vmware PRESS

VMware vCloud® Architecture Toolkit (vCAT)

Technical and Operational Guidance for Cloud Success

VMware vCAT Team

Foreword by **Pat Gelsinger**, CEO, VMware

VMware vCloud® Architecture Toolkit (vCAT)

Technical and Operational Guidance for Cloud Success

VMware Press is the official publisher of VMware books and training materials, which provide guidance on the critical topics facing today's technology professionals and students. Enterprises, as well as small- and medium-sized organizations, adopt virtualization as a more agile way of scaling IT to meet business needs. VMware Press provides proven, technically accurate information that will help them meet their goals for customizing, building, and maintaining their virtual environment.

With books, certification and study guides, video training, and learning tools produced by world-class architects and IT experts, VMware Press helps IT professionals master a diverse range of topics on virtualization and cloud computing and is the official source of reference materials for preparing for the VMware Certified Professional Examination.

VMware Press is also pleased to have localization partners that can publish its products into more than 42 languages, including Chinese (Simplified), Chinese (Traditional), French, German, Greek, Hindi, Japanese, Korean, Polish, Russian, and Spanish.

For more information about VMware Press, visit **vmwarepress.com**.

This page intentionally left blank

VMware vCloud® Architecture Toolkit (vCAT)

Technical and Operational Guidance for Cloud Success

VMware vCAT Team

vmware® PRESS

Upper Saddle River, NJ • Boston • Indianapolis • San Francisco
New York • Toronto • Montreal • London • Munich • Paris • Madrid
Cape Town • Sydney • Tokyo • Singapore • Mexico City

VMware vCloud Architecture Toolkit (vCAT)

Technical and Operational Guidance for Cloud Success

Published by Pearson Education, Inc.

Publishing as VMware Press

Library of Congress Control Number: 2013910623

ISBN-13: 978-0-321-91202-2

ISBN-10: 0-321-91202-0

Printed in the United States of America

First Printing: August 2013

Warning and Disclaimer

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 provided is on an "as is" basis. The authors, VMware Press, VMware, and the publisher 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 any digital content or programs accompanying it.

The opinions expressed in this book belong to the author and are not necessarily those of VMware.

Corporate and Government Sales

VMware Press 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
(800) 382-3419
corpsales@pearsontechgroup.com

For sales outside the United States, please contact:

International Sales
international@pearsoned.com

VMWARE PRESS PROGRAM MANAGERS
Eric Ullanderson
Anand Sundaram

ASSOCIATE PUBLISHER
David Dusthimer

AQUISITIONS EDITOR
Joan Murray

DEVELOPMENT EDITOR
Ellie Bru

MANAGING EDITOR
Sandra Schroeder

PROJECT EDITOR
Mandie Frank

COPY EDITOR
Krista Hansing
Editorial Services

PROOFREADER
Sheri Replin

INDEXER
Rebecca Salerno

EDITORIAL ASSISTANT
Vanessa Evans

DESIGNER
Chuti Prasertsith

COMPOSITOR
Trina Wurst

Contents

About the Contributors

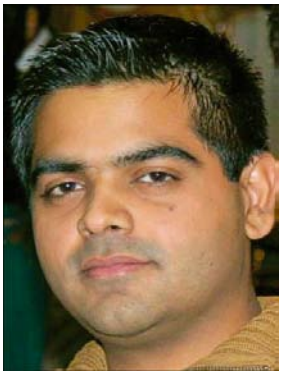
The following product owners have led the vCAT 3.x development effort upon which this VMware Press release is based.

John Arrasjid—John Arrasjid is a Principal Architect at VMware, Inc., where he started in 2003. John is part of the Global Technology Solutions team, is a VMware Ambassador, and is part of the Field Office of the CTO. John was awarded the vExpert 2012 designation, given to the top VMware evangelists in the industry, for his work on vCAT and the VCDX program. As lead architect and chief product owner of vCAT, John has led the development and release of vCAT since 2011. In his 10 years at VMware, John has co-authored four other books; *VCDX Boot Camp, Cloud Computing with vCloud Director, Foundation for Cloud Computing with vSphere 4, and Deploying the VMware Infrastructure*. John regularly presents at VMworld, VMware Partner Exchange, VMware vForum, USENIX LISA, and other industry conferences. His VCDX Boot Camp has been taught to more than 800 individuals since 2008. John holds a bachelor of science in computer science from SUNY at Buffalo and holds VCDX, ITIL Foundations, and CSPO certifications. He is a founding member of the Elastic Sky band, developer of the original vmsnap/vmres tool, and developer of several consulting engagements for security, business continuity, and performance. Find John on Twitter at @vcdx001.

Matthew Wood—Matthew Wood is an independent technical writer. Matthew has been a senior technical writer, editor, and manager for VMware Technical Services, and he was the lead editor for the vCAT project from 2010 until 2013. Matthew works with architects and consultants to produce IP for services kits and solutions kits related to all aspects of VMware technology. He also has written original documentation for the VMware Services Software Solutions group to support tools such as VMware HealthAnalyzer and Migration Manager. Matthew has 38 years of experience working with technology companies, focusing especially on UNIX, virtualization, and applications that support enterprise IT environments.

Wade Holmes—Wade Holmes is a Staff Solutions Architect at VMware, Inc., and holds VCDX, CISSP, CCSK, and CSPO certifications. He has more than 16 years of experience planning, teaching, and presenting on the architecture, design, and implementation of complex computing environments of all scopes and sizes. Wade has presented and taught at conferences such as VMworld, SXSW, USENIX LISA, and VMware User Group meetings. Wade was awarded the vExpert 2012 designation, given to the top VMware evangelists in the industry. He holds a bachelor's degree in information technology and a master's degree in information assurance. Find him on Twitter at @wholmes; he also maintains a blog at www.vwade.com.

Joe Sarabia—Joe Sarabia is a Cloud Architect at VMware, Inc., and holds industry certifications that include VCAP-DCD, VCAP-DCA, MCSE, NCDA, ITIL, and CSPO. Joe has had various roles in the information technology field. He initially focused on operational roles in areas of organizations that consumed services from the business. About 10 years ago, Joe's role pivoted to architecting and providing services on behalf of the business for business units and end users to consume. He has spent the last several years of his career as one of the leading hybrid cloud and SDDC architects in the industry, focusing on delivering business value to the globe's largest organizations through complex software systems. Joe has particularly established himself as a thought leader in the areas of component integration and end user portal experience. Find him on Twitter at @joesarabia.



Rohan Kalra—Rohan Kalra is a Business Solutions Architect who brings more than 14 years of IT service management consulting experience, including global operations process re-engineering for Fortune 500 clients (EMC, Kellogg's, Banco Santander, Goldman Sachs, Ricoh, and more). Rohan led the creation and release of operational readiness related IP assets available through VMware's professional services, partner channels, and Accelerate Strategy teams. Formerly an executive technology adviser at Accenture, he led the development of operational readiness and governance components of its next-generation infrastructure solution blueprint, focused on cloud computing and delivery of IT as a Service. Rohan holds ITIL and CSPO certifications. Find him on Twitter at @kalrarohan.

Rupen Sheth—Rupen Sheth is a Senior Solutions Engineering Manager on the Global Services team at VMware, Inc., where he is responsible for monetizing and scaling the Software Defined Data Center (SDDC) portfolio of solutions and services. Rupen holds VCDX, ITIL, CSPO, and TOGAF certifications. He has extensive experience in delivering enterprise business and virtualization/cloud solutions through the effective application of information technology, process management, and coordination and management of multidisciplinary teams. Rupen started as a consultant at VMware and now leads a team of solution architects responsible for SDDC solutions and services kits that are used by VMware field and partners worldwide. Rupen has presented and taught at VMworld, VMware Partner Exchange, and USENIX conferences. Find him on Twitter at @rupensheth.

Ian Perez-Ponce—As Senior Product Manager for VMware's vCloud Hybrid Service, Ian is responsible for service-creation and monetization efforts for the company's Infrastructure as a Service (IaaS) cloud solutions portfolio. With more than 14 years of service provider and information technology experience, Ian helps define VMware's premium hybrid cloud service strategy and oversees the development of the vCloud Service Provider partner ecosystem. Find him on Twitter at @iperezponce.

Christophe Decanini—Christophe Decanini is a Consulting Architect at VMware, Inc., where he started in 2007; currently, he is the technical lead for cloud orchestration. Based in Gland, Switzerland, Christophe is a global resource supporting customers in their orchestration and automation needs. He has presented orchestration solutions at conferences such as VMworld and is the main contributor on the www.vcoteam.info blog and in the official VMware Orchestrator community. Christophe was awarded the vExpert designation in 2011, given to the top VMware evangelists in the industry. He has 18 years of experience in IT automation and holds a bachelor's degree in computer science. Find him on Twitter at @vCOTeam.

Burke Azbill—Burke Azbill has been working in IT since the mid-1990s and for VMware since 2007. He has been an active member of the VMworld Hands On Labs and a leading contributor to the vCenter Orchestrator community with both his own blog (www.vcoteam.info) and his contributions to the Official VMware Orchestrator blog and the community in the VMware forums. Burke was awarded the vExpert designation in 2011 and 2012, given to the top VMware evangelists in the industry. His industry certifications include MCP+I, MCSE, MCSD, CNE, CCA, LPIC-1, and VCP. Find him on Twitter at @TechnicalValues.

Michael Haines—Michael Haines is a Senior Cloud Networking and Security Architect and Engineer for the Global Services Engineering team at VMware, Inc. He leads the security architecture and development of VMware's cloud solutions for service providers, enterprise customers, and partners throughout Europe and Asia Pacific. Michael is responsible for providing deep technical expertise and interfacing directly with Engineering and Product Management to support and develop current and future vCloud products and initiatives. He is also involved in prototyping vCloud solutions and frequently presents on VMware's vCloud vision. This includes presentations at VMworld, where he also acts as one of the Security Lab captains. Michael is the co-author of the following publications: *Cloud Computing with VMware vCloud Director, LDAP in the Solaris Operating Environment: Deploying Secure Directory Services, and Deploying LDAP in the Enterprise: Sun BluePrints Publications*. Find him on Twitter at @michaelahaines.

Dave Richey—Dave Richey holds a degree from Harvard and has developed software training materials for more than a decade, including a full curriculum for Mac programmers. He draws on his experience in software development and technical management to edit technical documentation at VMware, Inc., in the fields of virtualization and cloud computing.

Ben Lin—Ben is a Staff Systems Engineer for the Networking and Security Business Unit (NSBU) at VMware, Inc. He holds VCDX3/4/5 certifications and actively participates in VCDX panels and development activities. Ben graduated from the University of California, Berkeley with a bachelor of science in electrical engineering and computer science. Ben co-authored the book *Cloud Computing with VMware vCloud Director* and was closely involved with cloud designs and deployments since the inception of vCloud Director. He is also co-author of VCDX Boot Camp. He regularly presents at conferences such as VMworld, VMworld Europe, Partner Exchange, USENIX LISA, USENIX HotCloud, and vForum. Find him on Twitter at @blin23.

Christopher Knowles—Chris Knowles is a Staff Architect within the Global Center of Excellence (CoE) at VMware, where he works on hybrid cloud and Software Defined Datacenter architecture and integration. Within the CoE, Chris translates complex business requirements into real-world highly integrated infrastructure solutions. Chris leads the VMware LiVefire program, which enables VMware specialists and industry partners to deliver these advanced solutions in the field. When not balancing work and life with his wife, Erin, and two boys, Evan and Spencer, Chris is a regular speaker at VMworld and other industry events. Find him on Twitter at @sugeknowles.

Thomas Kraus—Thomas Kraus works as a Solution Architect in the VMware Networking and Security Business Unit (NSBU) at VMware, Inc., where he helps VMware's largest customers rationalize, understand, and deploy network virtualization and Software Defined Datacenters. Thomas is primarily focused on the architecture, troubleshooting, and optimization of complex cloud environments, with a focus on automation and integration. In addition to being a VCDX, his relevant certifications are RHCE and NetApp SVAP. Find him on Twitter at @tkrausjr.

David Hill—David Hill is an experienced entrepreneur, IT consultant, and architect who has worked in the IT industry for more than 16 years on projects across the public sector and financial institutions. David joined VMware in 2010 and is a Senior Solutions Architect in the Professional Services Engineering (PSE) team. There he develops cutting-edge technology best practices, design guidelines, and intellectual property for the company and partners. David holds VCP 3/4/5 and VCAP-DCD4 certifications. David is the author and owner of the cloud technical blog www.virtual-blog.com. Find him on Twitter at @davehill99.

This book is dedicated to our families, friends, co-workers, customers, and partners. With you, we have found the time, energy, and enthusiasm to raise the bar and produce something to educate many on the concepts, technology, and operations for cloud computing and the software-defined data center.

