



VMware vCloud[®] Architecture Toolkit

Private 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
2.	Service Definition	7
2.1	Service Definition Approach.....	7
2.2	Service Concepts and Terminology	7
2.3	Service Lifecycle	8
2.4	Service Quality	8
2.5	Service Characteristics	9
2.6	Service Objectives	9
2.7	Business Benefits.....	10
2.8	Stakeholders	10
2.9	User Roles	11
2.10	Use Cases.....	12
2.11	Service Offerings.....	13
2.12	Consumer Capabilities	15
2.13	Service Metering	15
2.14	vApp Catalog.....	16
2.15	Capacity Distribution	17
2.16	Service Level Agreement.....	18

List of Figures

Figure 1. Deployment Models	5
Figure 2. Service Models	6
Figure 3. Service Characteristics	9

List of Tables

Table 1. User Roles and Rights Example	11
Table 2. Workload Examples	12
Table 3. Service Offerings.....	13
Table 4. Example Committed Service Offering Definitions	14
Table 5. Workload Virtual Machine Sizing and Costing Examples	16
Table 6. Workload Virtual Machine Sizing and Utilization Examples	18

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. Cloud computing helps transform IT from a cost center into a service provider.

VMware vCloud® is the VMware solution for cloud computing. This document provides the service definition for an enterprise that provides private Infrastructure as a Service (IaaS) resources and uses public vCloud resources to extend its own private capacity.

This document has two primary goals:

- Provide an approach for creating a service definition.
- Provide a sample service definition for an enterprise vCloud that can be used as a starting point to create a customized service definition that meets specific business objectives.

1.1 Audience

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

- Enterprise IT as a provider of private vCloud services and a consumer 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 enterprise vCloud.

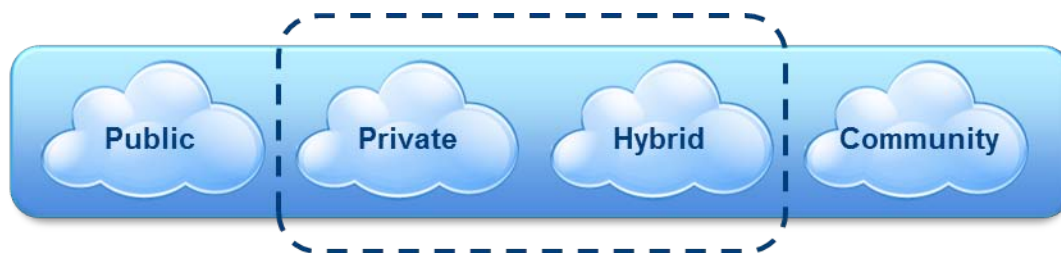
1.2 Scope

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

1.2.1 Deployment Model Scope

Figure 1 illustrates deployment models for cloud computing. For enterprises, the focus is on private and hybrid clouds.

Figure 1. Deployment Models



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

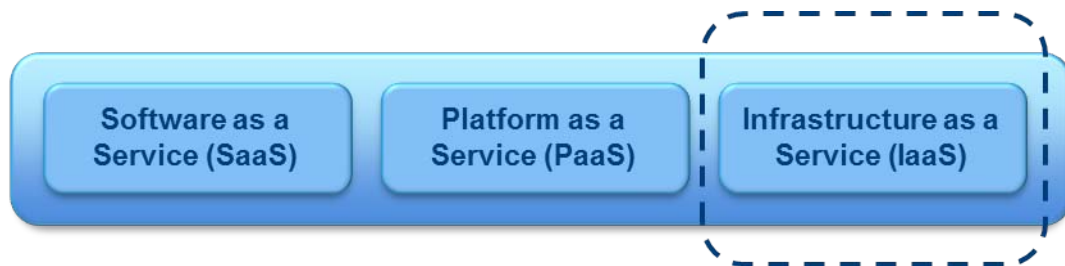
- *Private Cloud* – The cloud infrastructure is operated solely for an organization and may be managed by the organization or a third party. The cloud infrastructure may be on-premise or off-premise.
- *Public Cloud* – The cloud infrastructure is made available to the general public or to a large industry 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 an enterprise IT organization as a provider of services to the consumers of enterprise vCloud services—an enterprise private vCloud. The service definition also covers enterprise IT as a user or consumer of public vCloud services in order to extend its own private capacity—an enterprise hybrid vCloud.

1.2.2 Service Model Scope

Figure 2 shows the generally accepted service models for cloud computing.

Figure 2. Service Models



- Software as a Service (SaaS) – Business-focused software application 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 the application developers from a service catalog.
- Infrastructure as a Service (IaaS) – Infrastructure *containers* are presented to consumers in order to provide more agility, automation, and delivery of components.

