VMware vCloud® Architecture Toolkit
Service Definitions

Version 3.0

September 2012

vmware®

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

Contents

List of Figures

List of Tables

1. Introduction

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

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

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

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

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

1.1 Audience

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

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

1.2 Deployment Model

Figure 1 illustrates several deployment models for cloud computing.

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

Figure 1. Deployment Models

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

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

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

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

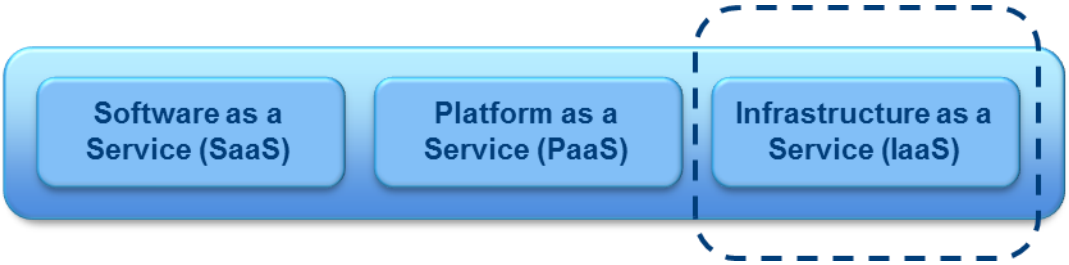
This document does not cover community vCloud service definition considerations or examples.

1.3 Service Model

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

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

Figure 2. Service Models

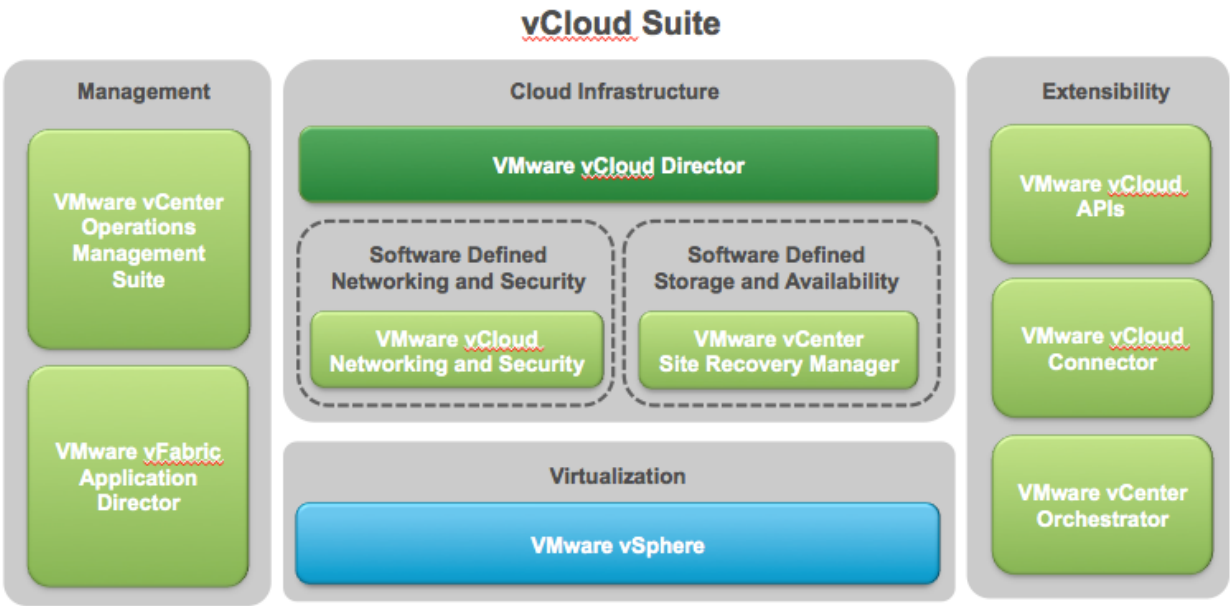


The service model for the service definition in this document is primarily IaaS, for an organization to provide Infrastructure as a Service to consumers of vCloud services through a catalog of predefined infrastructure containers. The IaaS service layer serves as a foundation for additional service offerings,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

1.4 Technology Mapping

vCloud services are delivered by the capabilities of the VMware technologies in the VMware vCloud Suite as shown in Figure 3.

