VMware vCloud® Architecture Toolkit™
for Service Providers

# Public Cloud Service Definition

Version 2.9

January 2018

Adrian Roberts

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

# Contents

## List of Figures

## List of Tables

# Introduction

The VMware Cloud Provider™ Program is a global network of approximately 4,000 service providers who have built their cloud and hosting services on VMware software. These service providers deliver world-class cloud and hosting services to their customers across the globe, offering value-add and differentiated services that support a wide choice of compliance requirements, performance, scale, market coverage, functional features, and so on. In this way, service providers give existing and new VMware enterprise customers many options when they choose to build out their unified hybrid cloud strategy.

The VMware vCloud® Architecture Toolkit™ for Service Providers supplies architectural guidance on how to build VMware based cloud platforms based on real world service models, implementation examples, use cases, and customer requirements.

This document enables service providers to define their cloud service, understand what use cases they want to support, and what services they want to take to market. From this document, the desired architecture can be positioned to build a VMware Powered Public Cloud service to offer infrastructure as a service (IaaS), platform as a service (PaaS), or software as a service (SaaS) to their customers.

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

- Service Providers of VMware powered cloud 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 VMware powered cloud services.

## 1.1 VMware Powered Public Cloud Overview

VMware Powered Public Cloud platforms are built on the same core technology that drives the VMware public cloud—VMware vCloud Air™. This enables the provider to offer their customers complete, secure multi-tenancy with unparalleled efficiency, security, performance, and scalability expected by cloud consumers.

A VMware Powered Public Cloud Is typically built with the following core principles:

- The cloud service must be built with VMware vSphere® and VMware vCloud Director® at its core.
- The vCloud APIs must be exposed to the cloud tenants.
- Cloud tenants must be able to upload and download virtual workloads packaged with the Open Virtualization Format (OVF) version 1.0.

Cloud providers can also obtain a certification badge which validates their implementation against a number of standards. For more information, go to http://vcloudproviders.vmware.com.

## 1.2 Deployment Model

Service providers typically have three different cloud deployment models that they can offer to their customers:

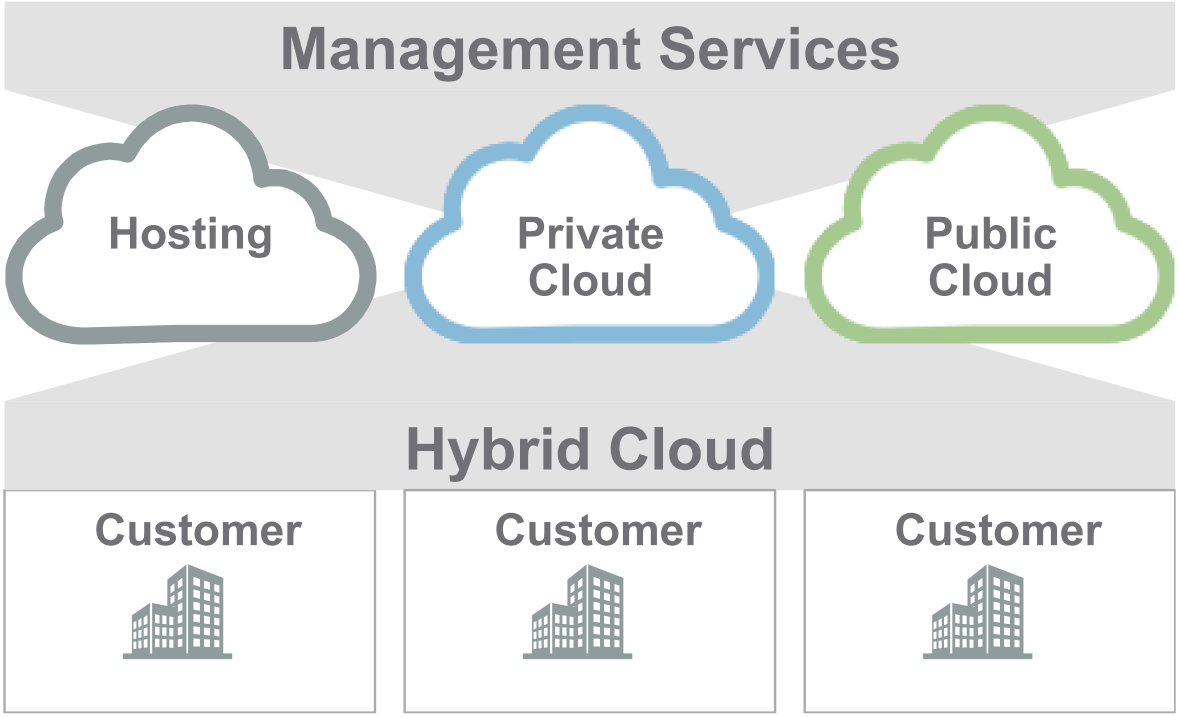
- Hosting (managed or unmanaged) – VMware Cloud Provider Program Powered Hosting Services offer all the benefits of a dedicated software-defined data center and are engineered on VMware vSphere to be fully compatible with customers' on-premises vSphere environments. This offers a unified hybrid cloud experience with the same advantages of improved availability, recoverability, performance and scalability to run your business critical applications with confidence. The hosting solution can either be managed by the provider or self-managed.
- Private Cloud (managed or unmanaged) – VMware Cloud Provider Program Powered Private Cloud Services are engineered on VMware vRealize® Suite, and is fully compatible with customers' on-premises vSphere environments. This provides a unified hybrid cloud experience and dedicated software-defined data centers, offering the required self-service consumption, availability, performance, and scalability to run your business critical applications in the cloud. The private cloud solution can either be managed by the provider or self-managed.
- Public Cloud – VMware Cloud Provider Program Powered Public Cloud Services are engineered on VMware vCloud Suite® with vSphere and VMware vCloud Director at the core. This unique combination provides complete multi-level security and a multi-tenant architecture that reduces complexity and supports policy implementation that can be consistent with your internal data center and vCloud Air, offering a unified hybrid cloud experience to the consumers.

All three models can be complimented with associated management services. The service provider can offer managed services on top of their core IaaS, PaaS, or SaaS offerings, such as:

- Professional services (managed creation)
- Patching
- SLAs
- Recoverability options
- Monitoring capabilities

This service definition focuses on the public cloud deployment model as shown in the following figure.

