



VMware vCloud[®] Architecture Toolkit

Operating a VMware vCloud

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. Overview 7
 - 1.1 Audience 7
 - 1.2 Scope 7
- 2. Operating a VMware vCloud 8
 - 2.1 vCloud Operations Framework 9
- 3. Organizing for vCloud Operations 11
 - 3.1 vCloud Center of Excellence 11
 - 3.2 Role of vCloud Center of Excellence in Standardization 12
 - 3.3 vCloud Service Management 13
 - 3.4 vCloud Operations Management 14
 - 3.5 vCloud Infrastructure Management 15
- 4. vCloud Service Management 19
 - 4.1 Service Catalog Management 19
 - 4.2 Service Level Management 23
- 5. vCloud Operations Management 27
 - 5.1 Configuration Management 27
 - 5.2 Orchestration Management 31
 - 5.3 Availability Management 33
 - 5.4 Continuity Management 35
- 6. vCloud Infrastructure Management 38
 - 6.1 Security and Compliance Management 38
 - 6.2 Capacity Management 41
 - 6.3 Performance Management 42
 - 6.4 Monitoring 43
- Appendix A: vCloud Director Cell Monitoring 45
- Appendix B: Compliance Considerations 53
- Appendix C: Capacity Planning 62
- Appendix D: Capacity Management 69
- Appendix E: Integrating with Existing Enterprise System Management 79
- Appendix F: Business Continuity 86

Appendix G: Upgrade Checklists 90

List of Figures

Figure 1. Cloud Computing Layers 8

Figure 2. vCloud Operations Framework 9

Figure 3. vCloud Operations Framework Mapped to Service Layers 10

Figure 4. vCloud Center of Excellence Ecosystem..... 12

Figure 5. Service Catalog Evolution..... 21

Figure 6. Example Organization with Public vCloud IaaS and Private vCloud PaaS/SaaS Layers 24

Figure 7. Configuration Management Interrelationships..... 28

Figure 8. Architectural Example Drawing..... 39

Figure 9. One Primary Function per Server..... 53

Figure 10. Log Collection in the vCloud Environment..... 56

Figure 11. Architecture of vCloud Components and Log Collection..... 57

Figure 12. Infrastructure Layers 60

Figure 13. vCloud Director Extension Overview 79

Figure 14. vCenter Orchestrator as a vCloud Director Extension..... 82

Figure 15. vCenter Orchestrator AMQP Subscription Policy 83

Figure 16. Credential Management Workflow..... 86

List of Tables

Table 1. Public Catalog Benefits	20
Table 2. Sample vApp Offerings	22
Table 3. vCloud vApp Requirements Checklist.....	36
Table 4. MBeans Used to Monitor vCloud Cells	45
Table 5. Audit Concerns Within the vCloud	53
Table 6. vCloud Component Logs	58
Table 7. Other Component Logs.....	58
Table 8. vSphere Host Variables	63
Table 9. Determining Redundancy Overhead.....	64
Table 10. Network Capacity Planning Items	68
Table 11. Capacity Monitoring Metrics.....	70
Table 12. Organization Virtual Datacenter Units of Consumption	72
Table 13. Recommended Organization Virtual Datacenter Capacity Thresholds	73
Table 14. Sample Organization Virtual Datacenter Resource Allocation	73
Table 15. Organization Virtual Datacenter Trending Information	75
Table 16. Organization Virtual Datacenter Capacity Trending Variables	76
Table 17. Sample Organization Virtual Datacenter Trending Information	77
Table 18. Approve a vApp workflow	84
Table 19. Recommended Protection Policies	88



1. Overview

Operating a VMware vCloud provides practical operations-focused, organizational, process, and supporting technology considerations and guidance based on the vCloud Operations Framework. The goal is to provide customers with the information they need to realize the benefits of proceeding along the journey of VMware vCloud® adoption and providing Infrastructure-as-a-Service (IaaS) using a service-focused, comprehensive operational framework. Both service providers and enterprises can use the guidelines in this document, with some variations depending on point of view.