—The vCAT Team

"Design is the fundamental soul of a human-made creation that ends up expressing itself in successive outer layers of the product or service."

—Steve Jobs

"I am constantly thinking about new and simple approaches to solving problems. As Albert Einstein said, 'Any intelligent fool can make things bigger and more complex. It takes a touch of genius to move in the opposite direction'. vCAT is a huge enabler for your service-oriented transformation efforts."

—Rupen Sheth, VCDX, ITIL, TOGAF certified

Acknowledgments

The vCATs

vCAT 3.1 Team members

Chief Product Owner

John Yani Arrasjid, Principal Architect

Project Leadership Team

Matthew Wood, Wade Holmes, Joe Sarabia, Rohan Kalra, Nira Metcalf, John Callaghan, Donna Colborn

Product Owners

John Arrasjid, VCDX
Burke Azbill
Christophe Decanini
Michael Haines
David Hill
Wade Holmes, VCDX
Rohan Kalra
Chistopher Knowles
Thomas Kraus, VCDX
Ian Perez-Ponce
Joe Sarabia
Rupen Sheth, VCDX
Matthew Wood

vCAT team logo concept by Catherine Arrasjid

Technical Publications

Matthew Wood, David Richey, Patrick Carri, Barbara Weinstein

Project Management

John Callaghan, Donna Colborn

Marketing & Web

Nira Metcalf, Kathleen Tandy, Adam Souza

Contributors

Deji Akomolafe
Richard Anderson
Kalen Arndt
Richard Benoit
Bill Call
Philip Callahan
Chris Colotti, VCDX
Aidan Dalgleish, VCDX
Massimo Re Ferre
Greg Herzog, VCDX
Bill Keenan
Kevin Lees
Ben Lin, VCDX
Matthew Meyer, VCDX
Hugo Phan, VCDX
Prasad Pimplaskar
Mahesh Rajani, VCDX
Tom Ralph, VCDX
Alan Renouf
Rawlinson Rivera, VCDX
Heman Smith
Timo Sugliani
Andy Troup
Raman Veeramraju

Sponsors

Scott Bajtos, Michael "Dino" Cicciarelli, Carl Eschenbach, Pat Gelsinger, Dr. Stephen Herrod, Paul Maritz, Ray O'Farrell, Raghu Raghuram, Rajagopal Ramanujam, Matthew Stepanski, Dan Smoot, Paul Strong, Yvonne Wassenar

The vCATs

Past vCAT Team members

The following are individuals who have worked on past releases of the vCloud Architecture Toolkit. This includes releases 1.6, 2.0, 2.0.1, and 3.0. Current vCAT 3.1 members have also participated in developing past releases.

As with the current vCAT team, we recognize the value in everyone's contribution, the dedication, and the sacrifices made to deliver this highly used resource.

As of pre-release, vCAT has had over 100,000 downloads used by architects, administrators, operators, consumers, customers, consultants, and vendors.

Product Owners

David Baldwin
Russel Callan
Pang Chen, VCDX
Chris Colotti, VCDX
Aidan Dalgleish, VCDX
Michael DiPetrillo
Ford Donald
Massimo Re Ferre
Jason Karnes
Kevin Lees
Ben Lin, VCDX
Mahesh Rajani, VCDX
Kamau Wanguhu, VCDX

Project Management

Darrel Carson
Bernie Clarke
Mary Toman
John McGinn
Monte Kingstone
Jamal Abdul Kadar

Marketing

Suzanne Ambiel

Contributors

Jason Carolan
Andrew Hald, VCDX
Jeremy Hunt
Randy Keener
Hany Michael
Phillippe Michel
Alex Mittell
Srinivas Muthu
Dushyanth Nataraj
Chirag Patel
Melanie Spencer
John Stanford
Hugo Strydom
Ben Thomas
Patrick Thomas
TJ Vatsa

Foreword

One thing I've learned in my career is that architecture really matters. Bad implementations can be thrown away, but architectures last a long time—sometimes forever! Having a framework and set of principles to both guide and enable innovation can determine success, and a lack thereof almost certainly signals failure. An example I'm intimately familiar with is Intel's x86 architecture. Putting the right framework in place has allowed the X86 design team to continue to create value and introduce new innovations to hundreds of millions of users to this day.

The VMware vCloud Architecture Toolkit (vCAT) serves a similar role for VMware. It provides the best of best practices that guide customers in assembling and operating a Cloud-capable, modern platform based on the Software Defined Data Center (SDDC).

The insight behind the software-defined strategy is that cloud-scale economics and agility require a radically simpler and more flexible approach to managing the hardware, network, storage, and security elements of the data center. This kind of agility requires that every technology layer be software defined and automatable. Networking, storage, compute, and security need to be abstracted, pooled, and made reconfigurable through instructions that are not bound to physical hardware. In a word, they need to be virtualized. VMware is applying its virtualization engineering capability to all the physical layers of the datacenter and extending these capabilities across multiple clouds. This gives our customers the most choice and control in how they deliver IT.

This latest vCAT release guides our customers in moving configuration management, policy management, and provisioning into the software layer. This simplifies the challenges companies increasingly face as software development and IT teams work together and the line between their roles becomes less distinct. vCAT also now supports *hybrid cloud deployments on partners' clouds or on the VMware Hybrid Cloud Service so that customers can deploy workloads on hardware they rent or own*. vCAT can also guide customers in laying the foundation for Desktop as a Service and Platform as a Service.

vCAT and other VMware reference architectures are developed as part of the VMware Validated Architecture program. Our engineering, support, and other technical teams review and validate these reference architectures in our labs and directly through our customers' deployments.

Each day, our customers and our partners come to depend more on VMware technology. This is both exciting and humbling for our company. As VMware continues to play a more central role in the IT industry, we've recognized the need to provide deep technical guidance that helps our customers realize success. We also recognize that our customers need to support existing investments and want to have the option to choose the best technology for their needs. To this end, we've created a way for other industry players to extend vCAT and integrate their products into the SDDC architecture. This also allows

partners to publish their own vCAT-compatible blueprints and design templates that guide our mutual customers in implementation and operation of solutions that incorporate those products. We believe that vCAT will continue to provide necessary and extensible architectural blueprints for the IT industry as it transitions to a software-defined approach to computing.

I heartily recommend this reference as a roadmap for anyone tasked with simplifying IT infrastructure and as an indispensable guide for those developing Software Defined Data Centers and vSphere/vCloud solutions.

Pat Gelsinger

Chief Executive Officer, VMware

Preface

"Technical skill is mastery of complexity, while creativity is mastery of simplicity."

Erik Christopher Zeeman

"What is to be sought in designs for the display of information is the clear portrayal of complexity. Not the complication of the simple; rather, the task of the designer is to give visual access to the subtle and the difficult—that is, revelation of the complex."

Edward Tufte

This book represents the work of more than 100 architects, consultants, administrators, engineers, project managers, technical editors, partners, and customers over multiple releases starting in 2010. A handful of people built the 1.x releases. For the 2.x release, approximately 72 individuals spent nearly 1,200 hours to produce 600 pages of content across eight documents. The 3.x releases saw about 42 individuals spend approximately 1,400 hours to produce 750 pages of content across nine documents. In your hands, you hold a compendium of these individual components in a single book format.

vCAT was created first as a reference architecture based on a limited set of use cases. The current release supports multiple use cases and, as such, has turned into a reference architecture toolkit that is part of a series titled VMware Validated Architectures (VVA).

The following sections present information on the owners of the product sections and a list of the contributors involved in the vCAT project since its inception. Approximately 50% of the development team holds VCDX certification. This material is thus not only a reference for SDDC and vCloud, but also a reference for those planning to achieve VCAP and VCDX certification.

You will notice our internal logo, a black cat on a white cloud. Catherine Arrasjid created this graphic to represent the project and the team of vCATs, as we are affectionately called. Our marketing team digitized it, and has become our internal team mascot.

It has been my pleasure to work on these releases—and to work with such an exceptional team of individuals, who are all recognized in the industry in their fields of specialization.

I want to call out the value of vCAT beyond just the cloud space. As you hear more about the Software Defined Data Center (SDDC) and related Software Defined components in networking, security, storage, and other areas, vCAT can provide the guidance you need. vCAT includes many of the components of SDDC—so what do you need to be aware of? We hope to include that in an addendum to vCAT that provides extensions in the SDDC area. Software Defined Networking and Security cover the areas currently represented by vCloud Networking and Security. Software Defined Storage will add relevant components on virtualization of storage. We expect a few other areas to come as the SDDC space continues to evolve.

VMware Validated Architectures, similar to vCAT, are designed to be easily integrated with third-party reference architectures. You will find references to these on the vendor sites. The goal is to allow ease of plug-and-play with other solutions, VMware, and third-party offerings.

As of publication, the vCAT site at www.vmware.com/go/vcat has more than 200,000 accesses and more than 100,000 downloads. These downloads were created by architects, administrators, operators, developers, project managers, solutions architects, and managers. The feedback has shown how vCAT is used and turned up suggestions to improve what we are producing. If you have input on improving this material, send your suggestions to IPfeedback@vmware.com. Please note that this book is printed in black and white to minimize cost and allow for wider adoption. Color versions of the original documentation in electronic and PDF format can be found at www.vmware.com/go/vcat in the Document Center tool.

We want to thank all participants on this project, with a special callout to our stakeholders who have supported this project and recognize the value it provides to our customers.

As you peruse this material, start by reading the Introduction, a guide to the material included in this book.

I wish you the best in your design and deployment of cloud and software-defined datacenters.

John Yani Arrasjid, VCDX-001

Principal Architect

VMware vCloud Architecture Toolkit R3.1 Release Notes

VMware vCloud Architecture Toolkit R3.1 Release Notes

The VMware® vCloud® Architecture Toolkit (vCAT) provides modular components so that you can design a vCloud reference architecture that supports your cloud use case. It includes design considerations and design patterns to support architects, operators, and consumers of a cloud computing solution based on VMware technologies.

For additional vCAT supporting material, visit the vCloud Architecture Toolkit page at vmware.com (www.vmware.com/go/vCAT). This is also where updates to vCAT will be posted.

vCAT 3.1 Documentation Packages

The following vCAT 3.1 packages are available:

- PDF package (~25MB)
- Documentation Center package (~50MB)

PDF Package

The PDF package is a zipped package that contains PDFs of all the documents in the toolkit. Use WinZip or a similar application to unzip the package, and use a PDF reader such as Adobe Reader to display and read the documents. You can print hard-copy documents from the PDFs.

Documentation Center Package

For a video overview on the Documentation Center packaging of vCAT, see the SME videos at www.vmware.com/go/vcat.

The documentation center package is a zipped package that contains a complete online help system that you can use to view all the documents in the toolkit from an easy-to-use interface. It offers powerful features such as the capability to search through the collection of vCAT documents, display a pregenerated PDF of a document, and, when served from a web server, access Google Translate to translate displayed pages into dozens of languages.

The vCAT 3.1 Documentation Center is also served from a website on vmware.com.

Browsers

The documentation center works with the following browsers:

- Google Chrome (preferred)
- Safari
- Internet Explorer (Search operates differently. Click the magnifying glass to search, enter a search term in the resulting text field, and press Enter or click Go).
- Firefox (works well except with Google Translate)

Installing the Documentation Center

The Documentation Center package is large, so it is recommended that you download the package over a high-speed link.

To install and display the vCAT 3.1 documentation center:

1. Download and unzip the package.
2. Double-click the index.html file to run it.
3. Allow blocked content, if prompted.

Offline Versus Online Capabilities

- If you install the Documentation Center package on your machine, you can use the toolkit offline. For example, you can install it on your laptop computer and review the documents while you are not connected to the Internet.
- If you install the Documentation Center package on a web server, the interface provides additional capabilities, such as access to Google Translate.
- The Documentation Center also optimizes the display for mobile devices. For example, using Safari on an iPad to access the documentation center works well.

Using the Documentation Center

- Click the folder icon to toggle display of the navigation pane on or off.
- Enter text in the search field and click the search icon (magnifying glass) to search for it. This is a client-side search implementation that can be used online or offline. It does not allow Boolean expressions.

If the documentation center is served from an Internet-connected web server, click the globe to display Google Translate. Select the language you want and click Translate. Each page is translated as it is displayed.