**Figure 1. Deployment Models**



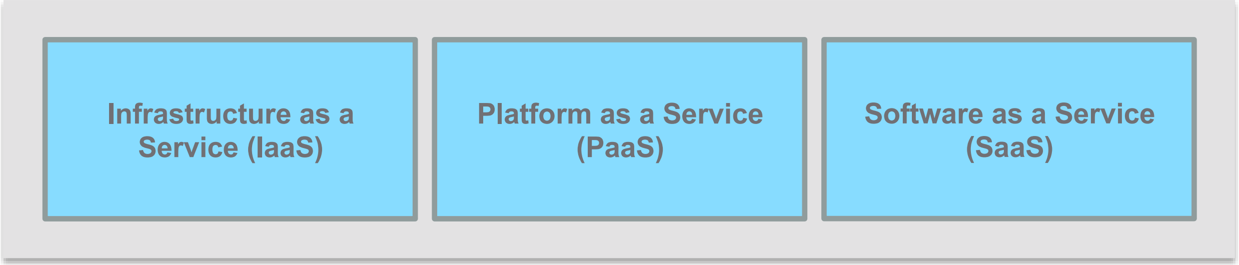
## 1.3 Service Model

Based on the hybrid model above, public cloud service can offer a multitude of services to customers. Typically, services can fall under one of three service models. VMware defines these service layers as:

- Infrastructure as a Service (IaaS) – Infrastructure containers are presented to consumers to provide agility, automation, and delivery of components.
- 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.

This service definition primarily focuses on Infrastructure as a Service. A service provider can, however, include additional "as a Service" offerings on top of the core cloud platform.

**Figure 2. Service Models**


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

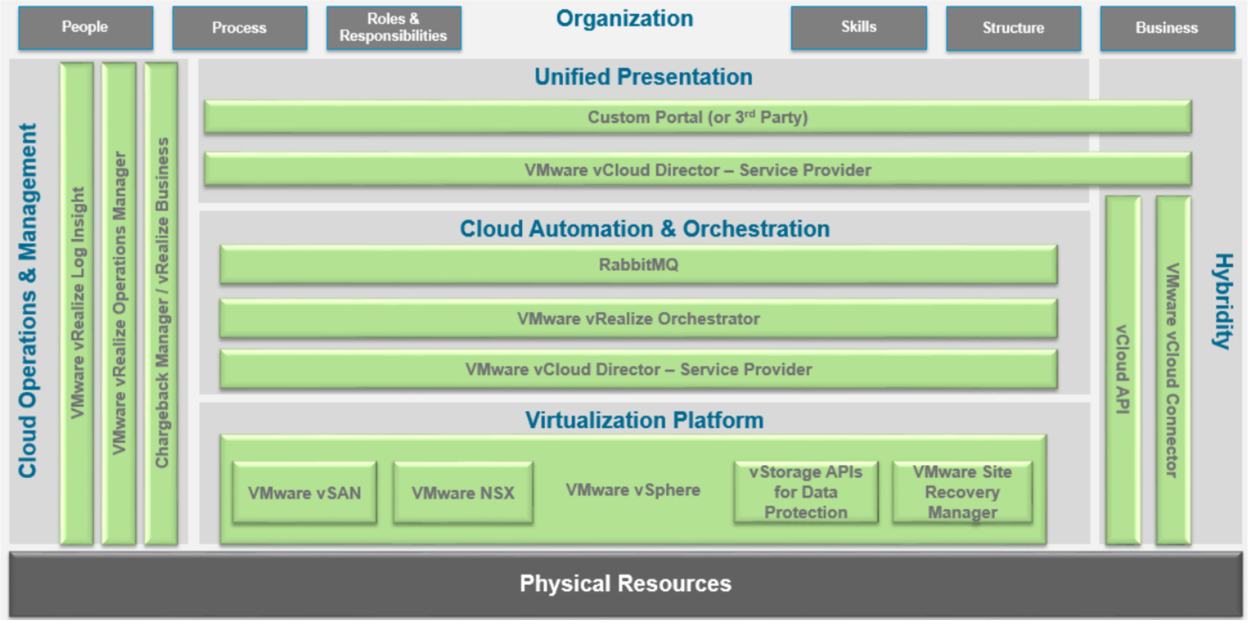
## 1.4 VMware Technology Mapping

The following list highlights the recommended VMware products required to build and operate a VMware powered public cloud platform:

- vSphere
- VMware vSAN™
- VMware NSX® for vSphere
- VMware vRealize Orchestrator™
- vCloud Director for Service Providers
- Custom portal or third-party
- VMware vRealize Operations Manager™
- VMware vRealize Log Insight™
- VMware vRealize Business™
- VMware Site Recovery Manager™
- VMware vCloud Connector®

Although this list is the recommended solution stack, some of the components are optional. For example, vSAN is not required and can be substituted by a traditional FC, ISCSI, or NFS-based storage array.

**Figure 3. Technology Mapping**


People
Process
Roles & Responsibilities
Organization
Skills
Structure
Business
Cloud Operations & Management
VMware vRealize Log Insight
VMware vRealize Operations Manager
Chargeback Manager / vRealize Business
Unified Presentation
Custom Portal (or 3rd Party)
VMware vCloud Director – Service Provider
Cloud Automation & Orchestration
RabbitMQ
VMware vRealize Orchestrator
VMware vCloud Director – Service Provider
Virtualization Platform
VMware vSAN
VMware NSX
VMware vSphere
vStorage APIs for Data Protection
VMware Site Recovery Manager
vCloud API
VMware vCloud Connector
Hybridity
Physical Resources

VMware vCloud Director for Service Providers 8.0 and forward requires vRealize Business for chargeback/showback functionality. Pre-8.0 releases of vCloud Director for Service Providers can leverage VMware vCenter® Chargeback Manager™.

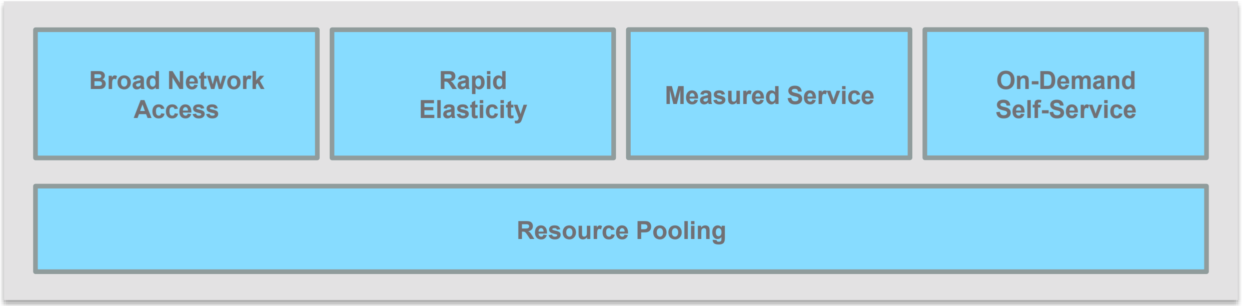
## 1.5 Service Characteristics

The NIST 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 a 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 multi-tenant 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.