Figure 3. Technology Mapping



1.5 Service Characteristics

The following essential cloud service characteristics are defined by the National Institute of Standards and Technology:

- 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 client or thick client platforms.
- Resource pooling – The provider's computing resources are pooled to serve multiple consumers, using a multitenant model with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. There is a sense of location independence because the subscriber generally has no knowledge of or control 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 provisioned to scale out quickly and to be released rapidly, in some cases automatically. Rapid elasticity allows resources both to scale out and scale in quickly. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be purchased in any quantity at any time.
- Measured service – Cloud systems automatically control and optimize resource usage by leveraging a metering capability at some level of abstraction appropriate to the type of service. Resource usage can be monitored, controlled, and reported, providing transparency for both the provider and consumer of the utilized service.

Figure 4. Service Characteristics

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

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

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

1.6 Service Development Approach

The approach for defining and designing vCloud services should include:

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

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

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

1.7 Concepts and Terminology

The key terms and service concepts are defined as follows:

- *Service* – A means of delivering value to consumers by facilitating outcomes that they want to achieve without the ownership of specific costs or risks.
- *vCloud* – A model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable resources that can be provisioned rapidly and released with minimal management effort.
- *vCloud Service Provider* (or *Provider*) – An entity that provides vCloud services to consumers.
- *Consumer* or *Customer* – Someone who consumes vCloud services and defines or agrees to service-level targets.
- *Service-Level Target* – A commitment that is documented in a service-level agreement. Service-level targets are based on service-level requirements and are needed so that the vCloud service design is fit for its purpose. Service-level targets should be *SMART* (Specific, Measurable, Actionable, Realistic, Time-bound) and are usually based on Key Performance Indicators (KPIs).
- *Service-Level Agreement* (SLA) – An agreement between a service consumer and the service provider that measures the quality and performance of the available services. The SLA is the entire agreement that specifies what service is to be provided, how it is supported, time, locations, cost, performance, and responsibilities of the parties involved.
- *Service-Level Objective* (SLO) – A negotiated document that defines the service to be delivered to the consumer, with one or more key performance indicators (KPIs). It provides a clear understanding of the nature of the service being offered, focusing on the contribution of the service to the business value chain. SLOs are specific, measurable characteristics of the SLA, such as availability, throughput, frequency, response time, or quality.
- *Operational-Level Agreement* (OLA) – An agreement internal to the service provider that details the interdependent relationships among the internal support groups of an organization working to support an SLA.
- *VMware vCloud Suite* – The suite of VMware technologies that provide the solution for vCloud computing.
- *VMware vCloud Services or vCloud Services* – vCloud computing services built with the VMware vCloud Suite.

2. 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 to understand the scope, limitations, and cost of service offerings.

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

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

2.1 Service Objectives

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

- Deliver a fully operational private or public vCloud infrastructure with hybrid capability.
- Provide secure multitenancy for vCloud infrastructure consumers.
- Provide compliance controls and transparency for the service.
- Maintain IT control of access to the system and resources.
- Provide differentiated tiers of scale to align with business needs.
- Allow for metering of the service for cost distribution.
- Establish a catalog of common infrastructure and application building blocks.
- Provide the following service offerings:
    - Basic (pay for resources used).
    - Committed (allocated resources).
    - Dedicated (reserved resources).
- Support a minimum of 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 enter and leave the vCloud easily with existing workloads.
- Provide a direct connection to the external network for applications with upstream dependencies.
- Provide an isolated network for applications that need to be isolated.
- Provide open, interoperable, and Internet-standard protocols for consuming vCloud resources.
- Provide for workload redundancy and data protection options.

2.2 Use Cases

The following use cases represent business problems, some general and some industry-specific, which can be addressed with vCloud services and represented by a service definition.

