



VMware vCloud[®] Architecture Toolkit

Public VMware vCloud Service Definition

Version 2.0.1

October 2011

© 2011 VMware, Inc. All rights reserved. This product is protected by U.S. and international copyright and intellectual property laws. This product is covered by one or more patents listed at <http://www.vmware.com/download/patents.html>.

VMware is a registered trademark or trademark of VMware, Inc. in the United States and/or other jurisdictions. All other marks and names mentioned herein may be trademarks of their respective companies.

VMware, Inc
3401 Hillview Ave
Palo Alto, CA 94304
www.vmware.com

Contents

1. Introduction	5
1.1 Audience	5
1.2 Scope	5
1.3 Public vCloud	6
2. Service Definition	8
2.1 Service Offerings.....	9
3. Compliance Definition	14
3.1 Enterprise Hybrid vCloud	14
3.2 Compliance Controls.....	15
3.3 Compliance Visibility and Transparency.....	16
3.4 Compliant Architecture.....	16
4. Architecture Definition	17

List of Figures

Figure 1. Deployment Models	5
Figure 2. Public vCloud Service Offerings	8
Figure 3. Relationship of Private, Public, and Hybrid Clouds	15

List of Tables

Table 1. Service Offerings.....	9
Table 2. Instance-Based Combinations	10
Table 3. Example of Committed Service Offering Definitions.....	12



1. Introduction

Cloud computing is an approach to computing that leverages the efficient pooling of on-demand, self-managed virtual infrastructure that is consumed as a service. The cloud computing approach—made possible by sophisticated automation, provisioning, management, and virtualization technologies—differs dramatically from the current IT model because it decouples data and software from the physical infrastructure that runs them.

VMware vCloud® is the VMware solution for cloud computing. This document provides the service definition for a service provider that provides public Infrastructure as a Service (IaaS) resources.

1.1 Audience

This service definition is intended for those involved in planning, defining, and designing appropriate vCloud services that a service provider can provide to consumers of these vCloud services. The intended audience includes the following roles:

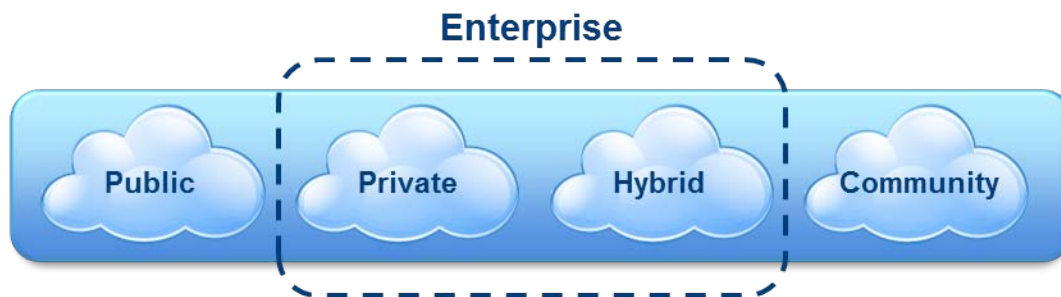
- Service provider as a provider of public vCloud services.
- Architects and planners responsible for driving architecture-level decisions.
- Technical decision makers who have business requirements that need IT support.
- Consultants, partners, and IT personnel who need to know how to create a service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

1.2 Scope

For cloud computing, commonly accepted definitions are defined for the deployment models, and there are several generally accepted service models.

Figure 1 illustrates deployment models for cloud computing.

Figure 1. Deployment Models



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

- *Private Cloud* – The cloud infrastructure is operated solely for an organization and may be managed by the organization or a third party. The cloud infrastructure may be on-premise or off-premise.
- *Public Cloud* – The cloud infrastructure is made available to the general public or to a large industry group and is owned by an organization that sells cloud services.

- *Hybrid Cloud* – The cloud infrastructure is a composition of two or more clouds (private and public) that remain unique entities, but are bound together by standardized technology. This enables data and application portability, for example, cloud bursting for load balancing between clouds. With a hybrid cloud, an organization gets the best of both worlds, gaining the ability to burst into the public cloud when needed while maintaining critical assets on-premise.
- *Community Cloud* – The cloud infrastructure is shared by several organizations and supports a specific community that has shared concerns (for example, mission, security requirements, policy, and compliance considerations). It may be managed by the organizations or a third party and may exist on-premise or off-premise.