The following figure shows the relationships among service characteristics.

**Figure 4. Service Characteristics**



To deliver business solutions using VMware public cloud services, the cloud 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 data center fabric, providing comprehensive infrastructure services capable of supporting any and all applications. A holistic infrastructure can support any workload, 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 cloud 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.

## 1.6 Service Development Methodology

The best practices approach for defining and designing VMware public cloud service:

- 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 lifecycle, including:
  - Setup, which includes definition and design
  - Request and approval
  - Provisioning
  - Consumption

  - Management and operations
  - Transition and termination

  There must be a conscious awareness of what consumers and the provider of the service experience at each stage of the service lifecycle 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.
- Understands the service's 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 SLA for the services and service components in the following areas:
  - Infrastructure
  - Application / VMware vSphere vApp™
  - Platform
  - Software
  - Business
- 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.

## 1.7 Concepts and Terminology

Key service terms and 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.
- VMware Powered Public Cloud – 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.
- Cloud service provider (or provider) – An entity that provides VMware Powered Public Cloud services to consumers.
- Consumer or customer – Someone who consumes VMware Powered Public Cloud 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 verify that the cloud service design is fit for its

purpose. Service-level targets must 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 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.
- vCloud Suite – The suite of VMware technologies that provides the solution for cloud computing.
- VMware vRealize Suite – The suite of VMware cloud management technologies that support the VMware Powered Public Cloud implementation model.

## 1.8 Target Markets

VMware Powered Public Cloud services are designed to provide enterprise-grade unified hybrid cloud capabilities to the provider’s customers, offering seamless extension of on-premises data center services to the cloud, business mobility options, and support for many different application architectures ranging from hybrid applications, development applications, and cloud native applications to Tier 1 business critical applications. This offers the customer the correct balance of on-demand agility with all the availability, business continuity, security, performance, and scalability that they have come to expect with VMware products.

# 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 into account the following considerations when developing a service definition. These considerations 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)

## 2.1 Service Objectives

Understanding the service objectives is an essential first step to creating a service definition. Service objectives must address the specific business challenges. The following are examples of service objectives for public cloud services:

- Deliver a fully operational public cloud infrastructure.
- Provide secure multi-tenancy for public cloud 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:
  - Virtual private cloud on-demand (pay for resources used)
  - Virtual private cloud (allocated resources)
  - Dedicated (reserved resources)
- Support a minimum of 1,500 virtual machines across the three service offerings with a plan to grow to a minimum of 5,000 virtual machines.
- Provide workload mobility between private and public cloud environments, enabling consumers to import and export workloads easily.
- Provide upstream network connectivity for applications with upstream dependencies.
- Provide an isolated network for applications that must be isolated.
- Provide open, interoperable, and Internet-standard protocols for consuming cloud resources.
- Provide workload redundancy and data protection options.

## 2.2 Use Cases

The use cases in the following tables represent business problems (some general and some industry-specific) that can be addressed with VMware Cloud Provider Program services and represented by a service definition.

**Table 1. Example: Use Case 1**

| **Use Case UC_01** | |
|---|---|
| Name | Business continuity and disaster recovery. |
| Problem Statement | The need to protect existing business services, processes, and applications in the event of a disaster. |
| Description | Business continuity and disaster recovery of business services, processes, and applications. |
| Requirements/Goal | • Protect virtualized infrastructure.<br>• Protect applications.<br>• Allow continuity of business processes in event of a disaster, |
| Risks | • Loss of business capability in the event of a disaster.<br>• Lack of compliance with disaster recovery mandates**.** |

**Table 2. Example: Use Case 2**

| **Use Case UC_02** | |
|---|---|
| Name | Increase 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.<br>• Enable IT to scale up, down, in, or out to support business demand.<br>• Scale within a short cycle of days in order to meet projected demand.<br>• Scale to off-premises capacity. |
| Risks | • Lost revenue due to lack of capacity.<br>• Lost customers from underperforming business services. |

Table 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 the development and test infrastructure. |
| Description | IT needs to be able to provide on-demand self-service provisioning of the development and test infrastructure to support the business to rapidly develop new products and services. |
| Requirements/Goal | • Give developers and test users access to a catalog of IT infrastructure services that they can rapidly provision and use.<br>• Provide self-service provisioning, with necessary approvals.<br>• Reduce time-to-market for products and services. |
| Risks | Products and services are late to market, resulting in loss of customers and market share. |

Table 4. Example: Use Case 4

| **Use Case UC_04** | |
|---|---|
| Name | Security and compliance assurance. |
| Problem Statement | The business is concerned about putting crucial financial applications and data on public cloud services. |
| Description | IT must be able to provide secure business services for financial applications and data, have controlled access, and be separated from other users of the cloud services. |
| Requirements/Goal | • Provide compliance controls and transparency for the service.<br>• Provide network isolation for applications that must be isolated. |
| Risks | Security and compliance breach. |

**Table 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 launch.<br>Give consumers access to a catalog of IT infrastructure services that they can rapidly provision and use. |
| Risks | Products and services are late to market, resulting in loss of customers and market share.<br>Lost opportunity cost. |

## 2.3 User Management and Identities

There are several built-in administration and user roles that can be associated with users or groups of users within vCloud Director for Service Providers. This is important because the architect must verify that appropriate user roles are associated with the correct users so that they can perform their business tasks.

This section discusses the different identity sources, user types, authentication controls, roles, and rights present in vCloud Director for Service Providers. An understanding of this information is required to properly secure the system and provide the correct access to the appropriate people.

### 2.3.1 About Users, Groups, Roles, and Rights

A user is a member of a single Organization or is a provider user. Users are assigned a role, and a role is assigned a set of rights. Users can be local users (only stored in the Oracle database) or LDAPv3 users imported into the database. Users can also be members of one or more groups imported from an LDAPv3 directory, potentially assigning an additional role for each group of which they are a member.