- Click a document in the navigation pane to display the sections in the document. Click a section to display content.
- Click the Page Forward or Page Back arrows to move from page to page.
- Click the Print icon to print the selected page to a printer.
- Select any document section and click the PDF icon to view a pregenerated PDF for the selected document. You can print the entire document from the PDF.
- Click the Email (envelope) icon or the link by the logo to send feedback to ipfeedback@vmware.com. The URL of the currently displayed page is automatically populated in the email Subject line.

vCAT 3.1.1 Changes and VMware Press Book Release

This book release combines all separate documents for vCAT 3.1 into one document. Each chapter in this book represents the nine separate documents. When we reference separate sections, see the associated chapter in the book format. We have not made specific updates to these release notes, to allow the material here and the material in Document Center to be synchronized.

There are several updates in this book that will apply to the updated web release.

- General
  - Updated graphics and screenshots.
  - Updated SSO material throughout.
- Chapter 1, Introduction:
  - Removed references to VMware Service Manager.
- Appendix D
  - Removed references to VMware Service Manager.

vCAT 3.1 Changes

For vCAT 3.1, most documents received additional edits, and graphics were improved for many figures. Content was updated as follows:

- *Chapter 1, Introduction:*
  - Links to brief video presentations were added for each document and topic area.
  - Figure 1.3 was updated.
- *Chapter 2, Service Definitions*:
  - The service offering examples were changed because of allocation model changes in vCloud Director 5.1.1.
  - Minor updates were made to the other service offering examples.
  - Other minor edits include the following:
    - The technology-mapping diagram was updated to show VMware vCloud Automation Center™.
    - vCloud API changed to VMware APIs.
    - VMware vCenter Operations Management Suite™ components are enumerated.
- *Chapter 3, Architecting a VMware vCloud*:
  - Information was added about vCloud Automation Center (a component of the vCloud Suite).
  - Section 3.8, "Multisite Considerations," was updated.
  - Allocation models guidance was updated.
  - Information about VMware metering was updated in Section 3.6, "vCloud Metering."
  - Hybrid vCloud considerations were updated in Section 3.9, "Hybrid vCloud Considerations."
- *Chapter 4, Operating a VMware vCloud*: Information was added about organizational structure and its evolution for vCloud in Section 4.5, "Organizing for vCloud Operations."
- *Chapter 5, Consuming a VMware vCloud*:
  - Updates were made to reflect the new network terminology in vCloud Director 5.1.
  - The text was updated to reflect new storage capabilities in vCloud Director 5.1.
  - Section 5.3.2, "vCloud Director Allocation Models," was updated to reflect changes in vCloud Director 5.1.

- Updates and clarifications were made to Section 5.4.3, "Working with Catalogs."
- Updates were made to Section 5.5.1.3, "vApp Migration," to reflect new capabilities in vCloud Director 5.1.
- Updates were made to Section 5.6.3 "What's New in the vCloud 5.1 API."

- *Chapter 6, Implementation Examples*: The following sections were extensively updated with the latest available information:
  - Section 6.3, "Organization Virtual Datacenter Examples"
  - Section 6.4.5, "VXLAN ORG Network for Disaster Recovery"
  - Section 6.7.3, "Implementing Signed Certificates from a Certificate Authority"
- *Chapter 7, Workflow Examples*: No content changes were made.
- *Chapter 8, Software Tools*: No content changes were made.
- *Chapter 9, Cloud Bursting*: No content changes were made.

Security information in Appendix B, "Security," was updated.

vCAT 3.0 Changes (Previous Release)

This section provides information on the changes that were made for the vCAT 3.0 release.

New documents were added to the toolkit, and in two cases, multiple documents were consolidated into one guide. Information about new components has been added, and information about other components has been updated.

New and Consolidated Documents

Workflow Examples, *Software Tools*, and *Cloud Bursting* are new documents with all new content.

The *vCAT 2.x Public VMware vCloud Service Definition* and *Private VMware vCloud Service Definition* have been consolidated into one *Service Definitions* document that covers public, private, and hybrid cases.

The vCAT 2.x documents, *Public VMware vCloud Implementation Examples*, *Private VMware vCloud Implementation Examples*, *and Hybrid Use Cases*, have been consolidated into one document titled *Implementation Examples* that covers public, private, and hybrid use cases. Many new implementation examples are provided.

New and Updated Components

vCAT 3.0 provided new and expanded coverage for architects, operators, and consumers.

- VMware vSphere®
- VMware vCloud Director®
- VMware vCenter™ Operations Management Suite™ (new):
  - VMware vCenter Chargeback Manager™
  - VMware vCenter Operations Manager™ (new)
  - VMware vCenter Infrastructure Navigator™ (new)
  - VMware vCenter Configuration Manager™ (new)
- VMware vCloud Networking and Security™ (formerly VMware vShield™):
  - VMware vCloud Networking and Security Edge™
  - VMware vCloud Networking and Security App™ (new)
  - VMware vCloud Networking and Security Data Security™ (new)
  - VMware vShield Endpoint™ (new)
- VMware vCloud Connector™
- VMware vCenter Orchestrator™
- VMware vSphere Service Manager™—Cloud Provisioning (new)
- VMware vCenter Site Recovery Manager™ (new)
- VMware vFabric™ RabbitMQ™ (new)
- VMware vFabric Application Director™ (new)
- VMware vFabric Application Performance Manager™ (new)
  - VMware vFabric Hyperic® (new)
  - VMware vFabric AppInsight™ (new)

VMware vCloud Networking and Security

VMware vShield has been renamed to VMware vCloud Networking and Security™. Note the following changes:

- VMware vShield Edge™ is now VMware vCloud Networking and Security Edge.
- VMware vShield App™ is now VMware vCloud Networking and Security App.
- VMware vShield Manager™ is now VMware vCloud Networking and Security Manager™.

The vCAT documents usually refer to vCloud Networking and Security, but some links to reference documents might still link to vShield documents on vmware.com. The vShield documents are being updated to reflect the new name.

Known Issues

Firefox generally works with the vCAT Documentation Center, but Google Translate does not work properly.

Providing Feedback

The usefulness of this architecture toolkit depends on feedback from customers and our network of partners. Send all feedback and IP submissions to ipfeedback@vmware.com.

From the documentation center interface, you can click the link next to the logo or click the Email (envelope) icon to send feedback.

Reader Services

Visit our website at www.informit.com/title/9780321912022 and register this book for convenient access to any updates, downloads, or errata that might be available for this book.

CHAPTER 1
Introduction

IN THIS CHAPTER

1.1 Overview

A *reference architecture* is an architecture template solution that addresses one use case in a particular domain. The VMware® vCloud® Architecture Toolkit (vCAT) provides modular components and documents to support multiple use cases, including design considerations and design patterns to support architects, operators, and consumers of cloud computing solutions based on VMware technologies.

vCAT is the first of several VMware Validated Architectures (VVA) VMware has released for customers, partners, vendors, and our internal teams. As a VVA, vCAT is supported by VMware and our support organization.

vCAT is vendor agnostic, but it does share vendor details when providing implementation examples. Vendors provide information about the use of their products with vCloud, including integration with vCAT, on the VMware Solutions Exchange (https://solutionexchange.vmware.com/store).

vCAT design guidelines cover multiple use cases. Instead of referring to *best practices* (a term subject to misinterpretation because best practices depend on use cases and are subject to many variables, including change over time), vCAT provides *design guidelines*. Architects must determine which design guidelines apply to the requirements, constraints, and characteristics of their projects and chosen technologies. When using the toolkit, consider the use case that best applies to your situation, and choose the design guidelines that support your design implementation.

This document covers the following topics:

- The vCAT documentation set
- Cloud computing and the VMware vCloud
- The journey to a mature vCloud implementation

For additional vCAT supporting material, visit the vCloud Architecture Toolkit page on vmware.com (www.vmware.com/go/vCAT). This is also where updates to vCAT are posted.

1.2 Using the vCAT Documentation Set

The vCloud Architecture Toolkit provides a set of documents to support the design of complex, integrated reference architectures for architects, operators, and consumers. Figure 1.1 shows the documents, and Table 1.1 briefly describes them.

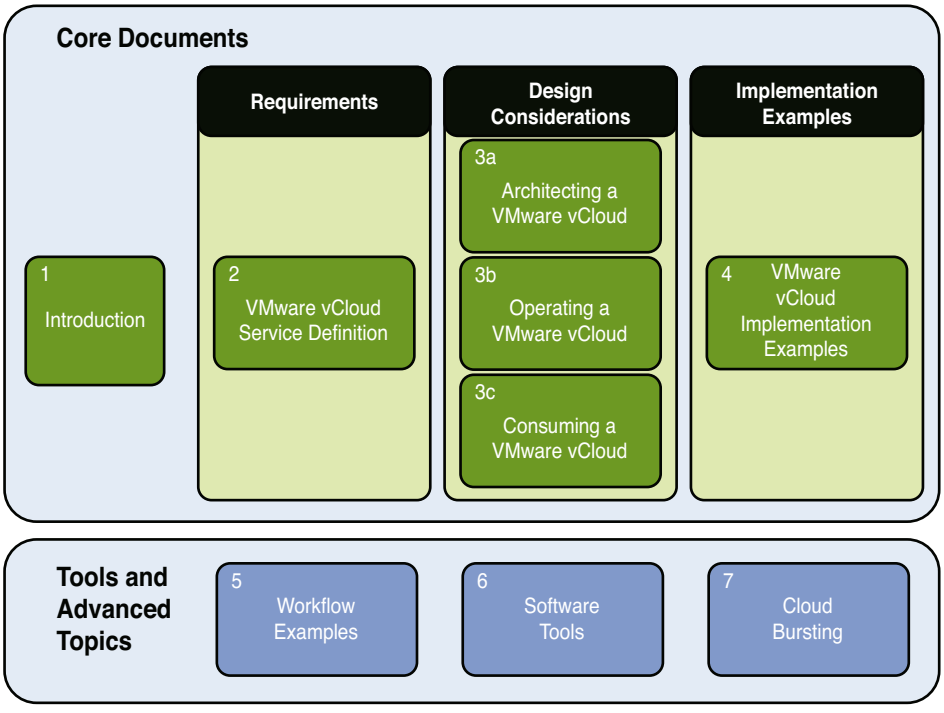


FIGURE 1.1 VMware vCloud Architecture Toolkit document map

Table 1.1 shows check marks in the first column to represent existence of an online video providing a brief presentation (<10 minutes) about a document and topic area.

TABLE 1.1 VMware vCloud Architecture Toolkit Documents

Video	Document	Description	Audience
	Release Notes	Includes information about the VMware Architecture Toolkit, toolkit packages, how to use the documentation center, and information about changes since the vCAT 3.0 release.	All
✓	*Introduction*	Covers the following topics: ▶ A brief summary of vCAT documents ▶ Suggested reading order, depending on audience or role ▶ Introduction to cloud computing and basic cloud computing requirements and definitions	All
✓	*Service Definitions*	Discusses service definition lifecycles, including specific considerations for private, public, and hybrid vCloud instances, and examples of service offerings designed to help you create service definitions that meet specific business objectives.	All
✓	*Architecting a VMware vCloud*	Details design considerations for architecting and building a VMware vCloud, including the basis for a reference architecture and guidance on requirements for implementing a VMware vCloud infrastructure.	Architects, IT operations
✓	*Operating a VMware vCloud*	Introduces high-level operational areas and discusses the evolution to support vCloud dynamics. Provides information about operational procedures, roles and responsibilities, setup, management, and monitoring of a vCloud. Also covers VMware management tools that support vCloud operations.	Architects, IT operations
✓	*Consuming a VMware vCloud*	Answers consumer questions such as the following: ▶ How do I handle the application lifecycle in a vCloud? ▶ How do I protect my workloads? ▶ How do I guarantee that workload resource requirements are met? Provides the consumer's perspective.	Architects, IT operations, consumers, end users
✓	*Implementation Examples*	Provides examples of how to build a vCloud.	Architects

Video	Document	Description	Audience
✓	*Workflow Examples*	Provides a description of useful scripts and workflows for VMware vCenter™ Orchestrator™. Other examples use technologies such as PowerCLI. Includes references to where these scripts can be found.	Architects, IT operations
✓	*Software Tools*	Includes information about software that can benefit architects and operators. Provides information about freely available technologies that have been created and used to assist in vCloud design, deployment, and operations. Also includes information about several powerful tools that are available only as part of a service engagement with VMware Professional Services or a VMware partner.	Architects, IT operations
✓	*Cloud Bursting*	Provides the theory behind autoscaling an enterprise cloud environment by using multiple cloud locations, including those owned by an enterprise and/or a service provider. This theory leverages VMware technologies but applies to other cloud technologies as well. This material is based on VMware field experience with customers and service providers.	Architects

Table 1.2 lists the typographical conventions used in all vCAT documents.