Table 1. Example: Use Case 1

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

Table 2. Example: Use Case 2

Use Case UC_02	
Name	Increase business capacity and scale rapidly.
Problem Statement	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 in order to support seasonal and periodic business demand.
Requirements/Goal	• Consumers should have access to scale capacity on demand. • IT must be able to scale up, down, in, or out to support business demand. • Scale within a short cycle of days in order to meet projected demand. • Scale to off-premises capacity.
Risks	• Lost revenue due to lack of capacity. • 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	Business is unable to develop new products and services rapidly because IT takes too long to provision development and test infrastructure.
Description	IT needs to be able to provide on demand self-service provisioning of development and test infrastructure to support the business to rapidly develop new products and services.
Requirements/Goal	• Developers and test users have access to a catalog of IT infrastructure that they can rapidly provision and use. • Self-service provisioning, with approvals if necessary. • Reduce time to market for products and services.
Risks	• Products and services are late to market, resulting in lost customers and market share.

Table 4. Example: Use Case 4

Use Case UC_04	
Name	Security and compliance assurance.
Problem Statement	Business is concerned about putting critical financial applications and data on vCloud services.
Description	IT must be able to provide secure business services for financial applications and data, which should have controlled access and be separated from other users of the vCloud services.
Requirements/Goal	• Provide compliance controls and transparency for the service. • Provide network isolation for applications that must be isolated.
Risks	• Security and compliance breach.

Table 5. Example: Use Case 5

Use Case UC_05	
Name	Business market launch.
Problem Statement	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	• Rapid service provisioning to support product and service launch. • Consumers should have access to a catalog of IT infrastructure that they can rapidly provision and use. • Self-service provisioning with approvals if necessary. • Reduce time to market for products and service.
Risks	• Products and services are late to market, resulting in lost customers and market share. • Lost opportunity cost.

2.3 User Roles

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

The following table provides a sample of the users and roles required for the enterprise vCloud solution.

Table 6. User Roles and Rights Example

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

2.4 Metering and Service Reporting

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

The following table provides examples of workload virtual machine sizing and costing.

Table 7. Workload Virtual Machine Sizing and Costing Examples

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

2.5 Security and Compliance

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

2.5.1 Compliance Definition

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

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

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

- Facilitating compliance through ISO 27001 certification and/or SSAE 16, SOC 2 reporting, based on a standard set of controls.

- Providing compliance logging and reports to service subscribers, for full visibility into their hosted vCloud environments.
- Architecting the service so that subscribers can control access to their vCloud environments.

2.5.2 Compliance Controls

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

- ISO 27001 certification, which certifies that security management processes are in place and have a relevant subset of the ISO 27001 controls, as specified by *VMware's 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 third-party audit. vCloud providers should make compliance certification types and status available so that subscribers understand what standards both the hosting environment and the services have been audited against.

2.5.3 Compliance Visibility and Transparency

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

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

- vCloud Director.
- vCloud Networking and Security Edge.

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

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

2.5.4 Compliant and Secure Architecture

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

2.6 Capacity Distribution and Allocation Models

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

- Pay-As-You-Go – No upfront resource allocation, and resources are reserved on demand per workload.
- Allocation Pool – Percentage of resources are reserved with overcommitment.
- Reservation Pool – 100% of resources are reservation-guaranteed.

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

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

The following table lists the virtual machine count for the various resource pools supporting the two example allocation models for the virtual datacenters.

Table 8. Definition of Resource Pool and Virtual Machine Split

Type of Resource Pool	Total Percentage	Total Virtual Machines
Pay-As-You-Go	50%	750
Small Reservation Pool	37.5%	563
Medium Reservation Pool	10%	150
Large Reservation Pool	2.5%	37
TOTAL	100%	1,500

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

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

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

Table 9. Workload Virtual Machine Sizing and Utilization Examples

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

2.7 Applications Catalog

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

Application workloads generally fall into the following categories:

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

Table 10. Applications Catalog Example

Application Type	Application Description
Operating Systems	• Microsoft Windows Server. • RHEL. • Centos. • SUSE Linux Enterprise Server. • Ubuntu Server.