The service model for the service definition in this document is for IaaS—for an enterprise IT organization to provide Infrastructure as a Service via a catalog of predefined infrastructure containers to consumers of these vCloud services. The IaaS service layer serves as a foundation for providing additional service offerings, such as PaaS, SaaS, Desktop as a Service, and so on.

2. Service Definition

To define an enterprise vCloud service, clear, complete, and concise terms of the service need to be defined. The following provides an approach for creating a service definition, and a sample of an enterprise vCloud service definition that can be used to create a customized service definition that meets specific business objectives.

2.1 Service Definition Approach

The approach for defining and designing the services offered by the enterprise vCloud should include:

- Involvement of all necessary stakeholders.
- Documenting the business drivers and requirements that can be translated into appropriate service definitions.
- Considering a holistic view of the entire service environment and service lifecycle including service setup, service request, service provisioning, service consumption, service management and operations, and service transition and termination.
- Defining the service scenarios and use cases.
- Representation of the service in order to understand the components of the service, interactions, and sequence of interrelated actions.
- Defining the users and roles involved with or interacting with the services so that user-centric services are created.
- Defining the service contract (Service Level Agreement or SLA) for the services and service components in the areas of infrastructure services, application/vApp services, platform services, software services, and business services. Service quality should be defined for the areas of performance, availability, continuity, scalability, manageability, security, compliance, and cost/pricing.
- Defining the business service catalog and supporting IT service catalog.

2.2 Service Concepts and Terminology

The following are the service concepts and terms used:

- *Service* – A means of delivering value to consumers by facilitating outcomes that consumers want to achieve without the ownership of specific costs or risks.
- *vCloud Service Provider (or Provider)* – An entity that provides vCloud services to consumers.
- *Consumer or Customer* – Someone who consumes vCloud services and defines or agrees to Service Level Targets.
- *Service Level Target* – A commitment that is documented in a Service Level Agreement. Service Level Targets are based on Service Level Requirements, and are needed to make sure that the vCloud service design is fit for its purpose. Service Level Targets should be *SMART* (Specific, Measurable, Actionable, Realistic, Time-bound), and are usually based on Key Performance Indicators (KPIs).

- *Service Level Agreement (SLA)* – An agreement between a service consumer and the service provider that measures the quality and performance of the available services.
- *Service Level Requirement* – A document recording the business requirements of a vCloud service.
- *Service Level Objective* – A negotiated document that defines the service that will be delivered to the consumer in qualitative terms, although a small number of KPIs may also be included. It provides a clearer understanding of the true nature of the service being offered, focusing on the contribution of the service to the business value chain.

2.3 Service Lifecycle

Although it might not be possible to consider every aspect of a service, it is important to keep the entire service environment and service lifecycle in mind when creating the service definition. This includes service setup, service request, service provisioning, service consumption, service management and operations, and service transition and termination. A conscious awareness of what consumers of the service and the provider of the service will experience at each stage of the service lifecycle must be taken into account. This helps create the necessary service definition elements for the consumer-facing (SLA) and internal-facing (operational-level agreement or OLA) criteria.

2.4 Service Quality

Define the following service quality areas for the services, as appropriate:

- Performance (application response time, bandwidth including burst, time to respond, time to resolution).
- Availability (uptime, backup, restore, data retention).
- Continuity (RPO, RTO).
- Scalability (staged burst to public vCloud).
- Manageability (user account management, metering/billing/reporting, supportability).
- Security (application/data access, management/control access, user accounts, authentication/authorization).
- Compliance (regulatory compliance, logging, auditing, data retention, reporting).
- Cost/pricing.

2.5 Service Characteristics

The following are the essential cloud service characteristics as defined by NIST.

Figure 3. Service Characteristics



- On-demand self service – A consumer can unilaterally provision computing capabilities as needed, automatically, without requiring human interaction with each service's provider.
- Broad network access – Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms.
- 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. There is a sense of location independence in that the subscriber generally has no control or knowledge over the exact location of the provided resources, but may be able to specify location at a higher level of abstraction.
- Rapid elasticity – Capabilities can be rapidly and elastically provisioned, in some cases automatically, to scale out quickly and be rapidly released to quickly scale in. 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 use by leveraging a metering capability at some level of abstraction appropriate to the type of service. Resource usage can be monitored, controlled, and reported, providing transparency for both the provider and consumer of the utilized service.

2.6 Service Objectives

The following are examples of service objectives for an enterprise vCloud service:

- Deliver a fully operational private vCloud 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 and application building blocks.
- Provide the following service offerings: Basic (pay for resources used), Committed (allocated resources), and Dedicated (reserved resources).
- Support a minimum of 1500 virtual machines across the three service offerings and have a plan to grow to a minimum of 5000 virtual machines.