TABLE 1.2 Document Typographical Conventions

Emphasis	Emphasizes information, introduces new terms, and identifies document and workflow names.
`Command`	Identifies system commands, filenames, and Registry keys.
`Code`	Indicates code snippets and scripts.
User Interface	Identifies UI objects such as tabs, buttons, and field labels with bold text.
Hyperlink	Uses blue, underlined text to indicate an active link (URL).
Note, Caution	Notes contain information related to the topic that is of possible interest to the reader. Cautions highlight important information on potential problems or actions that might cause unexpected results. A Caution alerts the user and indicates the possibility of significant data loss.

1.2.1 Recommended Reading Order

The documents can be read in the order shown in the document map or in the order recommended for a particular audience or role, such as one of the following:

- *vCloud providers* who offer the vCloud infrastructure and services. An *architect* has overall control over how a solution is designed and implemented in the environment.
- *vCloud operators* who are responsible for operation of the cloud. *Operators* are involved with the day-to-day running and administration of the vCloud environment. They need to understand operational procedures and how the vCloud components fit together.
- *vCloud consumers* who use cloud provider resources for application deployment. A consumer (organization or individual) is someone who consumes vCloud resources. Consumers want to run their workloads in the vCloud environment without concern for the underlying infrastructure or day-to-day administration.

Table 1.3 identifies the recommended documents for each role.

TABLE 1.3 vCAT Audience Reading Guidelines

VMware vCloud Architecture Toolkit (vCAT) Reading Recommendations	Architect	Admin/ Operator	Consumer
Introduction	✓	✓	✓
Service Definitions	✓	✓	✓
Architecting a VMware vCloud	✓	✓	
Consuming a VMware vCloud	✓	✓	✓
Implementation Examples	✓		
Workflow Examples	✓	✓	
Software Tools	✓	✓	
Cloud Bursting	✓		

1.3 Cloud Computing and VMware vCloud

Cloud computing leverages the efficient pooling of an on-demand, self-managed, virtual infrastructure that is consumed as a service. VMware vCloud is the VMware solution for cloud computing that enables delivery of *Infrastructure as a Service* (IaaS). Additional "as a Service" reference architectures can be layered on top of a VMware vCloud built using vCAT.

1.3.1 VMware vCloud Requirements

According to the National Institute of Standards and Technology (NIST), the key components of a cloud are on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service. VMware aligns with the definition of *cloud* as elastic, lightweight entry and exit, available over Internet protocols, and running on a shared infrastructure.

A cloud always starts with a shared, virtual infrastructure. If any resource is dedicated to only one customer, you have a *managed hosting platform*, not a cloud infrastructure. Similarly, it is not considered a cloud if the cloud administrator or service provider must perform manual procedures to provision cloud resources following a consumer request. This is why workflow automation and orchestration are included as part of a vCloud solution.

The VMware vCloud blueprint follows these basic NIST requirements as the foundation for an IaaS cloud:

- A cloud must be built on a *pooled, virtual infrastructure*. Pools include not only CPU and memory resources, but also storage, networking, and associated services.
- The cloud should provide *application mobility between clouds,* allowing the consumer to enter and leave the cloud easily with existing workloads. The ability to use existing consumer tools to migrate workloads to or from the cloud is highly desirable. Mobility of workloads between clouds requires cross-cloud resource management.
- The cloud should be *open and interoperable*, allowing the consumption of cloud resources over open, Internet-standard protocols. Access to cloud resources does not require any other specific network protocols or clients.
- Cloud consumers should pay only for resources they consume or commit to consuming.
- The cloud should be a secure, trusted location for running cloud consumer workloads.
- Cloud consumers should have the option and capability to protect their cloud-based workloads from data loss.
- Cloud consumers are not responsible for maintaining any part of the shared infrastructure and do not need to interact with the cloud provider to maintain the infrastructure. They are not responsible for storage and network maintenance, ongoing cloud infrastructure patches, or business continuity activities. The cloud should be available to run high-availability workloads, and any faults occurring in the cloud infrastructure should be transparent to cloud consumers as a result of *built-in availability, scalability, security, and performance guarantees*.

1.3.2 VMware Alignment to Standards

VMware continues to develop technologies that align with evolving cloud standards as defined by NIST and other global standards organizations.

vCloud solutions focus on the following areas:

- **People:** People who develop solutions, architect the design, operate the implementation, and consume the resources. (See *Operating a VMware vCloud* and *Consuming a VMware vCloud*.)
- **Process:** Processes for architects, operators, and consumers.
- **Technology:** Alignment with successful design, deployment, and integration considerations. VMware technologies address the relevant areas within the standards.

Standards are still evolving for private, public, community, hybrid, and other types of clouds. vCAT focuses on the most common core design areas. The technology is the same, but operations and vCloud resource consumption vary according to the type of vCloud, the type of vCloud provider, and specific consumer requirements.

- *A private vCloud* is operated by an organization and secured behind a firewall.
- *A public vCloud* is generally accessible to users on the Internet.
- *A community vCloud* is a specific public vCloud use case in which access is limited to specified groups that share a common set of requirements.
- *A hybrid vCloud* is characterized by a connection among multiple vCloud instances. Typically, a bridge between two private vCloud instances has a dedicated and secured connection. The underlying network resides behind an Internet-facing firewall.

As cloud computing continues to evolve, many cloud definitions will arise. The information in this toolkit is a valuable aid in support of your vCloud projects, regardless of your chosen definition.

1.3.3 vCloud Definitions

vCAT uses the terms *private vCloud*, *public vCloud*, and *hybrid vCloud,* based on a specific set of definitions that NIST provides.

- Private cloud:

  A private vCloud (also known as an *internal vCloud*) operates on private networks, where a single company maintains accessible resources behind the firewall. In many cases, all the tenants share one legal entity. For example, a university might offer IaaS to its medical and business schools, or a company might do the same for various groups or business units. The private vCloud can be managed by the enterprise and hosted on-premises or operated on a dedicated infrastructure provided by a vCloud service provider or systems integrator. In any case, a private vCloud must conform to the organizational security constraints.

- Public cloud:

  A public vCloud offers IT resources as a service through external service providers and is shared across multiple organizations or the Internet. This can be viewed as a vCloud infrastructure that one organization operates and that multiple, legally separated organizations use.

  A public vCloud is provisioned for open access and might be owned, managed, and operated by one or more entities.

  A public vCloud provider might also support a private, community, or hybrid vCloud.

- Hybrid cloud:

  A hybrid vCloud combines the benefits of the private and public vCloud, with flexibility and choice of deployment methods.

  A hybrid vCloud consists of multiple, linked vCloud infrastructures. These distinct vCloud infrastructures can be private, community, or public; but they must meet a set of requirements that the providers define and the consumers agree to. Connecting these vCloud instances requires data and application mobility, as well as management.

  When load-balancing between vCloud instances (*cloud bursting*), use a consistent monitoring and management approach when migrating an application or data workload. For the theory behind cloud bursting, see the *Cloud Bursting* document.

- Community cloud:

  A community vCloud is a specific public vCloud use case in which the cloud is shared, and typically owned, by a group of organizations with a common set of requirements. In many cases, the organizations also include some level of legal separation. Community vCloud resources are shared, with some parts under central control and other parts with defined autonomy. A vCloud built for government, education, or healthcare is an example of a community vCloud.

  A community vCloud can be offered by a traditional service provider, by a member of the community, or by a third-party vendor and hosted on one or more sites. It can be placed on-premises at one or more of the organizations' sites, off-premises at a vCloud provider site, or both on- and off-premises.

1.3.4 Solution Area to Technology Mapping

When considering various technology solutions for your vCloud architecture, evaluate the solution and operational requirements to provide justification for the proposed solution. As VMware continues to develop Software Defined Data Center (SDDC) technologies, we will update the matching Infrastructure as a Service component. Figure 1.2 shows the categories of design considerations for building both a cloud and the underlying SDDC, with the related product technology that is used.

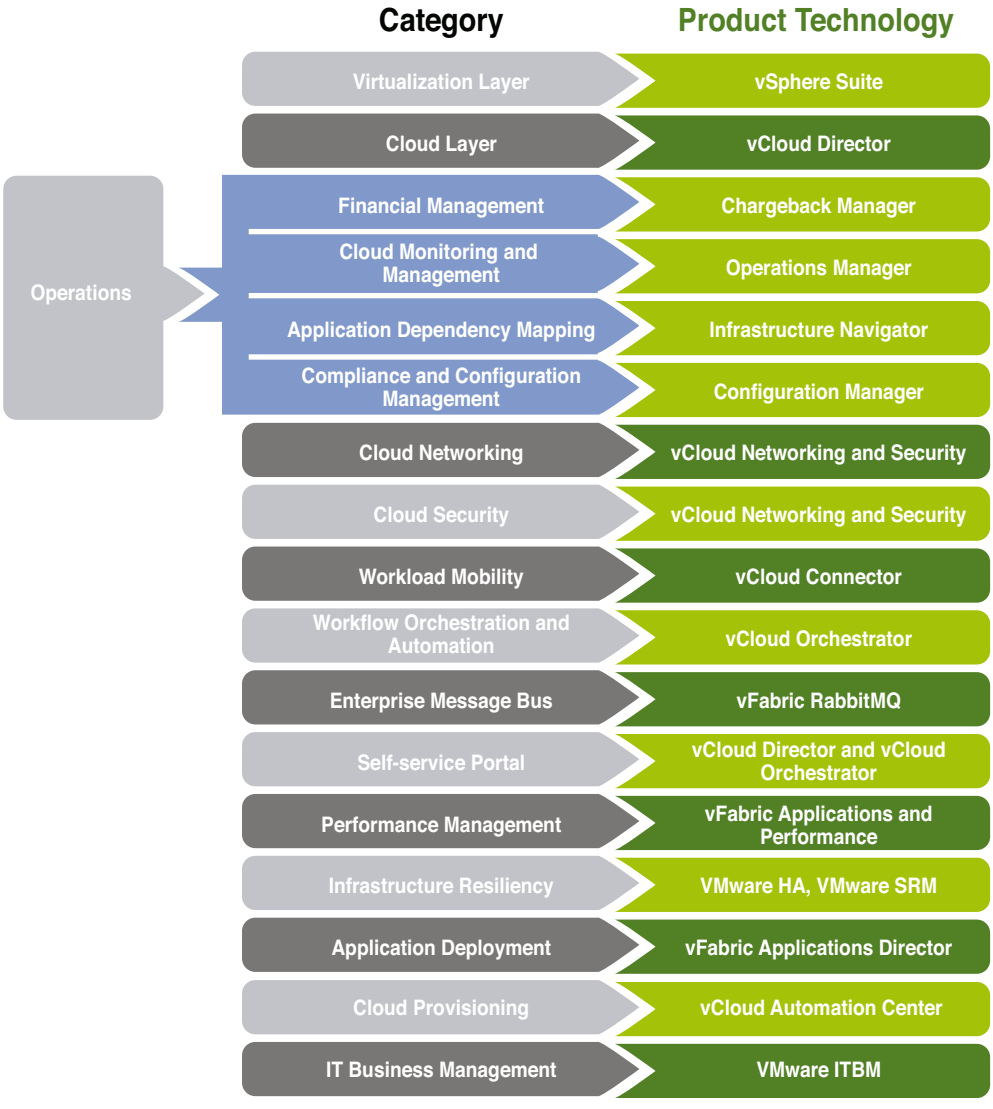


FIGURE 1.2 Technology areas

Figure 1.3 shows the technologies this vCAT release covers.

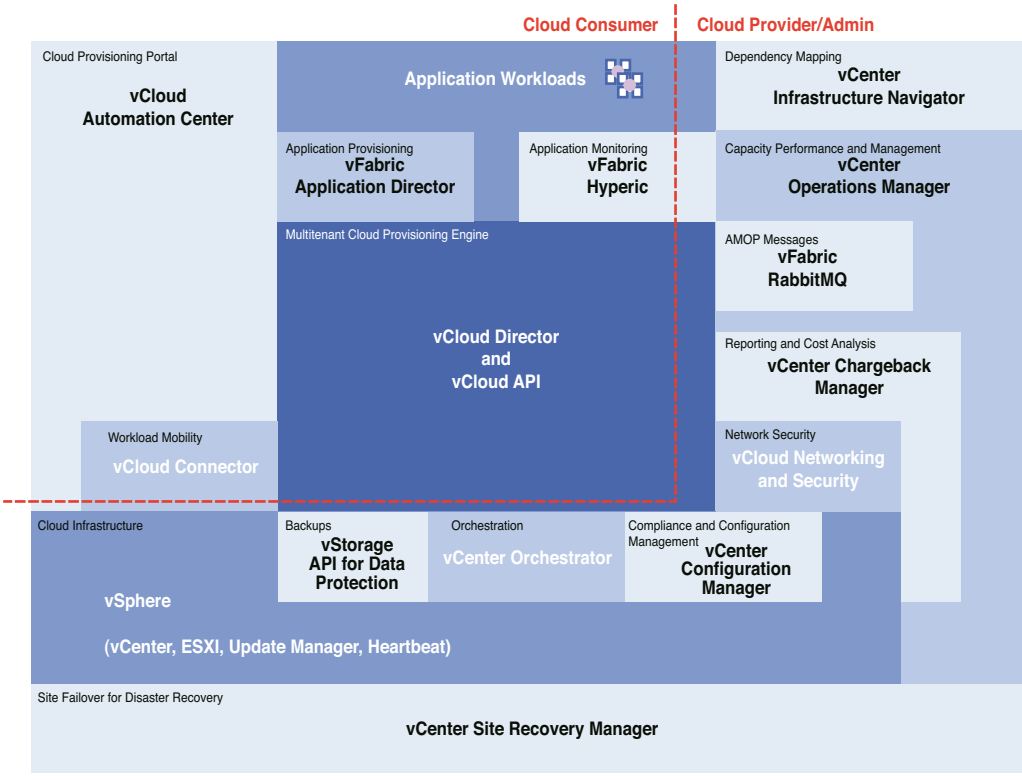


FIGURE 1.3 Technology areas in vCAT

NOTE

Except for the gray components, components that touch each other are integrated.