Infrastructure Applications	• Databases.
	  ○ Microsoft SQL Server.
	  ○ Oracle Database.
	  ○ MYSQL.
	• Distributed data management.
	  ○ VMware vFabric™ Gemfire®.
	• Web/application servers.
	  ○ Microsoft IIS.
	  ○ VMware vFabric™ tc Server™.
	  ○ Apache Tomcat.
	  ○ IBM WebsPhere Application Server.
	• Simple n-tier applications.
	  ○ 2-tier application with a web front end and database back end.
	  ○ 3-tier application with web, processing and database.
	  ○ Enhanced 3-tier with added monitoring.
	• Load balancer.

Application Frameworks	• Tomcat/Spring. • JBoss. • Cloudera/Hadoop.
Business Applications	• Microsoft SharePoint. • Microsoft Exchange. • VMware Zimbra.

2.8 Interoperability

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

2.9 Service-Level Agreement

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

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

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

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

SLA guarantees should extend to all facets of a provider's vCloud hosting infrastructure and individual service domains (for example, compute, network, storage, L4-L7 services, and management/control plane) that directly support vCloud services. Adherence to SLA requirements should also factor in the resiliency of the management framework, consisting of API and UI accessibility for service subscribers.

3. Service Offering Examples

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

The following is a summary of these service offerings:

- Basic – Based on the Pay-As-You-Go allocation model. This service offering lends itself to quick-start pilot projects or test and development application workloads that typically do not require persistent resource commitments or upfront resource reservations.
- Committed – Based on the Allocation Pool allocation model. This service offering provides consumers with a minimum initial commitment of resource capacity plus the added ability to burst above that minimum if additional infrastructure capacity is available at the time of need. The level of minimum commitment, expressed as a percentage of overall capacity per resource type, provides an extra layer of assurance to consumers who seek deterministic performance levels for their application workloads.
- Dedicated – Based on the Reservation Pool allocation model. This service offering provides consumers reserved resource capacity upfront, fully dedicated by individual tenant. The level of resource guarantee (always set to 100%) provides customers a higher degree of service assurance than the Committed service offering, plus additional layers of security and resource control for their application workloads.

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

To help decide which service offering makes the most sense for a particular set of business use cases, refer to the key service attributes summarized in Table 11. Additional details and reference examples for each service offering can be found in the following sections.

Table 11. Service Offering Matrix Example

	Basic Service Offering	**Committed Service Offering**	**Dedicated Service Offering**
Allocation Model	Pay-As-You-Go	Allocation Pool	Reservation Pool
Control Plane (Management)	Shared, multitenant	Shared, multitenant	Shared, multitenant
Cluster Resources	Shared, multitenant	Shared, multitenant	Dedicated, single-tenant
Unit of Consumption	vApp	Aggregate resource capacity allocated	Aggregate resource capacity reserved

	Basic Service Offering	**Committed Service Offering**	**Dedicated Service Offering**
Resource Allocation Settings (per Organization Virtual Datacenter)	NA	• CPU (GHz) • Memory (GB) • Storage (GB)	• CPU (GHz) • Memory (GB) • Storage (GB)
Resource Guarantee Settings	• % of CPU • vCPU speed • % of Memory • % of Storage	• % of CPU • % of Memory • % of Storage	• 100% of CPU • 100% of Memory • 100% of Storage
Limits (per Organization Virtual Datacenter)	Maximum number of virtual machines	Maximum number of virtual machines	Maximum number of virtual machines
Reporting/Billing Frequency	Per use	Monthly	Monthly or Annual
Metering Frequency	Hourly	Hourly	Hourly
Service Availability	99.95%	99.99%	99.99%
Target Workloads	Test and Development	Tier 2 and 3 Production	Tier 1 Production
Application Workload Examples	• Short-term or bursty workloads. • QA testing. • Integration testing. • New software version testing. • Short term data analytics.	• Static web content servers. • Lightly used app servers. • Active Directory Servers. • Infrastructure Servers (DNS, print, file). • Small/medium database servers. • Short term content collaboration. • Staging sites.	• Exchange and SharePoint servers. • Large database servers (high IOPS). • PCI related servers. • HPC workloads. • SaaS production applications. • CRM, EDA, ERP, and SCM applications. • Financial applications (high-compliance).