The documents, *Architecting a VMware vCloud*, *Operating a VMware vCloud*, and *Consuming a VMware vCloud* are designed to work together throughout the lifecycle of a VMware vCloud computing implementation with VMware technologies. By using all three documents together, combined with a private or public service definition, you can gain a comprehensive view of VMware vCloud computing.

- *Architecting a VMware vCloud* provides best practices, design considerations, and design patterns for constructing a vCloud environment from its constituent components.
- *Operating a VMware vCloud* includes best practices and considerations for operating and maintaining a vCloud environment. It covers the people, process, and technology involved in running a vCloud environment.
- *Consuming a VMware vCloud* covers the various considerations for the consumer when choosing to leverage vCloud computing resources.

This document is not a substitute for VMware product documentation, nor does it provide detailed implementation procedures for installing a vCloud.

1.1 Audience

This document is intended for, but not limited to, IT personnel responsible for or involved in the service, operations, and infrastructure management of one or more vCloud instances to deliver Infrastructure as a Service. It is assumed that the reader has knowledge of and familiarity with VMware vSphere® concepts.

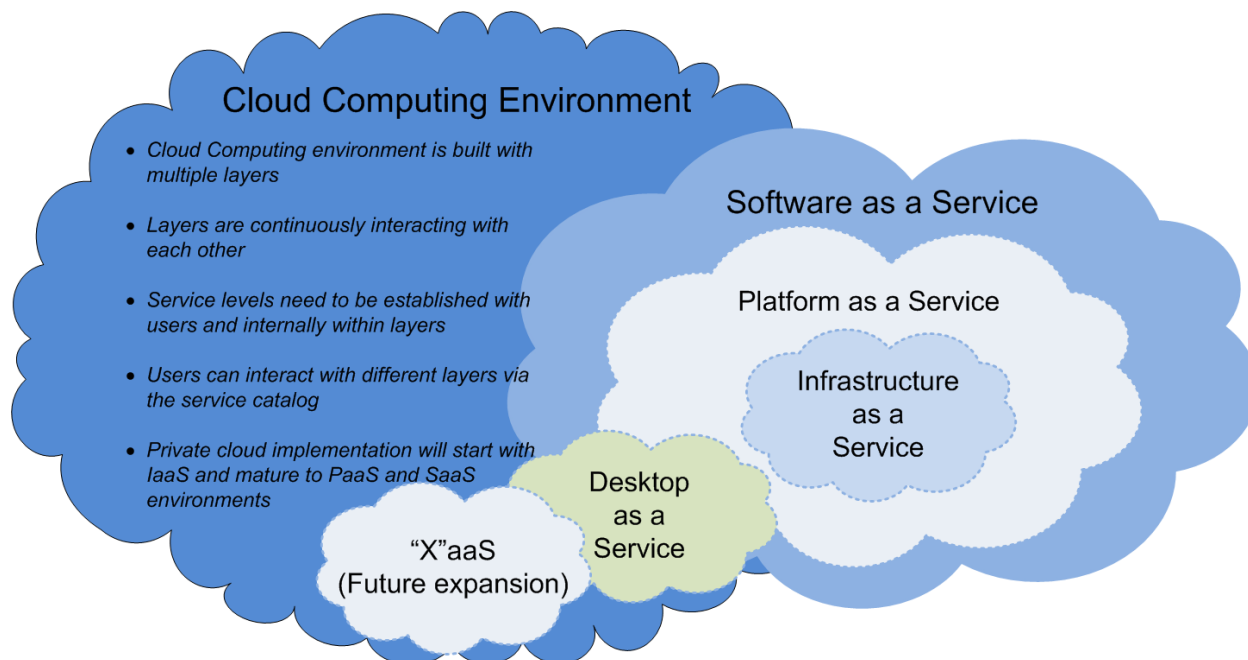
1.2 Scope

This document focuses on considerations for operating a vCloud from the perspectives of the organizational structure, service management, operations management, and infrastructure management.

2. Operating a VMware vCloud

Cloud computing is an approach to computing that leverages the efficient pooling of on-demand, self-managed virtual infrastructures, consuming them as a service. The NIST standard defines three such service layers within a cloud. Key cloud computing principles, along with the service layer paradigm are illustrated in Figure 1.

