



VMware vCloud™ Service Definition for a Private Cloud

Version 1.6

TECHNICAL WHITE PAPER

© 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

Table of Contents

1. Introduction	4
1.1 Phase I	4
1.2 Phase II	4
2. Workload Categories For Private Cloud	4
2.1 Transient	4
2.2 Highly Elastic	4
2.3 Infrastructure	5
3. Use Cases	5
4. Service Tiers	6
5. Roles and Rights	7
6. Metering Using Chargeback	7
7. Suggested vApp Catalog	8
7.1 Operating Systems	8
7.2 Infrastructure Apps	8
8. Other Considerations	8
8.1 Operational Components	8
8.2 Sizing Assumptions	9

1. Introduction

The service defined here will describe the creation of a fully virtualized, pooled compute platform for use as an enterprise private cloud, or business-internal organizational cloud computing service, commonly referred to as Infrastructure as a Service (IaaS). VMware technologies including vSphere, vShield Manager, vCenter Chargeback, and vCloud Director will be used to establish this service.

The result of this service includes secure multi-tenancy for lines of business, shared virtualized assets, a self-service user portal, standard catalogs of pre-defined virtual machines and applications with usage metering. Going forward, this service will be enhanced with greater capacity and the preparation for linking with other cloud services either within or beyond the boundaries of the enterprise.

1.1 Phase I

Goals:

- Deliver a fully operational private cloud infrastructure
- 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 internal cost distribution
- Establish a catalog of common infrastructure & application building blocks
- Provide a sizing capacity of 400 virtual machines

1.2 Phase II

Goals:

- Enhance the cloud service for federation to public cloud resources
- Provide for workload redundancy and continuity options
- Provide a sizing capacity of 1,000 virtual machines

2. Workload Categories For Private Cloud

Private cloud use cases generally fall into three different categories of workloads.

2.1 Transient

A transient application is one that is used infrequently, exists for a short time period, or is used for a specific task/need. It is then discarded. This type of workload is appropriate for a pay-as-you-go allocation model in VMware vCloud Director.

2.2 Highly Elastic

An elastic application is one that is dynamically growing and shrinking its resource consumption as it runs. An example of this would be a retail application that sees dramatically increased demand during holiday shopping seasons, or a travel booking application that expands rapidly as the fall travel season approaches. This bursty type of workload is appropriate for the allocation pool model in VMware vCloud Director.

2.3 Infrastructure

An infrastructure application is one that tends to run all the time at a predictably steady state. Examples of these include a print server, a directory server, an email relay server, or a systems monitoring engine. This type of workload is appropriate for a reservation pool model in VMware vCloud Director.

3. Use Cases

The table below lists common use cases that fall into the general cloud workload cases. These come from VMware customer engagements and have been found to be the most popular workloads.

EXAMPLE	
Software Testing	Software Development
Website Design	Custom Applications
File & Print Services	Email
Sales & Demonstration	Custom Applications
File & Print Services	Contractor Access

Table 1. Workload Examples

Commonly, virtual lab automation use cases are good fits for a private cloud, including integration and performance testing, dev-test-QA workflows for software authoring and general access for developers. Also, a private cloud is well suited for sales demonstrations, training staff and partners on software and easily accessing prebuilt configurations of complex application suites. Finally, a private cloud can be used to establish expanded capacity outside of the realm of a traditional MSP (hosting provider) for virtual lab use or lab bursting based on project needs.

With the knowledge of the workload types and common uses cases you can begin to define the service tiers that you will offer inside of your enterprise.

4. Service Tiers

Tiering of service within the enterprise can be an effective method of differentiating types, location, or owners of infrastructure as reflected in a private cloud. The following service tiers map to the needs defined by the three main workload types mentioned above.