- Provide workload mobility between vCloud instances allowing the consumer to easily enter the vCloud and exit the vCloud with existing workloads.
- Provide a direct connection to the external network for applications that have upstream dependencies.
- Provide an isolated network for applications that need to be isolated.
- Provide open, interoperable, and Internet standard protocols to consume vCloud resources.
- Provide for workload redundancy and data protection options.

2.7 Business Benefits

The resulting enterprise vCloud services provide the following benefits:

- Secure multi-tenancy for lines of business (LOB).
- Leveraging existing investments in both private and public clouds.
- Efficient provisioning, sharing, and management of virtualized resources.
- Self-service user portal with approval workflows.
- Standard catalog of business and IT services that can be created, managed, and maintained.
- Standard catalog of predefined virtual machines and applications with usage metering.
- Linking with other vCloud services within or beyond the boundaries of the enterprise.

2.8 Stakeholders

Include all stakeholders who are affected by or have influence over the enterprise vCloud service, including those involved in creating, providing, and consuming the service. This is *co-creation*, and facilitating co-creation in groups that are representative of the stakeholders is a vital aspect of vCloud service definition and design. Co-creation also facilitates a smooth interaction between the stakeholders during the actual vCloud service provisioning. Through co-creation, customers get the chance to add value to the vCloud service in partnership with the enterprise vCloud service provider early in the development of the vCloud service.

Examples of stakeholder categories include:

- Corporate functions:
 - CxO/Executive management
 - Enterprise security
 - Enterprise architecture
 - QA/Standards groups
 - Program management office
- End-User organization/LOB/BU:
 - Executives
 - Line management
 - Business domain experts
 - Data owners

- Enterprise vCloud project organization:
 - Executives
 - Line management
 - Business process/functional experts
 - Product specialist
 - Technical specialist
- Enterprise vCloud Operations:
 - Service management
 - Application management
 - Infrastructure management
- External:
 - Suppliers
 - Regulatory bodies

2.9 User Roles

There are several users and roles that apply to everyone who interacts with an enterprise vCloud service. Several roles are defined in the access model of the enterprise private vCloud service at the provider level and at the consumer level. There are also levels of rights granted to predefined roles that have an important impact on how users interact with the enterprise cloud service.

Table 1 provides a sample of the users and roles required for the enterprise vCloud solution.

Table 1. User Roles and Rights Example

User Role	Needs	Rights
Provider Cloud Administrator	One (minimum).	Highest level enterprise cloud provider administrator; has superuser rights.
Provider Catalog Author	As needed.	Provider user who creates and publishes new catalogs.
Consumer Organization Administrator	One per organization.	Administrator in the organization over systems and users.
Consumer Organization Author	One or more, as needed.	Allows vApp and catalog creation; no infrastructure management.
Consumer Organization User	One or more, as needed.	Allows consumer organization user to use vApps created by others.

2.10 Use Cases

Enterprise vCloud use cases generally fall into the following categories of workloads:

- **Transient** – A transient application is one that is used infrequently, exists for a short time period, or is used for a specific task or need. It is then discarded. This type of workload is appropriate for a pay for use consumption model.
- **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 consumption model.
- **Steady State** – A steady state application is one that tends to run all the time at a predictably steady state. This type of workload is appropriate for a reservation consumption model.

Table 2 lists common use cases that fall into the general vCloud workload cases.

Table 2. Workload Examples

Examples	
Software testing	Software development
Web applications	Messaging and collaboration applications
File and print services	Document management applications
Sales and demonstration	Custom applications
Training	Infrastructure applications
Low governance workloads	High governance workloads

With a knowledge of the workload types and common uses cases you can begin to define the service tiers that will be offered inside of your enterprise. The following are some example use cases:

- Deploy an enterprise private vCloud that deploys self-service and increases the speed with which consumer organizations can bring applications and services to market.
- Customer with a Web presence needs to extend their front-end server farm due to anticipated spikes in incoming traffic.

See the *Private VMware vCloud Implementation Example* and *Hybrid VMware vCloud Use Case* for further details.

2.11 Service Offerings

Service offerings within the enterprise can be an effective method of differentiating types, location, or owners of infrastructure as reflected in an enterprise vCloud. The following service offerings map to the business needs that an enterprise cloud provides to its consumers. A single consumer/Line-of-Business/Business Unit may consume one or more of these three service offerings:

- *Basic Service Offering* – Unreserved pay for use class. Designed for quick start pilot projects or for workloads such as software testing that can be charged based on resources used.
- *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.

Table 3. Service Offerings

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

2.11.1 Basic Service Offering

The Basic service offering is an instance-based pay for use resource consumption model. Each virtual machine provisioned in this virtual datacenter is charged separately and separate billing records are produced for each virtual machine. Consumers using the Basic service tier can be 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, 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 specific to the enterprise vCloud service provider's infrastructure, capabilities, location, and agreement with their consumers.

2.11.2 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.