1.3.4.1 VMware Professional Services

VMware offers professional services that align with vCloud use cases. These range from a proof of concept (POC) that might be used as a demonstration environment, to a production deployment that requires management, workflow automation, compliance enforcement, and validation. The following services are available:

- **VMware vCloud POC Jumpstart Service:** Provides knowledge transfer workshops and hands-on product installation, configuration, and use demonstrations for the vCloud solution.

- **VMware vCloud Accelerator Service:** Rapidly delivers a functioning VMware vCloud implementation suitable for deploying applications in a limited-scale preproduction environment. If all prerequisites are met, this service engagement can be completed in fewer than 30 business days.
- **VMware vCloud Design and Deploy Service:** Provides a comprehensive architectural design for VMware vCloud that addresses the customer's unique business requirements and operational demands, helping to pave the way to vCloud computing. This service is designed for enterprises that have a well-established, vSphere-based virtualization strategy for production workloads and that are ready to take the next step toward building their production vCloud infrastructure.
- **VMware Operational Readiness for Cloud Computing Service:** Offers a four- to six-week engagement in which VMware consultants examine existing operational practices to evaluate performance across more than 150 attributes in five key areas. They uncover unknown or hidden barriers to success and highlight areas in which additional focus on people or process can deliver increased productivity, streamline operations, and improve overall vCloud solution results.

Services can be combined or customized to meet your specific requirements.

1.4 Journey to a Mature vCloud Implementation

At every stage in the processes leading to a mature vCloud implementation, financial transparency, process maturity, organizational setup, and technology implementation are critical factors for success.

VMware defines three stages on the journey to a mature vCloud: Standardize, Service Broker, and Strategic Partner. Figure 1.4 depicts these, and the following sections describe them.

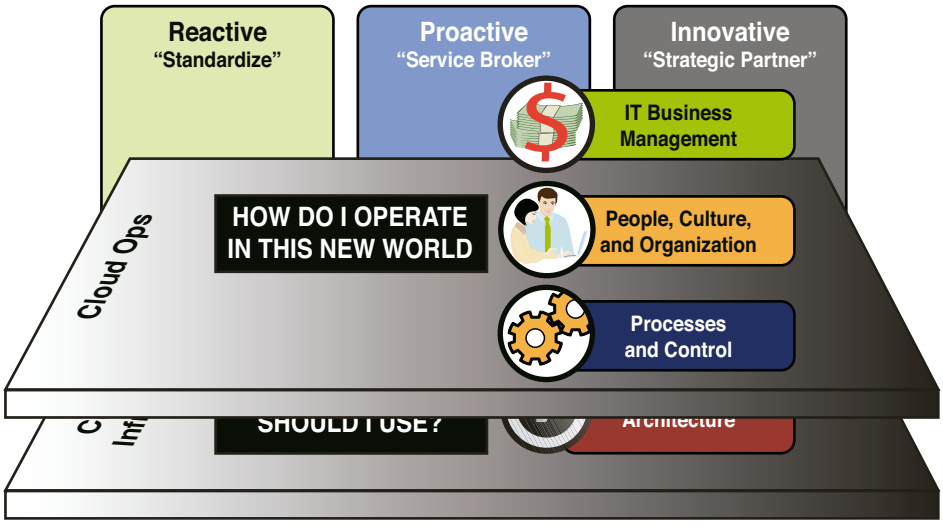


FIGURE 1.4 Journey states in vCAT

1.4.1 Stage 1: Standardize

Stage 1 often coincides with a more mature server virtualization environment, and the focus is on creating a working vCloud solution with an on-demand service catalog end users can directly access. The service catalog provides rapid deployment services for non-business-critical, development, and test environments, as well as for externally sourced applications. Implementing the service catalog promotes cloud acceptance by business users and also outlines a long-term vCloud implementation strategy with planning for operational and organizational change. The following capabilities are important for this stage:

- **Financial model and measurement:** Awareness and understanding of the costs of assets and underlying infrastructure capacity.
- **People and organization:** Specialized but shared roles for managing virtualized environments.
- No explicit virtualization Center of Excellence established.

NOTE

See "Organizing for vCloud Operations" in *Operating a VMware vCloud* for information about the Center of Excellence.

- **Process and control:**
  - IT processes are adapted for virtualization but are largely manual, with specific, customized interprocess integration.
  - The focus is on limited, continuous improvement.

- **Tools and technology:**
  - Online, self-service capability for development and test provisioning
  - Online, self-service capability for Software as a Service (SaaS)–based applications
  - Operational tools defined for virtualization environments
  - Some business workloads run in a virtualized environment, whether internal or provided by third parties

1.4.2 Stage 2: Service Broker

Stage 2 is the first service-driven stage for a vCloud. At this stage, IT has transformed from traditional models and is focused on delivering business services within a vCloud environment. This represents a cultural shift within the organization. To be successful, it requires enhanced IT operational maturity, an optimized IT organizational structure, and supporting cloud-management tools.

The term *service broker* implies that IT is organized at this stage to source internally and externally, such as adding external infrastructure capacity or providing access to vendor-based SaaS applications. The business is not necessarily aware of how the services are made available, but dramatically decreased development and provisioning times support business needs with increased quality of service and greater agility.

This stage focuses on the following goals:

- Gaining alignment and buy-in from key business stakeholders
- Creating service governance, lifecycle and service design, and development processes
- Providing service-based financial transparency
- Automating and integrating tools and technology in internal and external systems

Key capabilities for this stage include the following:

- Financial model and measurement:
  - Using usage metering and cost showback
  - Applying granular costing of underlying infrastructure assets
  - Educating IT customers about paying for services as an operating expense
  - Changing from project-based budgeting to demand-based budgeting
- People and organization: Establishing the Center of Excellence with dedicated, experienced, and knowledgeable staff
- Process and control:
  - Fully integrated IT operational processes adapted for virtualization and vCloud
  - Agile-based service design and development processes established
  - Service-level financial transparency

- Tools and technology:
  - Services defined and offered through an online consumer portal for self-service access to the service catalog
  - vCloud-level disaster recovery
  - Blueprint and policy-driven service development and provisioning
  - Purpose-built management tools for proactive vCloud operations

1.4.3 Stage 3: Strategic Differentiator

This stage is the final stage for a mature cloud. At this point, a highly efficient, scalable cloud with hybrid capability is available for an organization. IT is delivered as a service. Automated, policy-driven governance and control applies across the vCloud environment, with zero-touch operations supported by predictive and self-healing operational tool capabilities. True application mobility and device-independent access is available. The vCloud is considered to be the de facto model within the organization. The term *strategic differentiator* implies that IT has changed roles and become a business differentiator by increasing agility, resulting in faster time to market; increasing efficiency, resulting in reduced costs; and increasing reliability, resulting in dramatically increased quality of service. The following are key capabilities for this stage:

- Financial model and measurement:
  - Usage-based pricing and chargeback for services provided to business customers
  - Service demand–based budgeting
  - Priced catalog of service offerings
- People and organization: The Center of Excellence manages all elements of infrastructure, end-user, and application operations.
- Process and control:
  - Optimized, integrated, and fully automated IT processes that enhance business agility and efficiency
  - Continuous process, service, and performance improvements based on predictive capabilities
- Tools and technology:
  - Full hybrid capabilities
  - Tools that support single-pane-of-glass management across private and public vCloud environments
  - Service-level disaster recovery
  - Tools that support automated corrective actions for self-healing

CHAPTER 2
Service Definitions

IN THIS CHAPTER

2.1 Introduction

Businesses face constant pressure to introduce products and services rapidly into new and existing marketplaces, while users expect services to be easily accessible on demand and to scale with business growth. Management demands these services at a fair price. These pressures and demands all require information technology (IT) to become more service oriented. They also make it more important than ever for IT to improve its strategy to deliver services with the agility that businesses now expect. Cloud computing is central to a better IT strategy.

Virtualization has reduced costs and increased server efficiency, often dramatically, but it does not, by itself, deliver the level of automation and control required to achieve the efficiencies or agility associated with cloud computing. Cloud computing offers the opportunity to further improve cost efficiency, quality of service, and business agility. It enables IT to support a wide range of changing business objectives, from deployment of new tools, products, and services to expansion into new markets. Cloud computing transforms IT from a *cost center* into a *service provider*.

The VMware® vCloud® Suite is the VMware solution for cloud computing.

This book provides the information you need to create a service definition for an organization that provides Infrastructure as a Service (IaaS) resources for private, public, and hybrid vCloud instances. This book has the following goals:

- Acquaint you with what to consider when creating a service definition
- Provide examples that act as a starting point to create a service definition for service offerings that meet specific business objectives

2.1.1 Audience

This document is intended for those involved in planning, defining, designing, and providing VMware vCloud services to consumers. The intended audience includes the following roles:

- Providers and consumers of vCloud services
- Architects and planners responsible for driving architecture-level decisions
- Technical decision makers who have business requirements that need IT support
- Consultants, partners, and IT personnel who need to know how to create a service definition for their vCloud services

2.1.2 Deployment Model

Figure 2.1 illustrates several deployment models for cloud computing.

- For enterprises, the focus is on private and hybrid vCloud environments.
- For service providers, the focus is on public and hybrid vCloud environments.

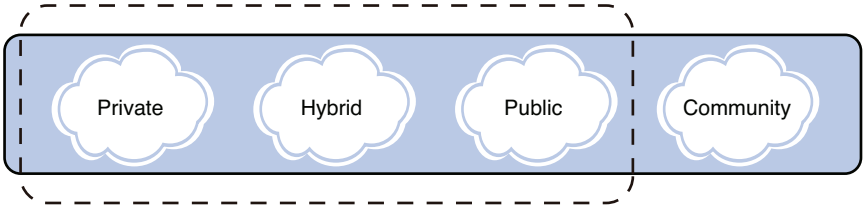


FIGURE 2.1 Deployment models

The following are the commonly accepted definitions for cloud computing deployment models:

- **Private vCloud:** The vCloud infrastructure is operated solely for an organization and can be managed by the organization or a third party. The infrastructure can be located on-premises or off-premises.
- **Public vCloud:** The vCloud infrastructure is made available to the general public or to a large industry group and is owned by an organization that sells vCloud services.
- **Hybrid vCloud:** The vCloud infrastructure is a composite of two or more vCloud instances (private and public) that remain unique entities but are bound together by standardized technology. This enables data and application portability, such as *cloud bursting* for load balancing between vCloud instances. With a hybrid vCloud, an organization gets the advantages of both, with the capability to burst into the public vCloud when needed while maintaining critical assets on-premises.

- **Community vCloud:** Several organizations share the vCloud infrastructure. This infrastructure supports a specific community that has shared concerns, such as mission, security requirements, policy, and compliance considerations. It can be managed by the organizations or a third party and can be located on-premises or off-premises.

This book covers the following private, public, and hybrid vCloud deployment models:

- **Private vCloud:** Enterprise IT as a provider of vCloud services to consumers
- **Hybrid vCloud:** Enterprise IT as a consumer of public vCloud services, extending its own private capacity
- **Public vCloud:** Service provider IT as a provider of vCloud services to a number of enterprise consumers

The book does not cover community vCloud service definition considerations and examples.