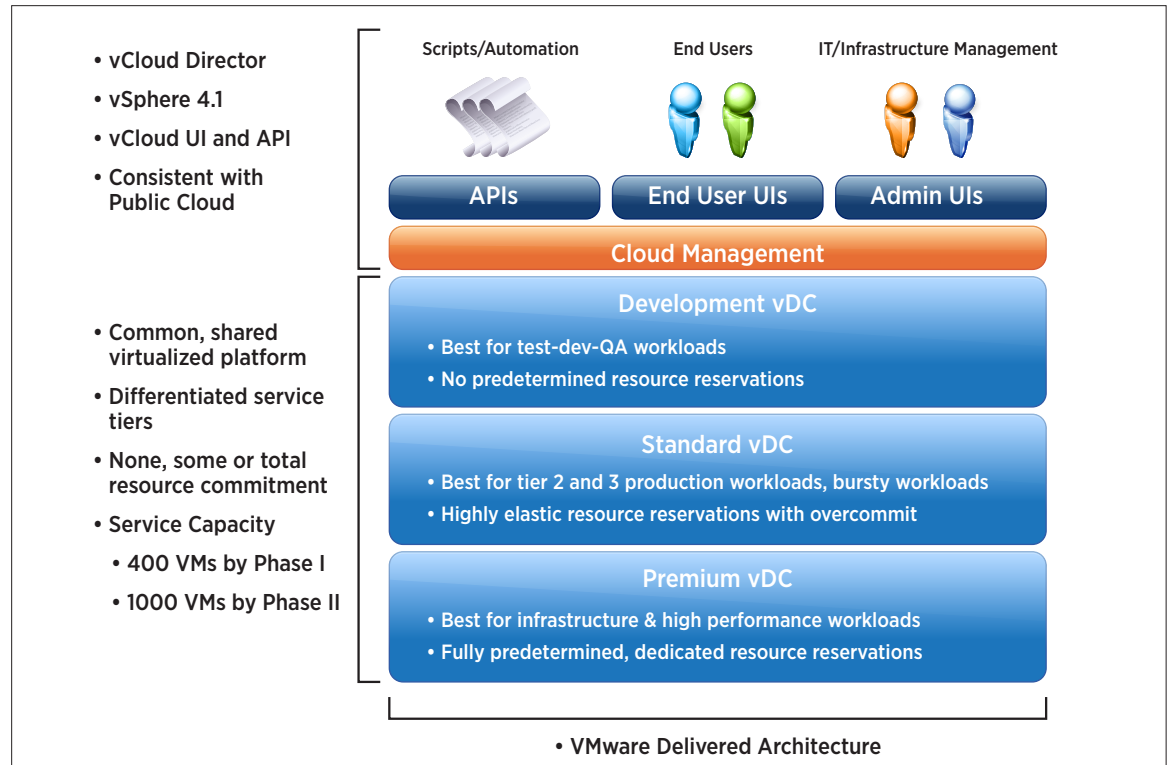


Figure 1. vCloud Service Tiers

	DEVELOPMENT vDC	STANDARD vDC	PREMIUM vDC
Consumption Model	Pay-as-you-go	Allocation pool	Reservation pool
Use Case(s)	Dev-Test-QA	Tier 2/3, Seasonal or Bursty	Tier 2, Print/Directory/Email
Workload Category	Transient	Highly Elastic	Infrastructure
Chargeable Unit	CPU, Memory, Storage	Resource pool	Resource pool
Catalog	Private	Private	Private

Table 2. Service Tier Descriptions

5. Roles and Rights

There are several roles defined in the access model of vCloud Director, including administrative roles at both the system level and at the Organization (vSphere's logical collection of users, groups, and computing resources) or private cloud level. Within the Organization, there are levels of rights granted to pre-defined roles that have an important impact on how users interact with the cloud UI.

At a minimum, you'll want to define a system administrator, an organizational administrator for each tenant, and a secondary, non-administrative user per tenant/private cloud.

ACCOUNT TYPE	NEEDS	PURPOSE
System Administrator	One (minimum)	Highest level administrator; has super user rights
Organization Administrator	One per Organization	Administrator in the Organization over systems and users
Organization Author	One or more, as needed	Allows vApp and Catalog creation; no infrastructure management

Table 3. Minimum Roles and Rights

6. Metering Using Chargeback

A concept highly desired by the enterprise, but historically very difficult to impossible to implement, is the notion of metering the use of the virtual infrastructure for the purposes of inter-departmental billing. Cloud computing enables this capability and thus should be considered as part of a standard service installation.

How this metering data is used falls to the architects of the infrastructure, but at a minimum the resource pools that back each provider virtual data center should be monitored with VMware vCenter Chargeback and the resulting reports provided to the line of business on a quarterly basis (for the purposes of 'showback').

The following table gives some examples only of workload virtual machine sizing and costing.

VIRTUAL MACHINE TYPE	SIZING	STORAGE	COST MODEL	
Large	4 vCPU x 8GB RAM	200G	Provision: \$400	Operate: \$200/mo.
Medium	2 vCPU x 2GB RAM	60G	Provision: \$300	Operate: \$100/mo.
Small	1 vCPU x 1GB RAM	30G	Provision: \$200	Operate: \$50/mo.

Table 4. Workload Virtual Machine Sizing and Costing Examples

7. Suggested vApp Catalog

The following is a list of suggested vApp templates that would help promote immediate use of the private cloud.

7.1 Operating Systems

- Microsoft Windows Server 2003 R2 Enterprise Edition
- Microsoft Windows Server 2008 R2 Enterprise Edition
- RHEL 5.x
- Centos 5.x
- Novel SUSE Linux Enterprise Server 11
- Ubuntu Server 10.04

7.2 Infrastructure Apps

- Databases
 - Microsoft SQL Server 2000/2005/2008
 - Oracle 11g
 - MYSQL 5.x
- Web/App Servers
 - Microsoft IIS
 - Spring tcServer
 - Apache Tomcat
 - IBM Websphere Application Server 7
- Simple n-Tier Apps
 - 2-Tier app with a web front-end and database backend
 - 3-Tier app with web, processing and database
 - Enhanced 3-Tier with added monitoring
- Load balancer

8. Other Considerations

8.1 Operational Components

When creating a private cloud there are other operational components that you will need, if you do not already have them in place, in order to maintain a good experience for the consumers of the cloud. Some of these operational components are:

- Ticketing
 - Helpdesk support (online or phone)
 - Optional: online ticket viewing
- Monitoring
 - Monitoring of all infrastructure components
- Metering
 - Quarterly showback reporting to the line of business
 - 12 months of metering history

8.2 Sizing Assumptions The following are assumptions that can be used for service capacity planning.

- Capacity assumption
 - 400 virtual machines by Phase I
 - 1,000 virtual machines by Phase II
 - 75% small / 20% medium / 5% large
- Distribution assumption
 - 45% small (1 GB, 1 vCPU)
 - 35% medium virtual machines (2 GB, 2 vCPU)
 - 15% large virtual machines (4 GB, 4 vCPU)
- Storage assumption
 - Average 60 GB per virtual machine

The following table gives some examples of workload virtual machine sizing and utilization.

VIRTUAL MACHINE TYPE	SIZING	CPU UTILIZATION	MEMORY UTILIZATION
Large	4 vCPU x 8GB RAM	>50% average	High (upwards of 90%)
Medium	2 vCPU x 2GB RAM	20-50% average	Moderate (50% - 75%)
Small	1 vCPU x 1GB RAM	10-15% average	Low (10% - 50%)

Table 5. Workload VM Sizing and Utilization Examples