The deployment model for the service definition in this document is for a service provider as a provider of services to consumers—a public vCloud.

1.3 Public vCloud

A public vCloud is designed to raise the bar and define a whole new class of enterprise-class cloud computing infrastructure services. In a market currently denominated by commodity, low performance, and insecure public cloud offerings, the public vCloud defines a new enterprise-class cloud computing segment. The public vCloud also can exist as a hybrid cloud solution, enabling enterprises to extend their private cloud to the public cloud with flexibility, scalability, security, and operational efficiency. Through a common platform built around VMware vSphere® and VMware vCloud® Director™, with common management and security models, in an environment that provides on-demand application portability, enterprise customers and leading global service providers are delivering cloud-compatible, connected and integrated hybrid clouds.

1.3.1 Target Markets and Use Cases

Public vCloud services are typically designed to serve corporate and departmental IT teams, government/federal sectors, and the general public. The service allows users to consume public vCloud resources in an evolutionary way, or to augment their private vCloud with public vCloud capacity to support new and existing workloads.

Public vCloud services support both new and existing workloads, as well as the following use cases:

- Pre-production environments – Development/Test/Stage
- Production environments
 - Web applications
 - Marketing/brochure sites
 - Multitiered Web applications
 - eCommerce Web sites
 - Corporate portals and Intranet sites
 - Messaging and collaboration applications
 - Microsoft SharePoint
 - Content/documentation management
 - Internal Wikis/blogs

1.3.2 Challenges Solved

The following are challenges faced by enterprises that want to adopt the public cloud today.

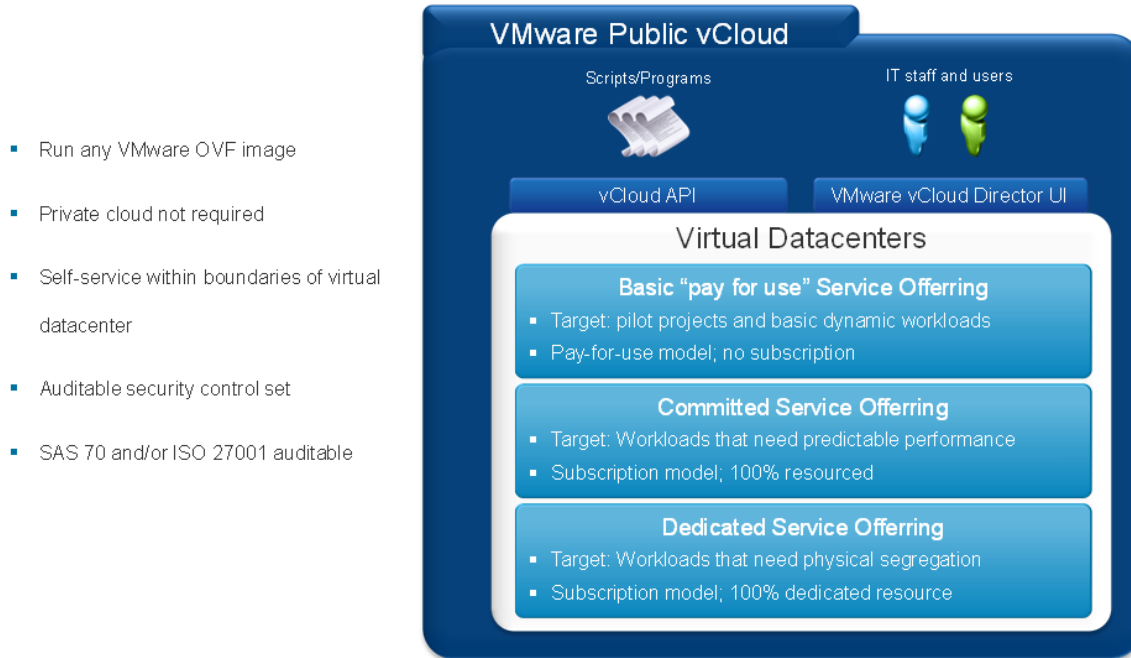
- Trust/Security
 - Alignment with existing processes and tools
 - Regulatory and standards compliance
 - Secure connectivity
 - Data location
- Management
 - Consistent identity and access management
 - Single pane of glass resource management
 - Compatible platforms and API

VMware vCloud provides solutions for these challenges.