3.1 Service Offering – Basic

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

3.1.1 Service Design Parameters

As part of the design process for the Basic service offering, providers should give special consideration to key service settings and values in vCloud Director that can impact service performance and consistency levels for a subscriber's organization virtual datacenter. Given the Pay-As-You-Go allocation model employed in this service, certain circumstances might arise that result in subscribers overcommitting resources over time. If not properly managed, performance for all application workloads could be negatively affected. The following table provides an example of these key service settings, values, and justifications.

Table 12. Resource Allocation Settings Example – Basic Service Offering

Resource Type	Value Range	Sample Setting	Justification
CPU resources guaranteed	0–100%	0%	The percentage of CPU resources that are guaranteed to a virtual machine running within the target organization virtual datacenter. This option controls over-commitment of CPU resources.
vCPU speed	0–8GHz	1GHz	This value defines what a virtual machine or vApp with one vCPU consumes at maximum when running within the target organization virtual datacenter. A virtual machine with two vCPUs consumes a maximum of twice this value.
Memory resources guaranteed	0–100%	75%	The percentage of memory that is guaranteed to a virtual machine running within in the target virtual datacenter. This option controls overcommitment of memory resources.
Maximum number of virtual machines	1–Unlimited	Unlimited	A safeguard that allows control over the total number of vApps or virtual machines created by a subscriber within the target virtual datacenter

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

3.1.2 Resource Allocation and Catalogs

The Pay-As-You-Go resource allocation model enables providers to deliver high levels of flexibility in the way resources are allocated, through published vApp catalogs in vCloud Director. vApp catalogs further enable providers to publish standard application images and pre-defined 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: vCPU quantity is based on a multiple of 1GHz, as provided in the example in Table 12. Any quantity of memory or vRAM assigned from Table 13 is reserved at 75%. Subscribers' ability to select specific quantities of resources, such as vCPU, memory, and storage for a given virtual machine or vApp dynamically may be governed as necessary by the provider. However, providers should first implement a pricing model commensurate with the range of scale for each resource type.

Table 13. Basic Service Offering Catalog Example

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

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

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

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

The maximum virtual machine instance size is derived from the maximum amount of vCPU and the maximum amount of memory that a physical host has available in the environment. Although the supported ranges for memory and storage shown in Table 13 indicate configuration maximums for a vSphere and vCloud Director environment, these ranges differ for different providers, given the variance in hosting architectures and physical infrastructure designs.

3.1.3 Service Metering

Subscribers to the Basic service offering are charged over time for the aggregate amount of resources consumed across their virtual machine and/or vApp inventory for a given organization virtual datacenter. The minimum standard time interval for billing and metering purposes is typically one hour. However, providers who have the means to do so are permitted to meter and charge subscribers for resource consumption on a sub-hourly basis. If subscribers opt to change the size of their virtual machine or vApp instances after initial setup, pricing changes retroactively, defaulting to the higher charge rate of either the new or the initial vCPU or memory setting. This is referred to as the *stepping function*—the virtual machine charge always steps up to the next instance size, measured by memory or vCPU, whichever charge rate is higher.

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

Table 14 lists the most common event triggers and resource states for vCloud Director. Columns marked with an "X" signify that the resource type is considered consumed when a virtual machine or vApp is in the associated state, and corresponding charges may then apply. These are meant to be illustrative only. Providers should rely on their own internal cost models and metering schemes for billing or showback.

Table 14. vCloud Director Event Triggers and States

API Operation	UI Operation	vCPU	RAM	Network (vNIC)	Storage
Instantiate/Compose	Add/New				X
Deploy	Start			X	X
Power On		X	X	X	X
Reset	Reset	X	X	X	X
Suspend (vApp)	Suspend				X
Suspend (virtual machine)				X	X
Shut Down	Stop				X
Reboot		X	X	X	X
Power Off				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.