Figure 1. Cloud Computing Layers



VMware defines the existing service layers as:

- Software as a Service (SaaS) – Business-focused services presented directly to users via a service catalog.
- Platform as a Service (PaaS) – Technology-focused services presented for application development and deployment presented directly to application developers via a service catalog.
- Infrastructure as a Service (IaaS) – Services providing infrastructure “containers” for various uses in order to provide better agility, automation, and delivery of components.

Additional service layers will be added as other services, such as Desktop as a Service, become a reality.

Companies embark on the journey to adopt cloud computing to realize increased quality of service, business agility, and operating cost efficiency.

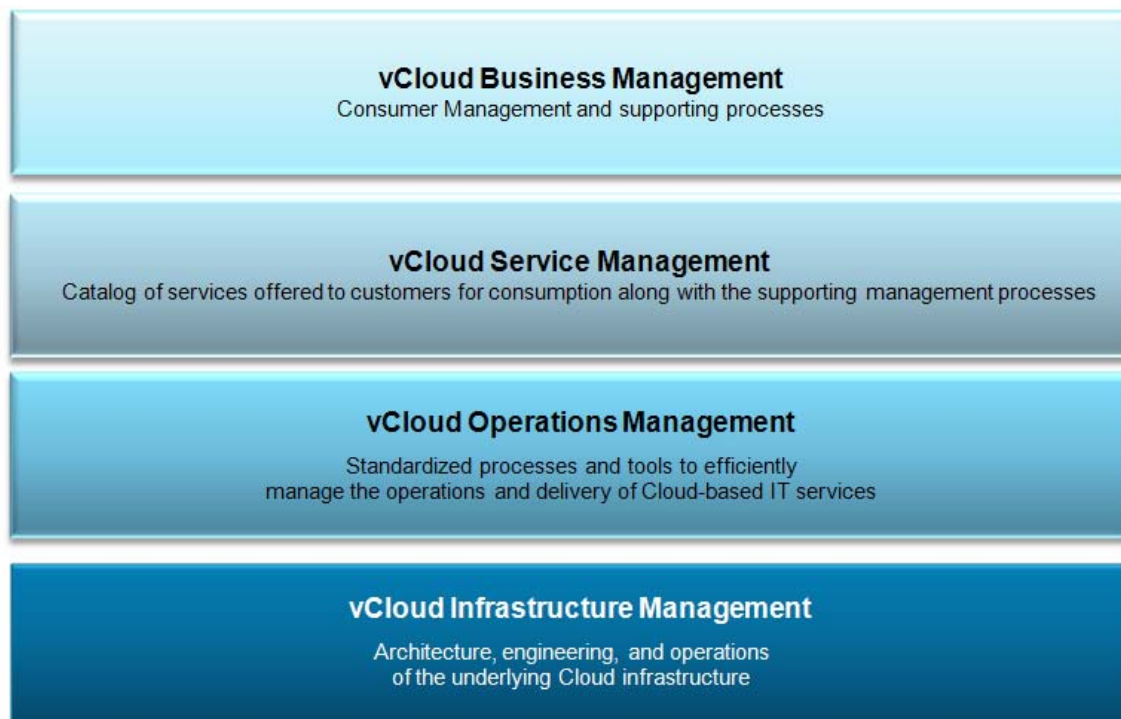
- Quality of service is achieved by providing standardized service offerings with associated availability levels and service management. Customers can expect the provision of a reliable VMware vCloud® service with predictable service levels so that end users are provided with the service they require in the manner they require within expected timeframes. To provide standardized, repeatable service, IT must introduce their own operational efficiencies to make sure that they are in control of the underlying infrastructure and applications, but are not restricting their use by over-managing it.

- Business agility is achieved by changing the way that IT thinks about managing services. Some end-user control must be provided so that IT does not become a bottleneck when provisioning services, but IT must retain some control to prevent the environment from becoming unmanageable. To provide business agility and speed to market, automation is key to the success of the vCloud. Automation of time-consuming, error-prone, and complex tasks is recommended to provide the reliable, rapid service expected from the vCloud.
- Increased cost efficiency can be realized by decreasing operational expenses. To achieve this, the current operational cost and burden of managing IT (approximately 70% operational expenses and 30% capital expenses) must change, especially as IT becomes more like a service provider providing Infrastructure as a Service. Reducing operational expenses can only be accomplished by enhancing IT operational processes for cloud computing, implementing the tools to support and automate the enhanced operational processes, and optimizing the organizational structure to most effectively align with managing the cloud computing infrastructure and IT services offered.