2. Service Definition

Public vCloud services typically deliver three classes of on-demand, self-service virtual datacenters as Basic, Committed, and Dedicated service offerings.

Figure 2. Public vCloud Service Offerings



The service makes it as easy as possible for enterprises to move their workloads to public vCloud services. Any existing VMware virtual machine or virtual application (vApp) can be run with little or no modification on a public vCloud service as compatibility with existing enterprise VMware deployments is a key design objective. There is no requirement for an enterprise to deploy a private cloud—any vSphere infrastructure is compatible.

Public vCloud services are built on vSphere, the most secure virtualization platform with EAL4+ and FISMA certifications, and VMware vCloud[®] Director[™], a cloud delivery platform offering secure multi-tenancy and organization isolation. With a public vCloud service, enterprises can exercise the defense-in-depth security best practice as 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 be performed against the enterprise's own LDAP/AD directory, so the enterprise can manage its own user base and provide role-based access according to its own policies. By following to the guidance in this service definition, all public VMware cloud services are auditable under SAS 70 Type II (SSAE16/SOC1/SOC3 or ISO27000 certifications). In addition, all public VMware vCloud services can provide customers relevant audit logs and compliance reports for their vCloud environments so that enterprises can meet their own internal audit requirements. Public vCloud service providers will always have at least one internet routable IP address with the option for adding an additional one, up to provider defined limits, on request.

2.1 Service Offerings

The following service offerings map to the business needs that a public vCloud provides to its consumers. A single consumer may use one or more of these service offerings:

- *Basic Service Offering* – Unreserved pay for use class. Designed for quick start pilot projects or for workloads such as software testing that does not need resource reservations or guarantees.
- *Committed Service Offering* – Provides reserved resources (subscription model) with the ability to burst above the committed levels if additional capacity is available. This offers predictable performance by reserving resources for workloads within a multi-tenant environment, while also allowing on-demand self service.
- *Dedicated Service Offering* – Provides dedicated compute resources (using resources specifically dedicated to a given customer). This offers predictable performance by reserving dedicated resources, which is useful for situations where security or compliance requirements require physical separation.
- *VPC Service Offering* – Provides a completely dedicated “air gapped” set of resources (including management resources) and enables a vCloud to be delivered and dedicated to a customer either on the customer's premise or within the provider's datacenter.

Table 1. Service Offerings

	Basic Service Offering	Committed Service Offering	Dedicated Service Offering	VPC Service Offering
Consumption Model	Pay for use model	Allocation model	Reservation model	Reservation model
Chargeable Unit	CPU, memory, storage	Resource pool	Resource pool	Resource pool
Targeted Use Cases (but not limited to)	Pre-production (test/dev/stage)	Production non-business critical	Production, high performance, security, compliance requirements	Production, high performance, security, compliance requirements
Metering	Hourly	Monthly	Monthly	Monthly
Infrastructure SLA	99.9%	99.9%	99.9%	99.9%

2.1.1 Basic Service Offering

The Basic service offering is an instance-based pay-as-you-go resource consumption model. Each virtual machine provisioned in this virtual datacenter is charged separately and separate billing records are produced for each virtual machine. Customers using the Basic virtual datacenter service are charged for each hour or part hour of consumption.

The following is an example of some of the considerations for a Basic service offering

- An instance-based model refers to the bundling of vCPU and memory together into a single virtual machine instance and price. Pricing can be charged by hour of consumption.
- Memory can be offered in the following unit options: 512MB, 1GB, 2GB, 4GB, 8GB, and 16GB (can offer up to 1TB memory).
- vCPU is offered in the following unit options: 1, 2, 4, and 8 (can offer up to 32 vCPU).
- Pricing can be charged by hour of consumption along the 24 possible combinations of memory and vCPU unit options.