3.2 Service Offering – Committed

The Committed service offering is based on the Allocation Pool allocation model in vCloud Director. It guarantees subscribers a minimum resource commitment through access to a shared management control plane in a multitenant service environment. Resource commitments for CPU (GHz), memory (GB), and storage (GB) are specified by capacity allocation for each tenant organization virtual datacenter, with a percentage guarantee for each resource type. This minimum guarantee provides deterministic performance for hosted workloads while offering tenants the ability to burst over the minimum guarantee level if additional infrastructure capacity is available.

3.2.1 Service Design Parameters

As part of the design process for the Committed service offering, providers should pay special consideration to key service settings and values in vCloud Director that can affect service performance and consistency levels for a subscriber's organization virtual datacenter. Despite use of the Allocation Pool allocation model in this service, circumstances may arise that can result in subscribers overcommitting resources over time. If not properly managed, performance for all application workloads can be negatively affected. Table 15 provides an example of these key service settings, values, and justifications.

Table 15. Resource Allocation Settings Example – Committed Service Offering

Allocation Type	Value Range	Sample Value	Justification
CPU allocation	Variable (GHz) based on physical host capacity	50GHz	The maximum amount of CPU available to the virtual machines running in the target organization virtual datacenter (taken from the supporting provider virtual datacenter) and the percentage of that resource guaranteed to be available to them.
CPU resources guaranteed	0–100%	75%	
Memory allocation	Variable (MB) based on physical host capacity	100GB	The maximum amount of memory available to the virtual machines running in the target organization virtual datacenter (taken from the supporting provider virtual datacenter) and the percentage of that resource guaranteed to be available to them.
Memory resources guaranteed	0–100%	75%	
Maximum number of virtual machines	1–Unlimited	Unlimited	A safeguard that allows control over the total number of vApps or virtual machines created by a subscriber within the target virtual datacenter.

In this example, the CPU allocation setting serves as a block or aggregate limit for the entire target organization virtual datacenter and has been configured as 50GHz (50,000 MHz). Of this 50GHz resource allocation, 75% (37.5GHz of total CPU capacity) is marked as guaranteed. This implies that the remaining 25% (12.5GHz of total CPU capacity) is available for bursting if sufficient infrastructure resources are available at the time of need.

The memory allocation setting also serves as a block, or aggregate, limit for the entire target organization virtual datacenter. It has been configured as double the CPU capacity, or 100GB. Of this 100GB resource

allocation, 75%—75GB of total memory capacity—is marked as guaranteed. This implies that 25%—25GB of total memory capacity—remains available for bursting if sufficient infrastructure resources are available at the time of need. The provider can use the combination of these settings to throttle back overcommitment from aggressive levels (for example, resource guarantees set to <100%) to more conservative levels (for example, resource guarantees always set to 100%), depending on SLAs in place or fluctuating service loads.

3.2.2 Resource Allocation and Catalogs

The Allocation Pool allocation model in the Committed service offering enables providers to aggregate resources in bulk, with a minimum upfront capacity guarantee for target service subscribers whose workloads demand more stringent service levels for capacity availability and performance. Like the Basic service offering, this service tier provides equal or better flexibility in the way resources are allocated through published vApp catalogs in vCloud Director. Subscribers gain flexibility through entitlements to create their own custom and private vApp catalogs as well as to use any public catalogs and standard application images that may be made available by the provider.

Table 16 provides an example of different sizing combinations that can be included in a Committed service offering (virtual datacenter) catalog. Values in the CPU Guarantee and Memory Guarantee columns reflect a 75% resource guarantee listed in the example in Table 15. Depending on the customizable sizing options made available by a provider of this service, subscribers may be able to specify non-standard combinations of resources such as CPU, memory, and storage types and quantities. It is still recommended, however, that providers first implement a pricing model commensurate with the range of scale for each resource type.