2.11.2.1. 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. For example, if a customer is allocated 10GHz of CPU resources, a virtual datacenter is created for the customer 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's 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's 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.

Table 4 provides an example of how virtual datacenter sizes are defined for the Committed service offering.

Table 4. Example Committed Service Offering Definitions

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

2.11.3 Dedicated Service Offering

The Dedicated service offering uses the reservation consumption-based model. An enterprise consumer/LOB/BU works with the enterprise cloud provider to provision a cluster of servers that's dedicated to this consumer. The hardware (network, storage, servers) is not shared with other consumers. The consumer gets full control over the reservation and limit of this set of resources. This service offering will likely be a fixed price monthly subscription.

2.12 Consumer Capabilities

The following is an example of the capabilities consumers of the enterprise vCloud are able to access for all of the service tier offerings:

- User Interface – A user interface portal is the primary customer interface to the enterprise vCloud service. It provides a consistent interface for interacting with the enterprise vCloud services.
- API – The full end-user vCloud API is implemented and exposed to the consumers to enable programmatic access to the enterprise vCloud services.
- Mobility/Elasticity – Enterprise IT can take advantage of vCloud service provider resources to extend its own private capacity. To the consumer, the capabilities appear to provide additional capacity that can be provisioned in any quantity at any time.
- Catalogs – Consumers can create and publish new catalog of vApp templates, and provision new vApps from catalogs.
- vApps – Consumers (or their administrators) can build new vApps, upload/download vApps, clone/create/delete vApps, and perform other operations to interact with the vApps (power on/off/suspend/resume, remote console, others)
- User Management – Consumers (or their administrators) can create or delete users, manage role assignments, and manage user quotas.
- Networking and Security – Consumers (or their administrators) can add or remove firewall rules, and configure DHCP settings.
- Billing – Consumers can view metering/billing reports from the past 12 months.
- Compliance – Consumers can download relevant log files from the past six months.

2.13 Service Metering

For vCloud environments, resource metering is essential for accurately measuring consumer usage and shaping consumer behavior through chargeback policies. Enterprises do not necessarily have the same cost pressures for an enterprise private vCloud as a public vCloud service provider. The requisite chargeback procedures or policies may not exist. An alternative to chargeback is *showback*, which attempts to raise awareness of the consumption usage and cost without involving format account procedures to bill the usage back to the consumer's department.

Table 5 provides examples of workload virtual machine sizing and costing.

Table 5. Workload Virtual Machine Sizing and Costing Examples

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

2.14 vApp Catalog

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

2.14.1 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

2.14.2 Infrastructure Apps

- Databases:
 - Microsoft SQL Server
 - Oracle Database 11g
 - MYSQL 5.x
- Distributed data management
 - VMware vFabric™ Gemfire®

- Web/App Servers:
 - Microsoft IIS
 - VMware vFabric™ tc Server™
 - Apache Tomcat
 - IBM Websphere Application Server 7
- Simple n-Tier Applications:
 - 2-Tier application with a Web front-end and database backend
 - 3-Tier application with Web, processing and database
 - Enhanced 3-Tier with added monitoring
- Load balancer.

2.14.3 Application Frameworks

- Tomcat/Spring
- JBoss
- Cloudera/Hadoop

2.14.4 Business Applications

- Microsoft SharePoint
- Microsoft Exchange
- VMware Zimbra

2.15 Capacity Distribution

To determine the appropriate standard resource units of consumption, the enterprise vCloud 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.

The following is an example of some of the assumptions that can be used for service capacity planning.

Virtual machines distribution assumption:

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

Table 6 lists some examples of workload virtual machine sizing and utilization.

Table 6. Workload Virtual Machine Sizing and Utilization Examples

Virtual Machine Type	Sizing	CPU Utilization	Memory Utilization
X-Large	8 vCPU x 8GB RAM (can offer up to 32 vCPU and 1TB RAM)	>50% average	High (upwards of 90%)
Large	4 vCPU x 4GB 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%)

2.16 Service Level Agreement

A service level agreement is a negotiated contract between the enterprise vCloud provider and the enterprise consumer for an enterprise private vCloud. Similarly, it is also an agreement between the enterprise vCloud provider (as a consumer of public vCloud resources) and a public vCloud service provider for an enterprise hybrid cloud.

The following is an example of some of the items that can be included in the SLA.

- Application performance – Application response time.
- Network – Bandwidth including burst.
- Support – Time to respond, time to resolution.
- Request fulfillment – Response time for provisioning and configuration requests.
- Availability – Uptime.
- Backups – backup schedule, restore time, data retention.
- Continuity – RPO, RTO.
- Manageability – User account management, metering parameters, reporting details/frequency/history.
- Security – Application/data access, management/control access, user accounts, authentication/authorization.
- Compliance – Regulatory compliance, logging, auditing, data retention, reporting.