No unauthenticated user is allowed to access any vCloud Director for Service Providers functionality, whether the access is through the vCloud API or the Web UI. Thus, all individuals that you want to access vCloud Director for Service Providers must be imported from LDAP, be members of LDAP groups you import into the system, or be managed by an Identity Provider (IdP). Each user authenticates using a user name and password. No other authentication methods are supported in this release of vCloud Director for Service Providers. It may be possible to proxy or layer a stronger authentication method in front of the vCloud API and the Web UI, but these configurations have not been tested by VMware and are not supported.

Groups are not created in vCloud Director for Service Providers. Instead, they are imported from the LDAPv3 directory associated with the system (provider) level or Organization. Groups allow users to authenticate to VMware vCloud Director for Service Providers without the need to create users in the database or manually import them from the Directory (LDAPv3) server. Instead, users can log in if they are a member of a group already imported from the Directory (LDAPv3) server. A user that is a member of multiple groups is assigned all the roles assigned to those groups.

Roles are groupings of rights that provide capabilities for the user assigned that role. The predefined roles are described in the “Roles and Rights” chapter of the *VMware vCloud Director Administrator’s Guide*. The administrator’s guide identifies which rights are assigned to each role to help you choose the appropriate role for each type of user.

For example, the vApp user role might be appropriate for an administrator that needs to power on and off virtual machines, but if they also need to edit the amount of memory assigned to a virtual machine, vApp Author would be a more appropriate role. These roles might not have the exact sets of rights relevant to your customers’ organizations, so you also have the ability to create custom roles. A description of what specific rights can be combined to create a useful custom role is outside the scope of this document.

### 2.3.2 Configuring Identity Sources

System administrators and Organization users have the option of importing users from an LDAPv3 directory or using an external Identity Provider (IdP), such as SafeNet. It is important to understand the differences between the user types. Local users are stored in the vCloud Director for Service Providers Oracle database with the hashes of their passwords. Those users are authenticated against their hashed entry in the database. Limited management functionality is available—database users can only be enabled or disabled and assigned to roles. A six-character password minimum is enforced for local database user accounts.

System LDAP users are stored in the LDAPv3 directory configured at the system level, and references to them are imported into the vCloud Director for Service Providers database where roles are assigned. LDAP users’ passwords are managed and maintained in the LDAPv3 directory, and authentication occurs against that directory using the settings specified in the LDAP configuration screen. All of the LDAPv3 directory’s controls for authentication and passwords are preserved, including authentication failure lockouts, password expiration, history, complexity, and so on, and are specific to the directory chosen. If an organization is configured to use the system LDAP, its users come from the OU specifically configured in that organization’s vCloud Director for Service Providers System LDAP Service settings.

Organization, or custom, LDAPv3 directories are unique to a specific organization and can be hosted in the cloud provider’s infrastructure or at the organization’s enterprise data center. A reference to these users is imported into the database as in the previous example.

VMware strongly recommends that the system administrators and organization users come from a directory server. At this time, users in the Oracle database are not managed with the password and authentication controls typically available in a directory, including authentication failure controls, password expiration, complexity and history, and integration with enterprise identity management systems. Creating local users adds additional management tasks when users change job functions or leave the company. Instead of managing those users using a single LDAPv3 directory, you must remember to change roles and disable users immediately upon a change in their roles or their termination. As an alternative, an enterprise identity management system can be extended to manage local users through the vCloud Admin API, the details of which are outside the scope of this document. For more information on the vCloud API, see the *VMware vCloud API Programming Guide.*

Some cloud providers might choose to allow organizations to use an OU within the system LDAP or to host the organization’s LDAP directory. In either case, appropriate management access to that directory must be provided so that users can be managed by the organization administrator. The lack of such control would provide an extra burden on the system administrator and hinder the organization from easily and properly controlling access to their virtual data centers. In the absence of such management controls, an organization must use only a private LDAPv3 directory that they themselves host and manage.

Finally, the VMware vCloud Director for Service Providers documentation recommends having at least one local user for the system and each organization so that if the LDAPv3 server is not accessible, administrators can still access the system. Organizations must carefully weigh the benefits and risks of such an approach. As mentioned earlier, local users do not have password strength or authentication lockout controls. Those users are thus open, if accessible from the Internet, to attacks. The cloud provider must carefully control which source IP addresses can authenticate to an organization’s cloud URL if local users are configured. Another alternative is the use of the system LDAP or locally-hosted LDAP system. If high availability is a concern, an alternative is the use of a replicated LDAP server, fronted by a load balancer, so as to reduce the possibility of the directory becoming unreachable.

### 2.3.3 Naming Service (LDAPv3) Connectivity

Connectivity from the vCloud Director for Service Provider cells to the system LDAP server and any Organization LDAP servers must be enabled for the software to properly authenticate users. VMware recommends that the system LDAPv3 server be located on the private management network, separated from the DMZ by a firewall. Some cloud providers and most IT organizations will run any Organization LDAP servers required, and those too would be on a private network, not the DMZ. Another option for an organization LDAP server is to have it hosted and managed outside of the cloud provider's environment and under the control of the organization. In that case, it must be exposed to the vCloud Director for Service Provider cells, potentially through the enterprise data center's own DMZ.

In all of these circumstances, opening the appropriate ports through the various firewalls in the path between the cells and the LDAPv3 server is required. By default, this port is 389/TCP for LDAP and 636/TCP for LDAPS. However, this port is customizable with most servers and in the LDAP settings in the Web UI. Also, a concern that arises when the organization is hosting their own LDAPv3 server is exposing it through their DMZ. It is not a service that needs to be accessible to the general public, so steps must be taken to limit access to only the VMware vCloud Director for Service Provider cells. One simple way to do this is to configure the LDAPv3 server and/or the external firewall to only allow access from IP addresses that belong to the vCloud Director for Service Provider cells as reported by the cloud provider. Other options include per-Organization site-to-site VPNs connecting those two sets of systems, hardened LDAP proxies or virtual directories, and so on.

Conversely, cloud providers must beware that organization-hosted LDAP servers managed by unscrupulous customers could be used as part of an attack against other Organizations. For example, one might conceive of an organization requesting an Organization name that is a common misspelling of another organization's name and using the similar-looking login URL in a phishing attack. The provider can take steps to protect against this and similar attacks by both limiting the source IP addresses of requests when possible to avoid inter-organization login attempts as well as ensuring that organization names it assigns are never too similar to one another.

### 2.3.4 Importing Groups

The purpose of importing groups into vCloud Director for Service Providers is to allow you to avoid manually importing individual users with the same role. When LDAPv3 users log in, their session is assigned the roles that are mapped to the groups of which they are members. As users' group memberships change based on updates to their duties within their organizations, the roles assigned to those users change automatically based on the group to role mapping. This allows organizations to easily integrate cloud roles with internal Organization groups/roles and the systems that provision and manage them.