Pricing is not specified here because it is variable and depends on the vCloud service provider's infrastructure, capabilities, location, and agreement with their consumers.

2.1.1.1. Instance-Based Model

The option for an instance-based model refers to the bundling of vCPU and memory into a single virtual machine instance and price. Table 2 shows the possible combinations of the Basic service offering. Each vCPU is equivalent to 1000MHz. Instance-based billing rounds up billing to the hour, so that if a customer uses a machine for 5 minutes they are charged for one hour of usage. If a customer uses a machine for 1 hour and 5 minutes they are charged for 2 hours. If a customer changes the virtual machine size after 5 minutes, this starts a new hour.

The following is only an example. Service providers should perform their own cost analysis to determine final pricing. For each possible combination of memory and vCPU, the service provider determines pricing.

Table 2. Instance-Based Combinations

	1 vCPU	2 vCPU	4 vCPU	8 vCPU	Max vCPU
512MB					
1GB					
2GB					
4GB					
8GB					
16GB					
Max Host Memory					

Pricing is determined by service provider.

The user can change the size of the virtual machine instance to arbitrary sizes, but the price of that virtual machine is always the price of the maximum amount of memory or vCPU. The charge begins when the virtual machine is deployed. This is referred to as the *stepping function*. The virtual machine charge always steps up to the next instance size, either by memory or vCPU, whichever price is greater.

The last row in Table 2, Max Host Memory, combined with Max vCPU, is designed so that there is an upper bound for the virtual machine. The maximum virtual machine size is the maximum amount of vCPU and the maximum amount of memory that a physical host has. The idea is to put a price on a virtual machine size that is maximum vCPU and maximum memory so that any virtual machine size above 16GB of memory and 8 vCPU has a specific charge.

As part of the service offering, there may also be other charges.

2.1.2 Committed Service Offering

The Committed service offering uses the allocation consumption model. A user is allocated a virtual datacenter that contains a certain amount of CPU (GHz), memory (GB), and storage (GB). The allocation consumption model is defined using two parameters: the reservation percentage and the total allocation (also called *limit*). The reservation percentage is how much resource is guaranteed or committed for the customer. The total allocation or limit is the maximum amount of resource the customer can consume.

2.1.2.1. Example

The following is an example of some of the considerations for a Committed service offering.

For the Committed service offering, the reservation percentage is set to 75% of the total allocation/limit. This means the customer can burst up to an additional 25% of resources they originally requested. So if a customer is allocated 10GHz of CPU resources, a virtual datacenter is created and 10GHz is allocated for the customer. This is the maximum amount of CPU the customer can ever consume. Of this 10GHz, the enterprise cloud service provider should reserve 75%, which is 7.5GHz. This is the amount of CPU that is guaranteed for the customer. The 25%, or 2.5GHz, is available to the customer if the underlying cluster has available resources. This model allows the service providers to charge a price that is generally higher than just 7.5GHz given the additional burstable capacity, but gives customers the benefit of potentially paying less than if the resources were fully guaranteed.

This consumption model can also be positioned to the customer where the percent reserved is the plan they are buying and they have the ability to burst beyond that capacity only if there are available resources—there are no guarantees that they can go above their reserved limit. If a customer needs additional capacity, they should consider moving to the next size virtual datacenter or request incremental reserved capacity.

Table 3 is an example of how virtual datacenter sizes are defined for the Committed service offering.

Table 3. Example of Committed Service Offering Definitions

	Small	Medium	Large	X-Large	Custom
CPU Reserved	7.5GHz	18.75GHz	37.5GHz	75GHz	N/A
CPU Limit	10GHz	25GHz	50GHz	100GHz	N/A
Memory Reserved	15GB	37.5GB	75GB	150GB	
Memory Limit	20GB	50GB	100GB	200GB	N/A
Storage	400GB	1TB	3TB	6TB	N/A
Approx Virtual Machines (Not limit)	10-20	25-50	50-100	100+	N/A

The approximate virtual machine count is calculated based on a distribution ratio assumption. In most virtualized environments, memory is the gating factor in terms of resource utilization. Based on the distribution ratio, on average a virtual machine consumes 2.15GB of memory. For a 20GB virtual datacenter, that is approximately 10 virtual machines. If a customer runs more small virtual machines, that count can increase. Thus, setting the virtual machine limit to approximately 20 gives customers some flexibility to run more small virtual machines, but still manage the capacity effectively.