Table 16. Committed Service Offering Catalog Example

Virtual Datacenter Instance Size	CPU Allocation (GHz)	CPU Guarantee (GHz)	Memory Allocation (GB)	Memory Guarantee (GB)	Storage Limit1 (GB)	Bandwidth2 (MBps)	Cost
Small	10GHz	7.5GHz	20GB	15GB	Variable	95th percentile	Set by provider
Medium	25GHz	18.75GHz	50GB	37.5GB	Variable	95th percentile	Set by provider
Large	50GHz	37.5GHz	100GB	75GB	Variable	95th percentile	Set by provider
Extra Large	100GHz	75GHz	200GB	150GB	Variable	95th percentile	Set by provider

[1] Storage allocations may be selected individually and can be customized for all virtual datacenter sizes from Small through Extra Large based on individual subscriber requirements.

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

3.2.3 Service Metering

Subscribers to the Committed service offering are charged for the minimum amount of resource capacity guaranteed over time for a given organization virtual datacenter, as negotiated at the point of sale. The time interval used for billing and metering purposes for the amount of guaranteed capacity can vary based on actual service terms set by the provider, but it is most often based on a monthly subscription period, with a minimum enrollment term set to longer than one month. When available, burst capacity may be consumed without specific terms and obligations, but it is up to the provider whether additional charges may be incurred for short-term bursting or temporary scale-out operations on behalf of the subscriber.

Subscribers whose workload requirements have consistently grown beyond their initial resource capacity guarantee should have the option to move into the next band of guaranteed capacity without being penalized in price. Providers, in fact, are encouraged to offer a progressive discount structure that incentivizes active subscribers to consume greater amounts of the Committed service offered as needed and with minimal disruption to service contract terms or operations.

Charges for resource consumption under the Committed service offering typically start when the target organization virtual datacenter has been fully provisioned and made available to the subscriber, with limited exceptions for certain resource types, such as storage, which may be reserved upfront without immediate use.

3.3 Service Offering – Dedicated

The Dedicated service offering is based on the Reservation Pool allocation model in vCloud Director. It provides subscribers of this service tier fully dedicated resources upfront through access to a shared management control plane in a multitenant service environment. In contrast to all other service offerings referenced in this document, the Dedicated service offering delivers premium levels of performance, security, and resource management by reserving 100% of physical CPU (GHz), memory (GB), and storage (GB) resource capacity for each target organization virtual datacenter, thereby enabling subscribers of the service to exercise granular control of overcommitment of those resources by specifying reservation, limit, and priority settings for individual virtual machines.

3.3.1 Service Design Parameters

As part of the design process for the Dedicated service offering, providers should give special consideration to key service settings and values in vCloud Director that can impact service performance and consistency levels for a subscriber's organization virtual datacenter. Although the Reservation Pool allocation model in this service provides stricter controls over the segmentation of resources allocated than the other models, circumstances can still arise that result in subscribers overcommitting resources over time, resulting in negative performance or availability implications, whether for subscribers' application workload environment or for their service settings, values, and justifications, or both. The following table provides an example of these key service settings, values, and justifications.

Table 17. Resource Allocation Settings Example – Dedicated Service Offering

Allocation Type	Value Range	Sample Setting	Justification
CPU allocation	Custom	76.8GHz	The amount of CPU resources reserved for this organization virtual datacenter (taken from the supporting provider virtual datacenter and assigned resource cluster).
Memory allocation	Custom	1024GB	The amount of memory resources reserved for this organization virtual datacenter (taken from the supporting provider virtual datacenter and assigned resource cluster).
Maximum number of virtual machines	1–Unlimited	Unlimited	A safeguard that allows control over the total number of vApps or virtual machines created by a subscriber within the target virtual datacenter.

In this table, the CPU allocation setting serves as a block or aggregate limit for the entire target organization virtual datacenter and has been configured as 76.8GHz (76,800 MHz). Of this 76.8GHz resource allocation, 100% of total CPU capacity is marked as reserved and guaranteed by default. The memory allocation setting also serves as a block or aggregate limit for the entire target organization virtual datacenter and has been configured as 1024GB. Of this 1024GB resource allocation, 100% of total memory capacity is also marked as reserved and guaranteed by default. This implies zero resource overcommitment by the provider for both compute and memory capacity, and it requires that the underlying provider virtual datacenter and associated physical resource clusters be 100% dedicated to each subscriber, to avoid any resource contention. The provider can use the combination of these settings to adjust CPU and memory capacity as needed, but the need to throttle back overcommitment for this service offering does not apply.