As an example, an Organization might decide to initially grant LDAPv3 users only the "Console Access Only" role to limit users' rights. To do so, all users that need this basic role are added to a single LDAPv3 group, and when that group is imported, the organization administrator assigns it the Console Access Only role. Then, those users who are required to perform additional job duties can be added to other LDAPv3 groups, also imported to vCloud Director for Service Providers, and assigned to these more privileged roles. For instance, users with a need to create catalogs could be added to the "Cloud A Catalog Author" group in the organization's LDAP server. Then the organization administrator can import the Cloud A Catalog Author group and map it to the predefined Catalog author role in vCloud Director for Service Providers.

For more information on available roles, see the "Predefined Roles and Their Rights" section in the *VMware vCloud Director Administrator's Guide* at http://pubs.vmware.com/vcd-56/index.jsp#com.vmware.vcloud.admin.doc_56/GUID-BC504F6B-3D38-4F25-AACF-ED584063754F.html.

## 2.4 Metering and Service Reporting

For service providers running VMware Powered Public Cloud environments, resource metering and service reporting are essential for calculating service costs that can be charged to their end customers through their billing mechanisms. It is extremely important to be able to report on metrics, such as compute usage and allocation, storage usage and allocation, network usage, and one-off costs, such as management and operations, where applicable.

The following table provides examples of workload virtual machine sizing and costing. The provider can create any combination of sizing and specification or alternatively allow the end user to configure explicit specifications.

**Table 6. Workload Virtual Machine Sizing and Cost Examples**

| VM Type | Sizing | Storage | Core Cost Model | Fixed Costs |
|---|---|---|---|---|
| Extra Large | 8 vCPU, 8 GB RAM | 400 GB | Provision cost ($) | Operate cost ($) |
| Large | 4 vCPU, 8 GB RAM | 200 GB | Provision cost ($) | Operate cost ($) |
| Medium | 2 vCPU, 2 GB RAM | 60 GB | Provision cost ($) | Operate cost ($) |
| Small | 1 vCPU, 1 GB RAM | 30 GB | Provision cost ($) | Operate cost ($) |
| Extra Small | 1 vCPU, 500 MB RAM | 15 GB | Provision cost ($) | Operate cost ($) |
| Custom | Any | Any | Provision cost ($) | Operate cost ($) |

## 2.5 Security, Compliance, and Cyber Risk

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

### 2.5.1 PCI DSS

The Payment Card Industry Data Security Standard (PCI DSS) is applicable to all types of environments that store, process, or transmit card holder data. This includes information such as personal account numbers (PANs), as well as any other information that has been defined as card holder data by the PCI DSS v3.1. Cloud computing is no exception to the PCI DSS audit process, and many of the cloud's advantages over earlier models, such as sharing of resources, workload mobility, consolidated management plane, and so on require that adequate controls are adopted to help meet the PCI DSS audit. PCI considerations are essential for assessors to help to understand what they might need to know about an environment to determine whether a PCI DSS requirement has been met. If payment card data is stored, processed, or transmitted in a cloud environment, PCI DSS applies to that environment, and typically involves validation of both the infrastructure and the applications running in that environment.

As an example, many enterprise computing environments in various vertical industries are subject to PCI DSS compliance, and typically those that deal in any kind of financial transaction for exchanging goods and services rely on VMware and VMware technology partner solutions to deliver those computing environments. As such, these customers seek ways to reduce overall IT budget while maintaining an appropriate risk posture for the in-scope environment. One of the greatest challenges in hosting the next-generation cloud computing environment is consolidating the many required modes of trust, such as those for a cardholder data environment (CDE) and a non-cardholder data environment.

For these reasons, VMware has enlisted its Audit Partners, such as Coalfire, a PCI DSS-approved Qualified Security Assessor, to engage in a programmatic approach to evaluate VMware products and solutions for PCI DSS control capabilities and to document these capabilities in a set of reference architecture documents. The first of these documents is this Product Applicability Guide, which contains a mapping of the VMware products and features to be considered for implementing PCI DSS controls. The next two documents that, together with this guide, comprise the PCI DSS Reference Architecture are the Architecture Design Guide and the Validated Reference Architecture, which provide guidance on the considerations to be made when designing a VMware cloud environment for PCI DSS.

### 2.5.2 HIPAA/HITECH

Information security design and architectural requirements, driven by regulatory compliance, are common but critical, aspects that organizations must consider when migrating from traditional IT environments to cloud computing environments. Helping organizations with the arduous tasks of meeting and maintaining HIPAA and the HITECH act regulatory compliance, VMware and its partners provide suites of industry-leading, virtualization solutions that address the confidentiality, integrity, and availability requirements of HIPAA/HITECH.

### 2.5.3 FedRAMP

The Federal Risk Authorization and Management Program (FedRAMP) was created to provide a streamlined and standardized process along with a “do once, use many times” approach to the authorization of commercial cloud services.

This program enables US Government agencies to take full advantage of the benefits of migrating their IT assets and infrastructure to the cloud, as they work to meet the goals of the Federal Cloud Computing Strategy published by the White House in February 2011.

The FedRAMP program provides an avenue for cloud service providers (CSPs) to obtain a provisional Authorization to Operate (p-ATO) after undergoing an independent third-party security assessment that has been reviewed by the JAB. By assessing security controls on candidate platforms, and providing p-ATOs on platforms that have acceptable risk, FedRAMP significantly reduces the time and cost to agencies by removing the assessment and authorization requirements of the underlying cloud vendor services on a system-by-system basis. This minimizes the work each consumer of FedRAMP cloud resources must undergo to receive an actual ATO for the workloads running applications that process sensitive data and transactions.

### 2.5.4 CJIS Compliance and Cyber Risk

The Federal Bureau of Investigation (FBI) established the Criminal Justice Information Services (CJIS) Division in 1992 to meet the need for criminal justice information to be available 24/7 in order for law enforcement, national security, and the intelligence community partners to protect the United States while preserving civil liberties.

Today, CJIS is FBI's largest division and processes millions of transactions on a daily basis, with response times ranging from minutes to seconds. The CJIS Division is responsible for many information technology-based systems like the National Crime Information Center (NCIC), National Instant Criminal Background Check System (NICS), Interstate Identification Index (III), National Data Exchange (N-DEx), Uniform Crime Reporting (UCR) Program, and the Next Generation Identification (NGI). These systems provide state, local, and federal law enforcement and criminal justice agencies with timely and secure access to critical, personal information such as fingerprint records, criminal histories, and sex offender registrations.