As part of the service offering, there are also other charges.

2.1.3 Dedicated Service Offering

The Dedicated service offering uses the reservation consumption-based model. A customer works with a service provider to provision a cluster of servers that is dedicated to this customer. The hardware (network, storage, servers) is not shared with other customers, but the service provider management cluster and interface is shared. The customer gets full control over the reservation and limit of this set of resources.

This service offering is likely to be a fixed price, monthly subscription. Given that this service offering is completely customized, no price guidance is provided. However, other charges may apply even in the Dedicated virtual datacenter model.

2.1.4 VPC Service Offering (Optional Service Level)

The VPC service offering uses the reservation consumption-based model. A customer works with a service provider to provision a cluster of servers that is dedicated to this customer. The hardware (network, storage, servers) and management cluster or interface is not shared with other customers. The customer gets full control over the reservation and limit of this set of resources.

This service offering is likely to be a fixed price monthly subscription. The VPC virtual datacenter may be deployed at the customer's site, within a service provider's datacenter, or in a contracted co-located space.

Given that this service offering is completely customized, no price guidance is provided. However, other charges may apply even in the VPC virtual datacenter model.

3. Compliance Definition

Security and compliance continues to be one of the biggest barriers to adoption of the public cloud by enterprise customers. Most regulations and mandates in the industry, including SOX, PCI, HIPAA, COBIT, and ISO, have two areas of requirements: transparency/visibility and control.

Transparency is an absolute requirement because cloud consumers must know who has accessed what data, when, where, and potentially why, based on documented evidence. PCI requirement #10 is a good example of the need for visibility and transparency.

Control is also a necessary component of compliance for cloud consumers. For example, cloud consumers must be able to control who can access, configure, and modify the cloud environment, what firewall ports are open, when to apply patches, and where the data resides. Cloud consumers, and especially enterprise customers, believe “you can outsource responsibility, but you can’t outsource accountability.” Ultimately, cloud consumers are accountable for compliance.

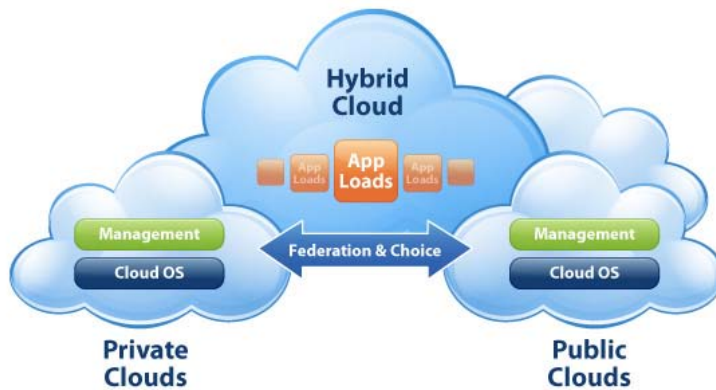
Public vCloud services can be designed to address this problem in the following ways:

- Facilitate compliance through ISO27001 certification or SAS70 Type II audit, based on a standard set of controls.
- Provide compliance logging and reports to the customers so that they have full visibility into their public vCloud environments.
- Architect the service so that customers can control the access to their vCloud environments

The following sections provide a set of high-level requirements for a public vCloud service.

3.1 Enterprise Hybrid vCloud

- A private vCloud is a cloud infrastructure operated solely for an organization. It may be managed by the organization or a third party, and it may exist on-premise or off-premise.
- A public vCloud is a cloud infrastructure made available to the general public or a large industry group, and is owned by a service provider that sells cloud services.
- A hybrid vCloud is a cloud infrastructure composed of two or more vCloud instances (private or public) that remain unique entities, but are bound together by standardized technology that enables data and application portability (for example, cloud bursting for load-balancing between clouds).

