cloud deployment is simply a combination of two or more of the previous deployment models with a management framework in place so that the environments appear as a single cloud, typically for the purposes of "cloud peering" or "bursting." Expect demand for hybrid cloud solutions in environments where strong requirements for security or regulatory compliance exist alongside requirements for price and performance.

Note that major cloud providers typically offer one or more of these types of deployment and service models. For example, Amazon AWS offers both PaaS and public cloud services. Terremark offers private and community clouds with specialized hybrid cloud offerings, colocation and exchange point services, and cost-efficient public cloud services through vCloud Express.[12]

Note

To determine the best cloud offering for your business, it is important to understand (or at least have a good idea of) your compute, storage, and networking requirements. It is helpful to know your budget and your total cost of ownership (TCO) metrics as well. Cloud computing providers will work with you to help you scope your environments for the purposes of sizing and capacity planning. Most providers will even help you determine an estimated return on investment (ROI) for your migration to the cloud.

While it is important for you to understand your infrastructure requirements, it is most critical for you to understand both your business processes and goals, and your underlying application architecture.

A strong knowledge of your critical data—where it lives and how you use it for business-critical decisions and customer success—will enable you to make a well-informed choice about cloud platforms and solutions.

Conclusion

In this chapter, we explored the standard definition of cloud computing to establish a baseline of common terminology. Understanding the essential characteristics of cloud computing platforms, as well as cloud deployment and service models, is critical for making informed decisions and for choosing the appropriate platform for your business needs.

12. See Terremark's most recent 10-K filing: www.faqs.org/sec-filings/100614/TERREMARK-WORLDWIDE-INC_10-K/.

Additionally in this chapter, we introduced Michael Porter's concept of the value chain and drew a comparison among IT infrastructure, application deployments, and manufacturing supply chains. These concepts are key components for understanding the costs (both CAPEX and OPEX) associated with traditional or legacy systems and the offsets potentially achieved by migrating to the cloud.

In the next chapter, we look at the business metrics most often used to measure the impact of technology adoption and implementation.

Index

D

E

T

U

V

W