2.1.3 Service Model

The National Institute of Standards and Technology (NIST) specifies three service layers in a cloud, as Figure 2.2 shows. VMware defines these service layers as follows:

- **Software as a Service (SaaS):** Business-focused services are presented directly to the consumer from a service catalog.
- **Platform as a Service (PaaS):** Technology-focused services are presented for application development and deployment to application developers from a service catalog.
- **Infrastructure as a Service (IaaS):** Infrastructure containers are presented to consumers to provide agility, automation, and delivery of components.

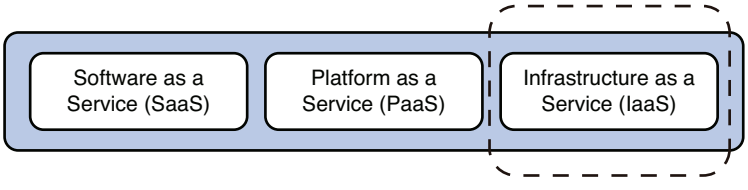


FIGURE 2.2 Service models

The service model for the service definition in this document is primarily IaaS, for an organization to provide Infrastructure as a Service to consumers of vCloud services through a catalog of predefined infrastructure containers. The IaaS service layer serves as a foundation for additional service offerings, such as PaaS, SaaS, and Desktop as a Service (DaaS).

2.1.4 Technology Mapping

vCloud services are delivered by the capabilities of the VMware technologies in the VMware vCloud Suite, as Figure 2.3 shows. The VMware vCenter™ Operations Management Suite™ includes the following technologies:

- VMware vCenter Operations Manager™
- VMware vCenter Chargeback Manager™
- VMware vCenter Configuration Manager™
- VMware vCenter Infrastructure Navigator™

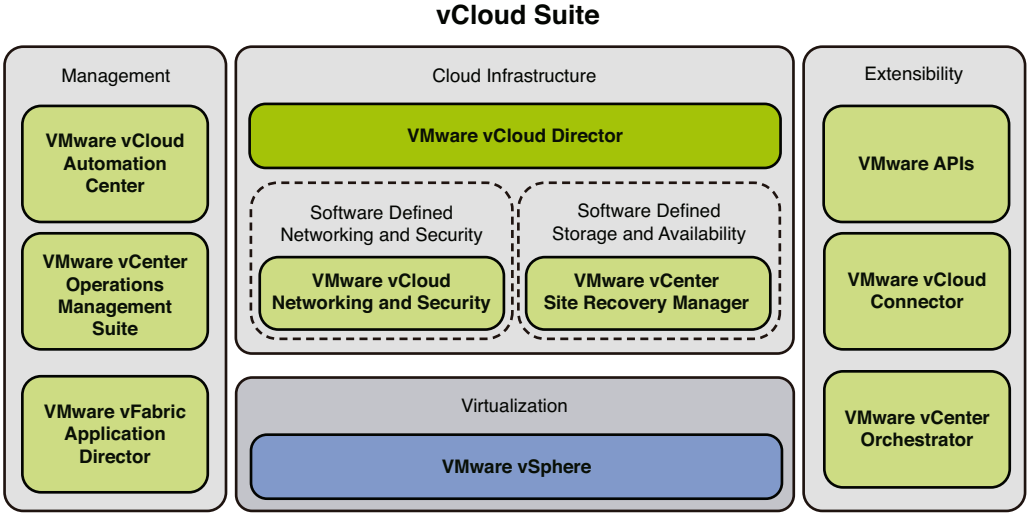


FIGURE 2.3 Technology mapping

2.1.5 Service Characteristics

The National Institute of Standards and Technology defines the following essential cloud service characteristics:

- **Broad network access:** Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin client or thick client platforms.
- **Rapid elasticity:** Capabilities can be provisioned to scale out quickly and to be released rapidly—in some cases, automatically. Rapid elasticity enables resources to both scale out and scale in quickly. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be purchased in any quantity at any time.

- **Measured service:** Cloud systems automatically control and optimize resource usage by leveraging a metering capability at some level of abstraction appropriate to the type of service. Resource usage can be monitored, controlled, and reported, providing transparency for both the provider and the consumer of the utilized service.
- **On-demand self-service:** A consumer can unilaterally automatically provision computing capabilities as needed without requiring human interaction with each service's provider.
- **Resource pooling:** The provider's computing resources are pooled to serve multiple consumers, using a multitenant model with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. A sense of location independence results because the subscriber generally has no knowledge of or control over the exact location of the provided resources, but the subscriber might be able to specify location at a higher level of abstraction.

Figure 2.4 illustrates the relationships among service characteristics and how they all relate to resource pooling.

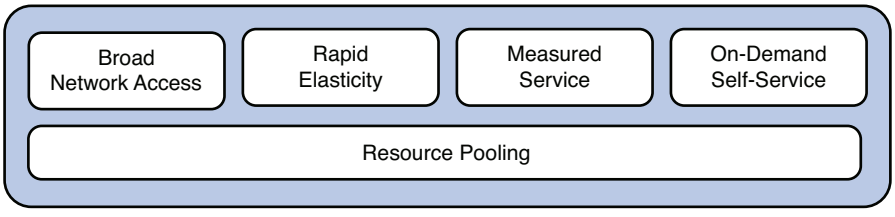


FIGURE 2.4 Service characteristics

To deliver business solutions using vCloud services, the vCloud infrastructure must have the following additional essential characteristics:

- **Standardized:** Homogeneous infrastructure delivered as software services across pools of standard x86 hardware. Homogeneity eliminates unnecessary complexity caused by operating system silos and the redundant tools and skill sets associated with them. It also eliminates costly, special-purpose hardware and enables a single, scalable approach to backup and recovery.
- **Holistic:** A platform optimized for the entire datacenter fabric, providing comprehensive infrastructure services capable of supporting any and all applications. A holistic infrastructure can support any workloads, with complete flexibility to balance the collective application demands, eliminating the need for diverse technology stacks.
- **Adaptive:** Infrastructure services are provided on demand, unconstrained by physical topology and dynamically adapting to application scale and location. The infrastructure platform configures and reconfigures the environment dynamically, based on collective application workload demands, enabling maximum throughput, agility, and efficiency.

- **Automated:** Built-in intelligence automates provisioning, placement, configuration, and control, based on defined policies. Intelligent infrastructure eliminates complex, brittle management scripts. Less manual intervention equates to scalability, speed, and cost savings. Intelligence in the infrastructure supports vCloud-scale operations.
- **Resilient:** A software-based architecture and approach compensates for failing hardware, providing failover, redundancy, and fault tolerance to critical operations. Intelligent automation provides resiliency without the need for manual intervention.

2.1.6 Service Development Approach

The approach for defining and designing vCloud services should have the following characteristics:

- Involves all necessary stakeholders.
- Documents business drivers and requirements that can be translated into appropriate service definitions.
- Takes a holistic view of the entire service environment and service lifecycle, including:
  - Service setup, which includes definition and design
  - Service request and approval
  - Service provisioning
  - Service consumption
  - Service management and operations
  - Service transition and termination

  A conscious awareness of what consumers of the service and the provider of the service experience at each stage of the service lifecycle must be taken into account, to create the necessary service definition elements for the consumer-facing service-level agreement (SLA) and internal-facing operational-level agreement (OLA) criteria.
- Defines the service scenarios and use cases.
- Represents the service to understand its components, interactions, and sequences of interrelated actions.
- Defines the users and roles involved with or interacting with the services so that the services created are user-centric.
- Defines the service contract (SLA) for the services and service components in the following areas:
  - Infrastructure services
  - Application/vApp services

- Platform services
- Software services
- Business services

- Defines service quality for these areas:
  - Performance
  - Availability
  - Continuity
  - Scalability
  - Manageability
  - Security
  - Compliance
  - Cost and pricing
- Defines the business service catalog and supporting IT service catalog

2.1.7 Concepts and Terminology

The key terms and service concepts are defined as follows:

- **Service:** A means of delivering value to consumers by facilitating outcomes that they want to achieve, without the ownership of specific costs or risks.
- **vCloud:** A model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable resources that can be provisioned rapidly and released with minimal management effort.
- **vCloud service provider (or provider):** An entity that provides vCloud services to consumers.
- **Consumer or customer:** Someone who consumes vCloud services and defines or agrees to service-level targets.
- **Service-level target:** A commitment that is documented in a service-level agreement. Service-level targets are based on service-level requirements and are needed so that the vCloud service design is fit for its purpose. Service-level targets should be *SMART* (Specific, Measurable, Actionable, Realistic, Time-bound) and are usually based on key performance indicators (KPIs).
- **Service-level agreement (SLA):** An agreement between a service consumer and the service provider that measures the quality and performance of the available services. The SLA is the entire agreement that specifies what service is to be provided, how it is supported, time, locations, cost, performance, and responsibilities of the parties involved.

- **Service-level objective (SLO):** A negotiated document that defines the service to be delivered to the consumer, with one or more key performance indicators (KPIs). It provides a clear understanding of the nature of the service being offered, focusing on the contribution of the service to the business value chain. SLOs are specific, measurable characteristics of the SLA, such as availability, throughput, frequency, response time, or quality.
- **Operational-level agreement (OLA):** An agreement internal to the service provider that details the interdependent relationships among the internal support groups of an organization working to support an SLA.
- **VMware vCloud Suite:** The suite of VMware technologies that provide the solution for vCloud computing.
- **VMware vCloud Services or vCloud Services:** vCloud computing services built with the VMware vCloud Suite.

2.2 Service Definition Considerations

Service definition is an important aspect of service design and management. It enables both the consumer and the service provider to know what to expect (or not to expect) from a service. Clearly defined services help customers understand the scope, limitations, and cost of service offerings.

Take the following considerations into account when developing a service definition. They are common to both private and public service definitions unless otherwise noted:

- Service objectives
- Use cases
- User roles that interact with the service
- Consumption model
- Service metering, reporting, and pricing
- Service offering details (infrastructure, applications)
- Other features that vary by offering type (backup, type of storage, availability, performance, continuity)

2.2.1 Service Objectives

Understanding the service objectives is an essential first step. Service objectives must address the specific business challenges. The following are examples of service objectives for vCloud services:

- Deliver a fully operational private or public vCloud infrastructure with hybrid capability
- Provide secure multitenancy for vCloud infrastructure consumers

- Provide compliance controls and transparency for the service
- Maintain IT control of access to the system and resources
- Provide differentiated tiers of scale to align with business needs
- Allow for metering of the service for cost distribution
- Establish a catalog of common infrastructure and application building blocks
- Provide the following service offerings:
  - Basic (pay for resources used)
  - Committed (allocated resources)
  - Dedicated (reserved resources)
- Support a minimum of 1,500 virtual machines across the three service offerings, and have a plan to grow to a minimum of 5,000 virtual machines
- Provide workload mobility between vCloud instances, allowing the consumer to enter and leave the vCloud easily with existing workloads
- Provide a direct connection to the external network for applications with upstream dependencies
- Provide an isolated network for applications that need to be isolated
- Provide open, interoperable, and Internet-standard protocols for consuming vCloud resources
- Provide for workload redundancy and data protection options

2.2.2 Use Cases

The use cases in Tables 2.1 through 2.5 represent business problems (some general and some industry-specific) that can be addressed with vCloud services and represented by a service definition.

TABLE 2.1 Example: Use Case 1

Use Case UC_01	
Name	**Modernization**
Problem Statement	Existing business services, processes, and legacy applications do not allow business to stay competitive.
Description	Modernization of business services, processes, and legacy applications
Requirements/Goal	Modernize infrastructure to make it service oriented ▶ Modernize applications ▶ Modernize business processes to improve speed to market
Risks	▶ Lost competitiveness and opportunities to support introduction of products and services in new or existing markets ▶ Increasing investment in maintaining legacy applications

TABLE 2.2 Example: Use Case 2

Use Case UC_02	
Name	**Increased business capacity and scale rapidly**
Problem Statement	The business is unable to scale up its operation because IT cannot scale up capacity rapidly to support the business.
Description	IT needs to be able to scale proactively to support seasonal and periodic business demand.
Requirements/Goal	▶ Give consumers access to scale capacity on demand ▶ Enable IT to scale up, down, in, or out to support business demand ▶ Scale within a short cycle of days to meet projected demand ▶ Scale to off-premises capacity
Risks	▶ Lost revenue due to lack of capacity ▶ Lost customers from underperforming business services

TABLE 2.3 Example: Use Case 3

Use Case UC_03	
Name	**Rapid provisioning of development and test services**
Problem Statement	The business cannot develop new products and services rapidly because IT takes too long to provision development and test infrastructure.
Description	IT needs to be able to provide on-demand self-service provisioning of development and test infrastructure to support the business in rapidly developing new products and services.
Requirements/Goal	▶ Give developers and test users access to a catalog of IT infrastructure that they can rapidly provision and use ▶ Self-service provisioning, with necessary approvals ▶ Reduce time to market for products and services
Risks	▶ Products and services that are late to market, resulting in lost customers and market share

TABLE 2.4 Example: Use Case 4

Use Case UC_04	
Name	**Security and compliance assurance**
Problem Statement	The business is concerned about putting critical financial applications and data on vCloud services.
Description	IT must be able to provide secure business services for financial applications and data, which should have controlled access and be separate from other users of the vCloud services.