### 2.5.5 Compliance Definition

Transparency enables VMware Powered Public Cloud 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 cloud consumers a necessary component of compliance by limiting access, based on a particular role and business need. Common auditor concerns include who can access, configure, and modify a cloud environment; what firewall ports are open; when to apply patches; and where the data resides. Cloud consumers—especially enterprise subscribers—believe that you can outsource responsibility, but you cannot outsource accountability. As evidenced in the PCI Security Standards Council *Assessor Update: July 2011*, active Qualified Security Assessors (QSAs) 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, VMware Powered Public Cloud services can 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 public cloud environments.

For these reasons, VMware has enlisted its audit partners such as Coalfire, a PCI DSS-approved Qualified Security Assessor, to engage in a programmatic approach to evaluate VMware products and solutions for PCI DSS control capabilities and to document these capabilities in a set of reference architecture documents.

### 2.5.6 Compliance Controls

For enterprise subscribers to feel secure and safe in the public cloud services domain and to have the information and visibility into the service for their internal audit requirements, providers of public cloud 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 in 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. VMware Powered Public Cloud providers must make compliance certification types and status available so that subscribers understand which standards both the hosting environment and the services have been audited against.

### 2.5.7 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 VMware Powered Public Cloud 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 VMware Powered Public Cloud service architecture for subscribers. To accomplish this, collect and maintain logs for periods of 6 and 12 months for relevant components of the hybrid cloud service and provide pertinent logs back to individual cloud subscribers on an as-needed basis.

Also, maintain and archive logs for the underlying multi-tenant hosting infrastructure, based on the same 6-month and 12-month periods. In the event of an audit, service providers must be able and willing to provide these logs to an auditor and/or individual subscriber.

In general, cloud service providers typically have logs covering the following components of a subscriber's environment and keep them readily available for subscriber access for periods of up to 6 and 12 months:

- VMware vCloud Director
- NSX for vSphere

vCloud Suite and vRealize Suite are based on a set of products that have been used in many secure environments. Products such as vCloud Director and NSX for vSphere 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. VMware vRealize Log Insight™ can be leveraged to support log generation and reporting for customer requirements.

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

### 2.5.8 Compliant and Secure Architecture

All VMware Powered Public Cloud services offer a secure platform. VMware vSphere, a core building block, offers a secure virtualization platform with EAL2+ and FISMA certifications. vCloud Director, a hybrid cloud delivery platform, offers secure multi-tenancy and organization isolation. The vCloud Suite enables service providers to exercise the defense-in-depth security best practice. The platform offers both per-organization edge firewalls and per-VM firewalls with the inclusion of VMware NSX, 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.

### 2.5.9 Auditing and Logging Compliance

#### 2.5.9.1 Introduction

Recording and monitoring the activities of users is an important part of overall system security. Most organizations have rules governing who is allowed to access and make changes to software and related hardware resources. Maintaining an audit log of significant activities enables the organization to verify compliance with rules, detect any violations, and initiate remediation activities. Some businesses are under external laws and regulations that require ongoing monitoring and verification of access and authorization rules.

An audit log can also be helpful in detecting attempts, whether successful or not, to gain illegitimate access to the system, probe its information, or disrupt its operation. Knowing an attack is attempted and the details of the attempt can help in mitigating the damage and preventing future attacks.

Whether or not it is required, it is part of good security practice to regularly examine logs for suspicious, unusual, or unauthorized activity. Routine log analysis also helps identify system misconfigurations and failures and help to verify adherence to SLAs.

The system audit log is maintained in the database and can be monitored through the vCloud Director for Service Providers web UI. Each organization administrator and the system administrator have a view into the log scoped to their specific area of control. A more comprehensive view of the audit log (and longterm persistence) is achieved through the use of remote syslog, described in following section. A variety of log management and Security Information and Event Management (SIEM) systems are available from a variety of vendors and open-source projects.

Diagnostic logs, described in the following section, contain information about system operation not defined as “audit events” and are stored as files in the local filesystem of each cell’s operating system.

#### 2.5.9.2 Logging in vCloud Director for Service Providers

vCloud Director for Service Providers includes two types of logs:

- Diagnostic logs that are maintained in each cell’s log directory
- Audit logs that are maintained in the database, and optionally, in a syslog server

Diagnostic logs can be useful for problem resolution but are not intended to preserve an audit trail of significant system interactions. Each vCloud Director for Service Providers cell creates several diagnostic log files described in the “Viewing the vCloud Director for Service Providers Logs” section of the VMware vCloud Director for Service Providers Administrator’s Guide.

The audit logs, on the other hand, do record significant actions, including login and logout. As detailed in the VMware vCloud Director for Service Providers Installation Guide, a syslog server can be set up during installation. Exporting logs to a syslog server is recommended for multiple reasons:

- Database logs are not retained after 90 days, while logs transmitted through syslog can be retained as long as desired.
- It allows audit logs from all cells to be viewed together in a central location at the same time.
- It protects the audit logs from loss on the local system due to failure, a lack of disk space, compromise, and so on.
- It supports forensics operations in the face of problems like those listed in the previous bullet.
- It is the method by which many log management and SIEM systems will integrate with vCloud Director for Service Providers. This allows:
    - Correlation of events and activities across vCloud Director for Service Providers, NSX for vSphere, VMware vCloud Networking and Security™, vSphere platform, and even the physical hardware layers of the stack.
    - Integration of cloud security operations with the rest of the cloud provider’s or enterprise’s security operations, cutting across physical, virtual, and cloud infrastructures.
- Logging to a remote system other than the system the cell is deployed on inhibits tampering with the logs. A compromise of the cell does not necessarily enable access to or alteration of the audit log information.

If you did not set up a syslog destination for logging at initial install time, you can configure it later by going to each cell, editing the `$VCLOUD_HOME/etc/global.properties` file, and restarting the cell.

The appropriate ports (514/UDP) must also be open from the vCloud Director for Service Providers host to the syslog server and properly configure the syslog server (which may be part of a larger log management or SIEM solution). The syslog server configuration details are system specific and outside the scope of this document. VMware recommends that the syslog server be configured with redundancy so that essential events are always logged.