Figure 3. Relationship of Private, Public, and Hybrid Clouds

A common misperception is that cloud computing implies an external cloud based on public cloud services such as Amazon. The fact is that cloud computing is how you approach IT. It is “a way of doing computing,” not a destination. Ultimately, most enterprises will benefit from adopting cloud computing within their own datacenters, building private clouds, and getting there in an evolutionary way through their existing virtualization journey.

Through a common platform built around VMware vSphere and VMware vCloud Director, and with common management and security models, service providers have the capabilities they need to provide on-demand application portability.

3.2 Compliance Controls

To make sure the enterprise customers feel secure and safe in the public vCloud, and that they have the necessary information and visibility into the service to meet their own internal audit requirements, public vCloud services should have one of the following certification or audit completed:

- ISO27001 certified, which proves that security management processes are in place, and have a relevant subset of the ISO27002 controls in place as specified by the VMware Compliance Architecture and Control Matrix.
- SSAE16/SOC1/3 audits based on the same relevant set of controls.

VMware supplies the standard set of compliance controls, and the service provider is responsible for the actual ISO or SSAE16/SOC1/3 audits using third-party auditors. The compliance controls should be published to the provider’s public vCloud customers so that they understand that the public vCloud is compliant and that the customers have full visibility into what controls the services were audited against.

3.3 Compliance Visibility and Transparency

Log management is required by many of the compliance frameworks such as ISO, HIPAA, PCI, and COBIT. Companies must meet the requirements of these audit standards. Enterprise customers demand that service providers provide them with visibility into their public vCloud environments. For example, enterprise customers must have access to all of the necessary logs and reports associated with user activities, access control, and firewall connections.

To meet the compliance requirements, service providers should provide vCloud consumers visibility and transparency into the public vCloud service. To accomplish this, service providers must be able to collect and maintain logs for all components of the vCloud service and provide relevant logs back to the consumers. The service provider may choose to keep the logs for the underlying infrastructure private, but in this case, the service provider must be willing to provide them to a customer for an audit. In general, public vCloud services should have logs that cover the following components of a customer's environment, and make them available to their customers in a proactive fashion:

- VMware vCloud Director
- VMware vShield Edge™

Public vCloud services are based on a set of products that were proven in many secure environments, and products such as vCloud Director and VMware vShield generate a set of logs that give customers visibility into all user activities and firewall connections. VMware provides the vCloud reference architecture and best practices so that service providers can capture this set of logs and provide them to customers.

In addition to logs, service providers should provide basic compliance reports to customers so that they understand all of the activities inside their vCloud environment. VMware provides a set of best practices to help the public vCloud service meet customer requirements. The service provider is responsible for maintaining logs for their public vCloud service as well as their customer's environments. This capability must be implemented and validated before the service is launched as generally available.

3.4 Compliant Architecture

All public vCloud services offer unparalleled security. Public vCloud services are built on VMware vSphere, the most secure virtualization platform with EAL4+ and FISMA certifications, and vCloud Director, a cloud delivery platform that offers secure multi-tenancy and organization isolation. With the public vCloud service, enterprises can exercise the defense-in-depth security best practice as 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 be performed against the enterprise's own LDAP/AD directory, which means that the enterprise can manage its own user base and provide role-based access according to its own policies.

4. Architecture Definition

To take full advantage of the hybrid vCloud, the vCloud reference architecture is designed to be compatible with the private vCloud stack that VMware advocates for enterprises. To support service providers in implementing this service, VMware provides a full set of reference blueprints to the service providers. This set of blueprints includes all of the documentation and best practices on understanding, architecting, sizing, and implementing an enterprise-class cloud infrastructure. It represents not only hundreds of person years of product knowledge, but also many person years of knowledge in building out scalable virtual and cloud infrastructures.

The public vCloud architecture is designed with the following products, all of which are required parts of the reference architecture.

- VMware vCloud Director
- VMware vSphere
- VMware vCenter Chargeback™
- VMware vShield Edge, which is embedded in VMware vCloud Director

The public vCloud services can also expose the vCloud Director user interface as the main provisioning and management interface for end users, as well as vCloud API for automation.