Use Case UC_04	
Name	**Security and compliance assurance**
Requirements/Goal	▶ Provide compliance controls and transparency for the service ▶ Provide network isolation for applications that must be isolated
Risks	▶ Security and compliance breach

TABLE 2.5 Example: Use Case 5

Use Case UC_05	
Name	**Business market launch**
Problem Statement	The business has insufficient resources and capacity to respond rapidly to marketplace needs, including seasonal events, although new opportunities have been identified.
Description	IT must be able to move at the speed of the business by rapidly providing the necessary infrastructure and services so that new applications, products, and services can be launched rapidly.
Requirements/Goal	▶ Provide rapid service provisioning to support product and service launches ▶ Give consumers access to a catalog of IT infrastructure that they can rapidly provision and use ▶ Self-service provisioning with necessary approvals ▶ Reduce time to market for products and services
Risks	▶ Products and services that are late to market, resulting in lost customers and market share ▶ Lost opportunity cost

2.2.3 User Roles

Several user roles apply to everyone who interacts with an enterprise vCloud service. Some roles are defined in the access model of the enterprise's private vCloud service at the provider level and at the consumer level. In addition, levels of privilege granted to predefined roles have an important impact on how users interact with the enterprise's vCloud service.

Table 2.6 provides a sample of the users and roles required for the enterprise vCloud solution.

TABLE 2.6 User Roles and Rights Example

User Role	Needs	Rights
Provider Cloud Administrator	One (minimum)	Highest-level enterprise vCloud provider administrator; has superuser privileges

User Role	Needs	Rights
Provider Catalog Author	As needed	Provider user who creates and publishes new catalogs
Consumer Organization Administrator	One per organization	Administrator over systems and users in the organization
Consumer Organization Author	One or more, as needed	User role that allows vApp and catalog creation, but no infrastructure management
Consumer Organization User	One or more, as needed	User role that allows a consumer organization user to use vApps created by others

2.2.4 Metering and Service Reporting

For vCloud environments, resource metering and service reporting are essential for calculating service costs. They also play an important role in accurately measuring consumer usage and shaping consumer behavior through chargeback policies. Enterprises might not necessarily have the same cost pressures for an enterprise private vCloud as for a public vCloud service provider. The requisite chargeback procedures or policies might not exist. An alternative to chargeback is *showback*, which tries to raise awareness of the consumption usage and cost without involving formal accounting procedures to bill the usage back to the consumer's department.

Table 2.7 provides examples of workload virtual machine sizing and costing.

TABLE 2.7 Workload Virtual Machine Sizing and Costing Examples

Virtual Machine Type	Sizing	Storage	Cost Model	
Extra large	8 vCPU, 8GB RAM (can offer up to 32 vCPU and 1TB RAM)	400GB	Provision cost ($)	Operate cost ($/mo)
Large	4 vCPU, 8GB RAM	200GB	Provision cost ($)	Operate cost ($/mo)
Medium	2 vCPU, 2GB RAM	60GB	Provision cost ($)	Operate cost ($/mo)
Small	1 vCPU, 1GB RAM	30GB	Provision cost ($)	Operate cost ($/mo)

2.2.5 Security and Compliance

Security and compliance continue to be concerns for enterprise subscribers seeking to adopt vCloud services. Most regulations and mandates in the industry, such as SOX, PCI DSS, and HIPAA/HITECH, have two general areas of requirements: transparency and control.

2.2.5.1 Compliance Definition

Transparency enables vCloud consumers to know who has accessed what data, when, and where. Payment Card Industry (PCI) requirement #10.3 is a good example of the need for transparency. It states that logs must contain sufficient detail for each event to be traced to a source by user, time, and origin.

Control gives vCloud consumers a necessary component of compliance by limiting access, based on a particular role and business need. Who can access, configure, and modify a vCloud environment; what firewall ports are open; when to apply patches; and where the data resides are common concerns from auditors. Cloud consumers—especially enterprise subscribers—believe that you can outsource responsibility, but you can't outsource accountability. As evidenced in the PCI Security Standards Council *Assessor Update: July 2011*, active Qualified Security Assessors (QSA) have the ultimate responsibility for their client's assessment and the evidence provided in the Report on Compliance. Both vCloud consumers and their auditors retain final accountability for their compliance and enforcement.

By design, vCloud services are intended to address common security and compliance concerns with transparency and control by doing the following:

- Facilitating compliance through ISO 27001 certification and/or SSAE 16, SOC 2 reporting, based on a standard set of controls
- Providing compliance logging and reports to service subscribers, for full visibility into their hosted vCloud environments
- Architecting the service so that subscribers can control access to their vCloud environments

2.2.5.2 Compliance Controls

For enterprise subscribers to feel secure and safe in the vCloud services domain, and to have the information and visibility into the service needed for their own internal audit requirements, providers of vCloud services must actively pursue one of the following certifications as part of their general service availability plans:

- ISO 27001 certification, which certifies that security management processes are in place and have a relevant subset of the ISO 27001 controls, as specified by the *VMware Compliance Architecture and Control Matrix*
- SSAE 16, SOC 2 report, based on the same relevant set of controls

VMware can provide documented guidance on how to meet the standard set of compliance controls, but providers are directly responsible for achieving ISO 27001 and/or SSAE 16, SOC 2 certification status for their service environments through a third-party audit. vCloud providers should make compliance certification types and status available so that subscribers understand what standards both the hosting environment and the services have been audited against.

2.2.5.3 Compliance Visibility and Transparency

Log management is often built into many of the compliance frameworks, such as ISO 27002, HIPAA/HITECH, PCI DSS, and COBIT. Enterprise subscribers not only need visibility into their private vCloud instances, but they also demand that providers give them visibility into their public vCloud environments. For example, enterprise subscribers must collect and archive logs and reports related to user activities and access controls such as firewalls.

To meet the requirements of being compliant with the controls, providers must enable reasonable visibility and transparency into their vCloud service architecture for subscribers. To accomplish this, service providers should collect and maintain logs for periods of 6 and 12 months for relevant components of the vCloud service and should be able to provide pertinent logs back to individual vCloud subscribers on an as-needed basis. Service providers should also maintain and archive logs for the underlying multitenant hosting infrastructure, based on the same 6- and 12-month periods. In the event of an audit, service providers should be able and willing to provide these logs to an auditor and/or individual subscriber. In general, vCloud service providers should have logs covering the following components of a subscriber's environment and should keep them readily available for subscriber access for periods of up to 6 and 12 months:

- VMware vCloud Director®
- VMware vCloud Networking and Security Edge™

The VMware vCloud Suite is based on a set of products that have been used in many secure environments. Products such as VMware vCloud Director and VMware vCloud Networking and Security™ generate a set of logs that give subscribers visibility into all user activities and firewall connections. VMware provides the necessary blueprints and best practices so that providers can best standardize and capture these sets of logs and provide subscribers with the capability to access them.

In addition to logs, service providers should provide basic compliance reports to their subscribers so that they understand all the activities and risks in their vCloud environment. VMware provides design guidelines in this area so that vCloud service providers can meet common enterprise subscriber requirements. Service providers are responsible for logging their vCloud services as well as their subscriber environments. These capabilities should be implemented and validated before any vCloud service is made generally available.

2.2.5.4 Compliant and Secure Architecture

All vCloud services offer a secure platform. VMware vSphere, a core building block, offers a secure virtualization platform with EAL4+ and FISMA certifications. vCloud Director, a vCloud delivery platform, offers secure multitenancy and organization isolation. The vCloud Suite enables enterprises to exercise the defense-in-depth security best practice. The platform offers both per-organization firewalls and per-vApp firewalls, and all organizations are isolated with their own Layer 2 networks. Access and authentication can optionally be performed against an enterprise organization's own directory using LDAP or Active Directory; the enterprise can thus self-manage its user base and provide role-based access according to its own policies.

2.2.6 Capacity Distribution and Allocation Models

To support the service offerings, determining the infrastructure's capacity and scalability is important. The following models determine how the resources are allocated:

- **Pay as you go:** No upfront resource allocation; resources are reserved on demand per workload.
- **Allocation pool:** A percentage of resources is reserved with overcommitment.
- **Reservation pool:** All resources are reservation guaranteed.

To determine the appropriate standard units of resource consumption, the vCloud service provider can analyze current environment usage, user demand, trends, and business requirements. Use this information to determine an appropriate capacity distribution that meets business requirements. If this information is not readily available, predicting the infrastructure capacity can be difficult because it depends on the expected customer uptake and usage of the workloads. However, understanding the infrastructure capacity required, based on an estimate of the different allocation models and capacity distribution of the workloads, is useful. The capacity distribution and resulting infrastructure resources allocated can be adjusted based on utilization and demand.

The following example distributes capacity based on 50% of the virtual machines for the reservation pool allocation model and 50% of the virtual machines for the pay as you go model. The reservation pool model is applied to small, medium, and large pools, with a respective split of 75%, 20%, and 5%. Therefore, *small* represents 37.5% of the total, *medium* represents 10% of the total, and *large* represents 2.5% of the total number of virtual machines in the environment.

Table 2.8 lists the virtual machine count for the various resource pools supporting the two example allocation models for the virtual datacenters.

TABLE 2.8 Definition of Resource Pool and Virtual Machine Split

Type of Resource Pool	Total Percentage	Total Virtual Machines
Pay as you go	50%	750
Small reservation pool	37.5%	563
Medium reservation pool	10%	150
Large reservation pool	2.5%	37
Total	100%	1,500

The following virtual machine distribution is used in the service capacity planning example:

- 45% small virtual machines (1GB, 1 vCPU, 30GB of storage)
- 35% medium virtual machines (2GB, 2 vCPU, 40GB of storage)
- 15% large virtual machines (4GB, 4 vCPU, 50GB of storage)
- 5% extra-large virtual machines (8+GB, 8+ vCPU, 60GB of storage)

Table 2.9 lists some examples of workload virtual machine sizing and utilization.

TABLE 2.9 Workload Virtual Machine Sizing and Utilization Examples

Virtual Machine Type	Sizing	CPU Utilization	Memory Utilization
Small	1 vCPU, 1GB RAM	10%–15% average	Low (10%–50%)
Medium	2 vCPU, 2GB RAM	20%–50% average	Moderate (50%–75%)
Large	4 vCPU, 4GB RAM	>50% average	High (more than 90%)
Extra large	8 vCPU, 8GB RAM (can offer up to 32 vCPU and 1TB RAM)	>50% average	High (more than 90%)

2.2.7 Applications Catalog

Supply a list of suggested applications and vApps that the private and public vCloud should provide to the consumers. The goal is to help consumers accelerate the adoption of the vCloud service. The vApp templates provided to the consumers can be compliant based on the security policies and also must take license subscription into consideration.

Application workloads generally fall into the following categories:

- **Transient:** A transient application is used infrequently, exists for a short time, or is used for a specific task or need. It is then discarded. This type of workload is appropriate for a pay as you go consumption model.
- **Highly elastic:** An elastic application dynamically grows and shrinks its resource consumption as it runs. Examples include a retail application that sees dramatically increased demand during holiday shopping seasons and a travel-booking application that expands rapidly as the fall travel season approaches. This *bursty* type of workload is appropriate for an allocation consumption model.
- **Steady state:** A steady state application tends to run all the time in a predictably steady state. This type of workload is appropriate for a reservation consumption model.

Table 2.10 illustrates the types of applications in a service catalog, by category.

TABLE 2.10 Applications Catalog Example

Application Type	Application Description
Operating systems	▶ Microsoft Windows Server ▶ RHEL ▶ Centos ▶ SUSe Linux Enterprise Server ▶ Ubuntu Server

Application Type	Application Description
Infrastructure applications	▶ Databases: ▶ Microsoft SQL Server ▶ Oracle Database ▶ MySQL ▶ Distributed data management: VMware vFabric™ GemFire® ▶ Web/application servers ▶ Microsoft IIS ▶ VMware vFabric tc Server ▶ Apache Tomcat ▶ IBM WebSphere Application Server ▶ Simple n-tier applications ▶ Two-tier application with a web front end and database back end ▶ Three-tier application with web, processing, and database ▶ Enhanced three-tier application with added monitoring ▶ Load balancer
Application frameworks	▶ Tomcat/Spring ▶ JBoss ▶ Cloudera/Hadoop
Business applications	▶ Microsoft SharePoint ▶ Microsoft Exchange ▶ VMware Zimbra®

2.2.8 Interoperability

Interoperability aspects of the service definition should list the areas in which the solution must integrate and interact with external systems. For example, a chargeback capability of the solution might need to interoperate with financial and reporting systems. Alternatively, interoperability between vCloud instances built to the vCloud API standards might be required.