This discussion covers only sending the audit log to a syslog server. Security Operations and IT Operations organizations might also benefit from the centralized aggregation and management of the diagnostic logs. There are a variety of methods for collecting those logs, including scheduling a job to periodically copy the logs to a centralized location, setting an additional logger in the `log4j.properties` file (`$VCLOUD_HOME/etc/log4j.properties`) to a central syslog server, or using a log-collection utility to monitor and copy the log files to a centralized location. The configuration of these options is dependent on which system you prefer to use in your environment and is outside the scope of this document.

#### 2.5.9.3 Diagnostic Logging and Log Rollover

The Jetty request log file (`$VCLOUD_HOME/logs/yyyy_mm_dd.request.log`) is programmatically controlled by the Jetty (HTTP) server, but does not come with a maximum size limit. For this reason, there is a risk of unbounded log file growth. A log entry is added to the current file for each HTTP request served by Jetty. VMware recommends that you use log rotate or similar methods to control the size of logs and the number of old log files to keep.

The other diagnostic log files are limited to 400 MB total. Verify that you have sufficient free disk space to accommodate those files plus the size that you allow the Jetty request logs to consume. Centralized logging makes sure that you do not lose valuable diagnostic information as the 400 MB log file total is reached and files are rotated and deleted.

## 2.6 Capacity Distribution and Allocation Models

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

- Virtual Private Cloud (VPC) on-demand – No resources are allocated up-front. Resources are reserved on-demand per workload. Aligning to vCloud Director for Service Provider Pay as you go allocation model.
- VPC – A percentage of resources are reserved with over-commitment from a shared resource pool. Aligning with vCloud Director for Service Providers Allocation Model.
- Dedicated cloud – All resources are reserved up-front regardless of utilization. Aligning with vCloud Director for Service Providers Reservation model.

To determine the appropriate standard units of resource consumption, the VMware Powered Public Cloud 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. VMware vRealize Operations can help the cloud provider’s operations teams accurately forecast capacity usage and trends.

The example in the following table distributes capacity based on 50 percent of the virtual machines for the dedicated allocation model and 50 percent of the virtual machines for the on-demand model. The reservation pool model is applied to small, medium, and large pools, with a respective split of 75 percent, 20 percent, and 5 percent.

Therefore, small represents 37.5 percent of the total, medium represents 10 percent of the total, and large represents 2.5 percent of the total number of virtual machines in the environment.   The table lists the virtual machine count for the various resource pools supporting the two example allocation models for the virtual data centers.

**Table 7. Example Definition of Resource Pool and Virtual Machine Split**

| Type of Resource Pool | Total Percentage | Total Virtual Machines |
|---|---|---|
| On-demand | 50% | 750 |
| Small dedicated pool | 37.5% | 563 |
| Medium dedicated pool | 10% | 150 |
| Large dedicated 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 (1 GB, 1 vCPU, 30 GB of storage)
- 35% medium virtual machines (2 GB, 2 vCPU, 40 GB of storage)
- 15% large virtual machines (4 GB, 4 vCPU, 50 GB of storage)
- 5% extra-large virtual machines (8+ GB, 8+ vCPU, 60 GB of storage)

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

**Table 8. Workload Virtual Machine Sizing and Utilization Examples**

| Virtual Machine Type | Sizing | CPU Utilization | Memory Utilization |
|---|---|---|---|
| Extra Small | 1 vCPU, 500 MB RAM | 5-10% average | Low (5–50%) |
| Small | 1 vCPU, 1 GB RAM | 10-15% average | Low (10–50%) |
| Medium | 2 vCPU, 2 GB RAM | 20-50% average | Moderate (50–75%) |
| Large | 4 vCPU, 4 GB RAM | >50% average | High (more than 90%) |
| Extra Large | 8 vCPU, 8 GB RAM | >50% average | High (more than 90%) |

## 2.7 Service Catalog

Supply a list of suggested services and applications that the VMware Powered Public Cloud service will provide to the consumers. The goal is to help consumers accelerate the adoption of the VMware Powered Public Cloud service. The application and service templates provided to the consumers can be compliant based on the security policies and must take into consideration license subscription.

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 an on-demand consumption model. Cloud native applications also fall in to this space when developers require infrastructure for short periods as they move through their software development lifecycle (SDLC).
- 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 a VPC 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 dedicated consumption model.

The following table lists the types of applications in a service catalog.

**Table 9. Service and Application Catalog Example**

| Service Type | Service Description |
|---|---|
| Operating systems | Microsoft Windows server<br>RHEL<br>CentOS<br>SUSE Linux Enterprise Server (SLES)<br>Ubuntu server |
| Infrastructure services | Databases<br>Microsoft SQL<br>Oracle databases<br>MySQL<br>Distributed data management: VMware vFabric® GemFire®<br>Web application servers<br>Microsoft IIS<br>VMware vFabric tc Server<br>Apache Tomcat<br>IBM WebSphere application server<br>Tiered applications:<br>2 / 3 tier applications (web, application, database with networking and security)<br>Networking services<br>Edge routers, NAT, FW, routing, and so on |

| Service Type | Service Description |
|---|---|
| | Load-balancing services (VMware NSX or third party)<br>Desktops<br>Horizon DaaS<br>DR services<br>DRaaS |
| Application frameworks | Tomcat/Spring<br>JBoss<br>Cloudera/Hadoop |
| Business applications | Microsoft SharePoint<br>Microsoft Exchange |
| Professional Services | Architecture and Design<br>Configuration Services<br>Integration Services |

## 2.8 Service Continuity and Recoverability

### 2.8.1 Data Protection

The VMware Powered Public Cloud service provides data protection that offers a backup and recovery solution to the virtual machines at image level through either VMware vSphere Storage APIs – Data Protection™, integrated third-party services, or VMware vSphere Data Protection Advanced.

### 2.8.2 Disaster Recovery

Disaster recovery must be available to the consumer of the cloud service as an optional feature. This allows end users to leverage the cloud services as disaster recovery targets for their own data centers.

## 2.9 Service Migration and Mobility

VMware Powered Public Cloud VMware Cloud Providers have two options for workload migration and mobility— hybrid mobility through VMware vCloud Connector®, and Offline Data Transfer.

### 2.9.1 Offline Data Transfer Services

Offline Data Transfer is an optional data migration service for transferring large numbers of virtual machines, VMware vSphere vApps™, or templates from your local private vSphere or vCloud Director environments to your VMware Powered Public Cloud VMware Cloud Provider environments. vCloud Connector is used to invoke the service.

As part of this service, a VMware Cloud Provider does the following:

- Ships a physical storage device to the consumer to load VMs, vApps, or templates onto that device and ships it back to the VMware Cloud Provider using a preferred carrier. The content that you load to the device is encrypted by vCloud Connector. The decryption key is stored in the VMware Cloud Provider's vCloud Connector infrastructure, thereby providing security of consumer content during transfer.
- Transfers the data from the device to the appropriate service-offering instance.