3.3.2 Resource Allocation and Catalogs

Using the Reservation Pool allocation model in the Dedicated service offering enables providers to aggregate resources in bulk, with a 100% upfront physical capacity guarantee for target service subscribers whose workloads demand the most stringent service levels for capacity availability, performance, and security isolation. Like the Committed service offering, this service tier provides equal flexibility in how resources are allocated through published vApp catalogs in vCloud Director. Subscribers are entitled to the same privileges for creating their own custom and private vApp catalogs, in addition to use of any public catalogs and standard application images that may be made available by the provider.

The following table provides an example of different sizing combinations that can be included in a Dedicated service offering (virtual datacenter) catalog. Values in the CPU Capacity Reserved and Memory Capacity Reserved columns imply a 100% dedicated resource guarantee upfront. Depending on the customizable sizing options made available by a provider of this service, subscribers may be able to specify non-standard combinations of resource types, such as CPU, memory, storage, and quantities. It is still recommended, however, that providers first implement a pricing model commensurate with the range of scale for each resource type.

Table 18. A Dedicated Service Offering Catalog

Virtual Datacenter Instance Size	Compute Nodes Reserved[1]	CPU Capacity Reserved[2] (GHz)	Memory Capacity Reserved[3] (GB)	Storage Limit[4] (GB)	Bandwidth[5] (MBps)	Cost
Small	2–4	9.6–19.2GHz	512–1024GB	Variable	95th percentile	Set by provider
Medium	4–8	19.2–38.4GHz	1024–2048GB	Variable	95th percentile	Set by provider
Large	8–16	38.4–76.8GHz	2048–4096GB	Variable	95th percentile	Set by provider
Extra Large	16–32	76.8–153.6GHz	4096–8192GB	Variable	95th percentile	Set by provider

[1] Compute node form factor can be either blade servers or rackmount servers. This example assumes a generic (x86) dual-processor blade server.

[2] CPU capacities in this example are based on each compute node having dual Intel 2.4GHz Xeon E7-2870 processors.

[3] Memory capacities in this example are based on the each compute node having 256GB of physical memory.

[4] Storage allocations may be selected individually and can be customized for all virtual datacenter sizes from Small through Extra Large based on individual subscriber requirements.

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

Although resources assigned at the organization virtual datacenter layer in the Dedicated service offering are fully reserved and dedicated, subscribers to the service are entitled to control resource overcommitment through reservation and limit settings for each individual virtual machine. The following table provides an example of how these resource allocation settings might be configured.

Table 19. Resource Allocation Settings per Virtual Machine

Resource Type	Priority	Reservation	Limit
CPU	• Low (500 Shares) • Normal (1000 Shares) • High (2000 Shares) • Custom	Custom (GHz) per virtual datacenter allocation	• Unlimited • Maximum (custom)
Memory	• Low (1280 Shares) • Normal (2560 Shares) • High (5120 Shares) • Custom	Custom (GB) per virtual datacenter allocation	• Unlimited • Maximum (custom)

3.3.3 Service Metering

Subscribers to the Dedicated service offering, like subscribers to the Committed service offering, are charged for the full amount of resource capacity guaranteed over time for a given organization virtual datacenter, as negotiated at the point of sale. The time interval used for billing and metering purposes for the amount of guaranteed capacity can vary based on actual service terms set by the provider, but it is most often based on a monthly subscription period, with a minimum enrollment term set to longer than one month.

Subscribers whose workload requirements have consistently grown beyond their initial resource capacity guarantee should have the option to move into the next band of reserved capacity without being penalized in price. Providers, again, are encouraged to offer a progressive discount structure that incentivizes active subscribers to consume greater amounts of the Dedicated service offered as needed and with minimal disruption to service contract terms or operations.

Charges for resource consumption under the Dedicated service offering typically start when the target organization virtual datacenter has been fully provisioned and made available to the subscriber, with limited exceptions for certain resource types, such as storage, which may be reserved upfront without immediate use.