2.2.9 Service-Level Agreement

A service-level agreement (SLA) is a negotiated contract or agreement between a vCloud service provider and the consumer that documents the services, service-level guarantees, responsibilities, and limits between the two parties.

General guidelines for vCloud service providers require that any service offering made available carry a comprehensive SLA guarantee that is equal to or exceeds three 9s (99.9%) for availability and reliability, and includes special considerations for overall service performance and customer support handling and responsiveness. An SLA can be either a negotiated contract or a standard contract between a vCloud provider and subscriber that defines responsibilities and limitations associated with the services:

- Availability (uptime)
- Backups (schedule, restore time, data retention)
- Serviceability (time to respond, time to resolution)
- Performance (application performance, network performance)
- Compliance (regulatory compliance, logging, auditing, data retention, reporting)
- Operations (user account management, metering parameters, response time for requests)
- Billing (reporting details, frequency, history)
- Service credits or penalties

Although detailed guidance on how to calculate the level of availability and performance for all vCloud service elements is beyond the scope of this document, it is anticipated that service providers have an SLA framework in place that can be leveraged or augmented to support vCloud services.

SLA guarantees should extend to all facets of a provider's vCloud hosting infrastructure and individual service domains (for example, compute, network, storage, Layer 4–7 services, and management/control plane) that directly support vCloud services. Adherence to SLA requirements should also factor in the resiliency of the management framework, consisting of API and UI accessibility for service subscribers.

2.3 Service Offering Examples

Service offerings and their inherent virtual datacenter constructs provide an effective means of creating service differentiation within a broader vCloud service landscape. They deliver consistent service levels that invariably align with unique business use case requirements, as presented by individual tenants in either a private or public vCloud setting. The service offerings presented in this section serve as a reference for building a differentiated IaaS service model. They also try to address the full spectrum of enterprise workload requirements observed in the vCloud services market today.

The following is a summary of these service offerings:

- **Basic:** Based on the pay as you go allocation model. This service offering lends itself to quick-start pilot projects or test and development application workloads that typically do not require persistent resource commitments or upfront resource reservations.
- **Committed:** Based on the allocation pool allocation model. This service offering gives consumers a minimum initial commitment of resource capacity, plus the added capability to burst above that minimum if additional infrastructure capacity is available at the time of need. The level of minimum commitment, expressed as a percentage of overall capacity per resource type, provides an extra layer of assurance to consumers who seek deterministic performance levels for their application workloads.

- **Dedicated:** Based on the reservation pool allocation model. This service offering gives consumers reserved resource capacity up front, fully dedicated by individual tenant. The level of resource guarantee (always set to 100%) gives customers a higher degree of service assurance than the Committed service offering, along with additional layers of security and resource control for their application workloads.

Because of often unpredictable business demands and the elastic nature of vCloud service consumption models, it is not unreasonable for providers of private or public vCloud instances to seed a service environment with a single service offering type and adapt that service over time, given proper business justification. This approach is not only common, but also recommended, regardless of the number of service offering examples made available for consideration.

To help decide which service offering makes the most sense for a particular set of business use cases, refer to the key service attributes summarized in Table 2.11. The following sections give additional details and reference examples for each service offering.

TABLE 2.11 Service Offering Matrix Example

	Basic Service Offering	Committed Service Offering	Dedicated Service Offering
Allocation Model	Pay as you go	Allocation pool	Reservation pool
Control Plane (Management)	Shared, multitenant	Shared, multitenant	Shared, multitenant
Cluster Resources	Shared, multitenant	Shared, multitenant	Dedicated, single tenant
Unit of Consumption	vApp	Aggregate resource capacity allocated	Aggregate resource capacity reserved
Resource Allocation Settings (per Organization Virtual Datacenter)	—	▶ CPU (GHz) ▶ Memory (GB) ▶ Storage (GB)	▶ CPU (GHz) ▶ Memory (GB) ▶ Storage (GB)
Resource Guarantee Settings	▶ % of CPU ▶ vCPU speed ▶ % of memory ▶ % of storage	▶ % of CPU ▶ vCPU speed ▶ % of memory ▶ % of storage	▶ 100% of CPU ▶ 100% of memory ▶ 100% of storage
Limits (per Organization Virtual Datacenter)	Maximum number of virtual machines CPU (GHz) Memory (GB)	Maximum number of virtual machines CPU (GHz) Memory (GB)	Maximum number of virtual machines CPU (GHz) Memory (GB)
Reporting/ Billing Frequency	Per use	Monthly	Monthly or annually

	Basic Service Offering	Committed Service Offering	Dedicated Service Offering
Metering Frequency	Hourly	Hourly	Hourly
Service Availability	99.95%	99.99%	99.99%
Target Workloads	Test and development	Tier 2 and 3 production	Tier 1 production
Application Workload Examples	▶ Short-term or bursty workloads ▶ QA testing ▶ Integration testing ▶ New software version testing ▶ Short-term data analytics	▶ Static web content servers ▶ Lightly used app servers ▶ Active Directory servers ▶ Infrastructure servers (DNS, print, file) ▶ Small/medium database servers ▶ Short-term content collaboration ▶ Staging sites	▶ Exchange and SharePoint servers ▶ Large database servers (high IOPS) ▶ PCI-related servers ▶ HPC workloads ▶ SaaS production applications ▶ CRM, EDA, ERP, and SCM applications ▶ Financial applications (high compliance)

2.3.1 Service Offering—Basic

The Basic service offering is based on the pay as you go allocation model in vCloud Director. It gives subscribers instant, committed capacity on demand through access to a shared management control plane in a multitenant service environment. Resource commitments for CPU (GHz), memory (GB), and storage (GB) are committed only when virtual machines or vApps are instantiated within the target organization virtual datacenter in vCloud Director. This service is designed for quick-start pilot projects and test and development application workloads that do not typically require persistent resource commitments or upfront resource reservations.

2.3.1.1 Service Design Parameters

As part of the design process for the Basic service offering, providers should give special consideration to key service settings and values in vCloud Director that can impact service performance and consistency levels for a subscriber's organization virtual datacenter. Given the pay as you go allocation model employed in this service, certain circumstances might arise that result in subscribers overcommitting resources over time. If not properly managed, these circumstances can negatively affect performance for all application workloads. Table 2.12 provides an example of these key service settings, values, and justifications.

TABLE 2.12 Resource Allocation Settings Example—Basic Service Offering

Resource Type	Value Range	Sample Setting	Justification
CPU allocation	Variable (GHz), based on physical host capacity	50GHz	The maximum amount of CPU available to the virtual machines running in the target organization virtual datacenter (taken from the supporting provider virtual datacenter).
CPU resources guaranteed	0%–100%	0%	The percentage of CPU resources that are guaranteed to a virtual machine running within the target organization virtual datacenter. This option controls overcommitment of CPU resources.
vCPU speed	0–8GHz	1GHz	This value defines what a virtual machine or vApp with one vCPU consumes at most when running within the target organization virtual datacenter. A virtual machine with two vCPUs consumes a maximum of twice this value.
Memory resources guaranteed	0–100%	75%	The percentage of memory that is guaranteed to a virtual machine running within the target virtual datacenter. This option controls overcommitment of memory resources.
Maximum number of virtual machines	1–unlimited	Unlimited	A safeguard that allows control over the total number of vApps or virtual machines a subscriber can create within the target virtual datacenter.

In this example, the minimum vCPU speed setting is configured as 1GHz (1000MHz), with a memory resource guarantee of 75%. CPU resource guarantees and limitations on the maximum number of virtual machines supported per tenant are optional and can be implemented at the provider's discretion. The provider can use the combination of these settings to change overcommitment from aggressive levels (for example, resource guarantees set to <100%) to more conservative levels (for example, resource guarantees always set to 100%), depending on SLAs in place or fluctuating service loads.

2.3.1.2 Resource Allocation and Catalogs

The pay as you go resource allocation model enables providers to deliver high levels of flexibility in the way resources are allocated, through published vApp catalogs in vCloud Director. vApp catalogs further enable providers to publish standard application images and predefined resource profiles that subscribers can customize, based on a given set of application workload requirements.

Table 2.13 provides an example of different sizing combinations that can be included in a vApp catalog with the Basic service offering.

NOTE

vCPU quantity is based on a multiple of 1GHz, as in the example in Table 2.12. Any quantity of memory or vRAM assigned from Table 2.13 is reserved at 75%. The provider can govern subscribers' capability to select specific quantities of resources, such as vCPU, memory, and storage for a given virtual machine or vApp dynamically, as necessary. However, providers should first implement a pricing model commensurate with the range of scale for each resource type.

TABLE 2.13 Basic Service Offering Catalog Example

vApp Instance Size	vCPU/GHz	OS Bit Mode	Memory[1] (MB)	Storage[2] (GB)	Bandwidth[3] (MBps)	Cost
Extra small	1.0/1GHz	32-/64-bit	500–100,000	10–2,000	Variable	Set by provider
Small	1.0/1GHz	32-/64-bit	500–100,000	10–2,000	Variable	Set by provider
Medium	2.0/2GHz	64-bit	500–100,000	10–2,000	Variable	Set by provider
Large	4.0/4GHz	64-bit	500–100,000	10–2,000	Variable	Set by provider
Extra large	8.0/8GHz	64-bit	500–100,000	10–2,000	Variable	Set by provider

[1] Virtual memory allocation can be customized for all virtual machine instances from small through extra large. The range provided takes into account the maximum amount of memory that can be allocated per virtual machine or vApp in vCloud Director.

[2] Storage allocations can be selected individually and are customizable for all virtual machine instances from small through extra large, based on individual subscriber requirements. The range provided takes into account the maximum amount of storage that can be allocated per virtual machine or vApp in vCloud Director.

[3] Ingress/egress bandwidth allocation can be customized for all virtual machine instances from Small through Extra Large, based on individual subscriber requirements and the Internet service capabilities available at the provider.

The maximum virtual machine instance size is derived from the maximum amount of vCPU and the maximum amount of memory that a physical host has available in the environment. Although the supported ranges for memory and storage in Table 2.13 indicate configuration maximums for a vSphere and vCloud Director environment, these ranges differ for different providers, given the variance in hosting architectures and physical infrastructure designs.

2.3.1.3 Service Metering

Subscribers to the Basic service offering are charged over time for the aggregate amount of resources consumed across their virtual machine and/or vApp inventory for a given organization virtual datacenter. The minimum standard time interval for billing and metering purposes is typically one hour. However, providers who have the means to do so are permitted to meter and charge subscribers for resource consumption on a subhourly basis. If subscribers opt to change the size of their virtual machine or vApp instances after initial

setup, pricing changes retroactively, defaulting to the higher charge rate of either the new or the initial vCPU or memory setting. This is referred to as the *stepping function*—the virtual machine charge always steps up to the next instance size, measured by memory or vCPU, whichever charge rate is higher.

Charges for resource consumption typically begin when the virtual machine is deployed, with limited exceptions for certain resource types such as storage, which may be reserved in advance without immediate use. It is important for providers to understand how different resource states, such as *provisioned* and *reserved,* can be used to determine a chargeable event in a service billing scheme.

Table 2.14 lists the most common event triggers and resource states for vCloud Director. Columns marked with an X signify that the resource type is considered consumed when a virtual machine or vApp is in the associated state; corresponding charges then apply. These are meant to be illustrative only. Providers should rely on their own internal cost models and metering schemes for billing or showback.

TABLE 2.14 vCloud Director Event Triggers and States

API Operation	UI Operation	vCPU	RAM	Network (vNIC)	Storage
Instantiate/ compose	Add/new				X
Deploy	Start			X	X
Power on		X	X	X	X
Reset	Reset	X	X	X	X
Suspend (vApp)	Suspend				X
Suspend (virtual machine)				X	X
Shut down					X
Reboot		X	X	X	X
Power off	Stop			X	X
Undeploy					X
Delete	Delete				
Expire/deploy					X
Expire/storage (mark)					X
Expire/storage (delete[1])					

[1] *The Delete or Expire/storage state means that all resources have been both deactivated and decommissioned, and no further charges should be applied at that point.*

2.3.2 Service Offering—Committed

The Committed service offering is based on the allocation pool allocation model in vCloud Director. It guarantees subscribers a minimum resource commitment through access to a shared management control plane in a multitenant service environment. Resource commitments for CPU (GHz), memory (GB), and storage (GB) are specified by