### 2.9.2 vCloud Connector

vCloud Connector supports migration of VMs, vApps, and templates between the VMware vCloud Air IaaS and other vSphere or vCloud Director environments, such as data centers or vCloud Air evaluation environments. Export, transport, and import can use vCloud Connector or Open Virtual Machine Format (OVF).

These migration capabilities support onboarding to the VMware Cloud Provider Program IaaS, export from the service offering, and synchronization of templates between the VMware Cloud Provider Program IaaS and on-premises data centers.

In addition to the basic network-based copy operation of VMs, vApps, and templates between vSphere, vCloud Director, the VMware Cloud Provider, and vCloud Air IaaS, vCloud Connector also supports the following use cases:

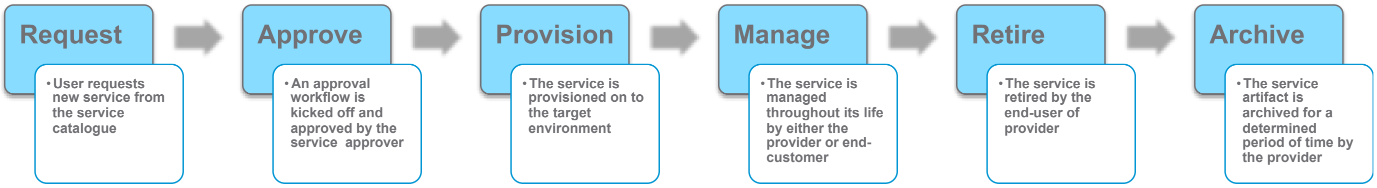
- Extend a single Layer 2 network from your private vSphere and vCloud Director environments to a VMware Cloud Provider Program vCloud Director based cloud so you can migrate VMs or vApps to the cloud while retaining the same IP and MAC address. This allows the VMs or vApps to communicate with other virtual machines or vApps in the private vSphere or vCloud Director environments.
- Synchronize your VMware Cloud Provider catalog with your private vSphere folder or vCloud Director catalog so that all authorized users of your private vSphere or vCloud Director and the VMware Cloud Provider use the same templates.

## 2.10 Service Lifecycle

As a cloud service provider, it is important to understand the service lifecycle—from request to provision, manage, retire, and archive—and the implications this has on the design of the cloud service and integrations with systems such as billing, CMDB, and operations management, where applicable.

The following figure provides an example of a service lifecycle.

**Figure 5. Example Service Lifecycle**


Request
• User requests new service from the service catalogue
Approve
• An approval workflow is kicked off and approved by the service approver
Provision
• The service is provisioned on to the target environment
Manage
• The service is managed throughout its life by either the provider or end-customer
Retire
• The service is retired by the end-user of provider
Archive
• The service artifact is archived for a determined period of time by the provider

## 2.11 Interoperability and Integration

Interoperability aspects of the service definition must 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 VMware Powered Public Cloud instances built to the vCloud API standards might be required.

Integration with external systems is typically achieved through vRealize Orchestrator. vRealize Orchestrator can integrate with many different external systems either through a supported plug-in or API interfaces. Some of the most common integration points are:

- IPAM services
- CMDB services
- Service management tools (ServiceNow, BMC Remedy, and so on)
- Patching services
- Monitoring services

## 2.12 Service Level Agreements

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

General guidelines for VMware public cloud 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 defines responsibilities and limitations associated with the services, including:

- 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

Detailed guidance on how to calculate the level of availability and performance for all VMware Powered Public Cloud service elements is beyond the scope of this document. However, it is anticipated that service providers have an SLA framework in place that can be leveraged or augmented to support cloud services. SLA guarantees must extend to all facets of a provider’s cloud hosting infrastructure and individual service domains (for example, compute, network, storage, Layer 4–7 services, and management/control plane) that directly support VMware Powered Public Cloud services. Adherence to SLA requirements must also factor in the resiliency of the management framework, which provides API and user interface accessibility for service subscribers.

# VMware Powered Public Cloud Service Examples

## 3.1 Virtual Private Cloud On-Demand Offering

The Virtual Private Cloud on-demand 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 multi-tenant 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 data center 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.

### 3.1.1 Service Design Parameters

As part of the design process for the Virtual Private Cloud on-demand service offering, 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 data center. 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. The following table provides an example of these key service settings, values, and justifications.

**Table 10. Resource Allocation Settings Example – VPC On-Demand Service Offering**

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

In this example, the minimum vCPU speed setting is configured at 1 GHz (1000 MHz), with a memory resource guarantee of 75 percent. 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 over-commitment from aggressive levels (for example, resource guarantees set to less than100 percent) to more conservative levels (for example, resource guarantees always set to 100 percent), depending on SLAs in place or fluctuating service loads.

### 3.1.2 Resource Allocation and Catalogs

The VPC on-demand 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.

The following table provides an example of different sizing combinations that can be included in a vApp catalog with the basic service offering.

**Note** In the following table, vCPU quantity is based on a multiple of 1 GHz. Any quantity of memory or vRAM assigned from Table 12 is reserved at 75 percent. The provider can give subscribers the capability to select specific quantities of resources, such as vCPU, memory, and storage for a given virtual machine or vApp dynamically, as needed. However, providers must first implement a pricing model commensurate with the range of scale for each resource type.

**Table 11. VPC On-Demand Service Offering Catalog Example**

| Instance Size | vCPU/GHz | Memory (MB) | Storage (GB) | Bandwidth (MBps) | Cost |
|---|---|---|---|---|---|
| Extra Small | 1.0/1 GHz | 500–100,000 | 10–2,000 | Variable | Set by provider |
| Small | 1.0/1 GHz | 500–100,000 | 10–2,000 | Variable | Set by provider |
| Medium | 2.0/2 GHz | 500–100,000 | 10–2,000 | Variable | Set by provider |
| Large | 4.0/4 GHz | 500–100,000 | 10–2,000 | Variable | Set by provider |
| Extra-Large | 8.0/8 GHz | 500–100,000 | 10–2,000 | Variable | Set by 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 12 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.

### 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's virtual data center. 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 sub-hourly basis. If subscribers opt to change the size of their virtual machine or vApp instances after initial setup, the 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 might 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.

The following table 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 examples are meant only to be illustrative. Providers must rely on their own internal cost models and metering schemes for billing or showback.

**Table 12. 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.