2.1 vCloud Operations Framework

The underlying VMware vCloud Operations Framework within which organizational structure and critical processes can be defined is shown in Figure 2.

Figure 2. vCloud Operations Framework

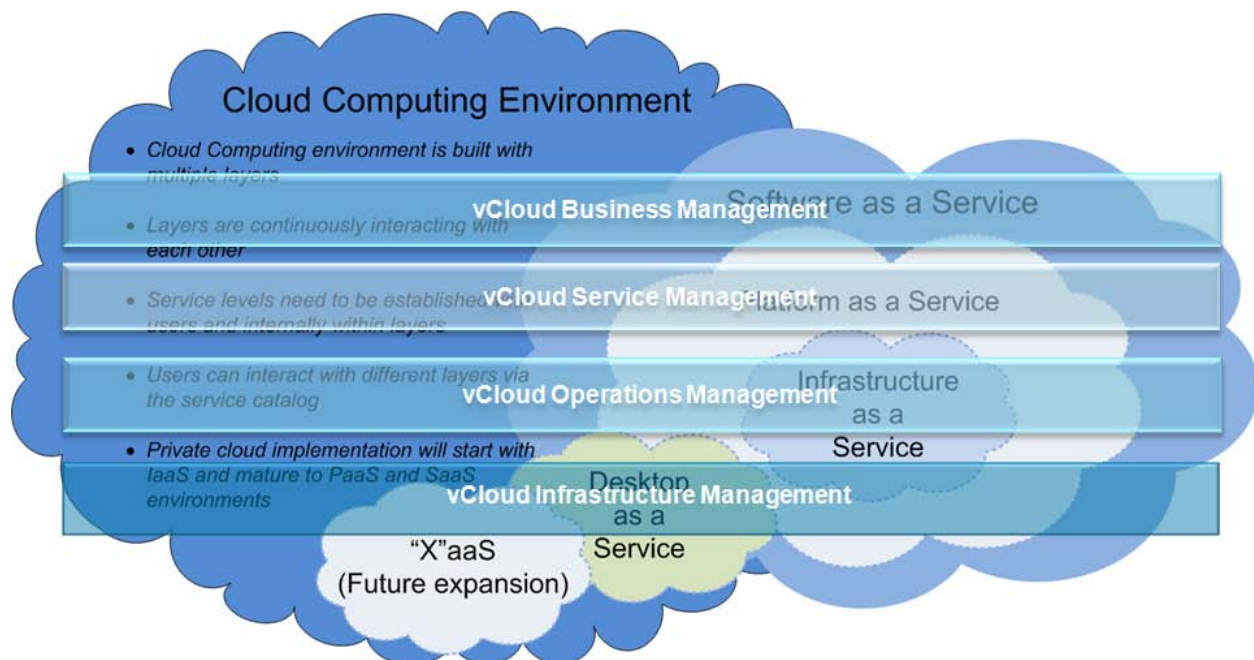


The vCloud Operations Framework consists of the following layers:

- vCloud Business Management – Addresses the source of business drivers and requirements for vCloud services to be offered, along with management of the businesses’ or line of business’ consumers and its supporting processes.
- vCloud Service Management – Converts the business drivers and requirements into vCloud service definitions, manages service development and transition, creates and reports on Service Level Agreements back to the business and its consumers, and manages the service catalog lifecycle.
- vCloud Operations Management – Defines, develops, and delivers standardized IT Service Management processes and tools to support them, to manage the operations and delivery of vCloud services.
- vCloud Infrastructure Management – Architects, deploys, and manages the underlying vCloud infrastructure upon which the vCloud services are deployed and delivered.

These layers are required regardless of the specific vCloud service layer being addressed.

Figure 3. vCloud Operations Framework Mapped to Service Layers



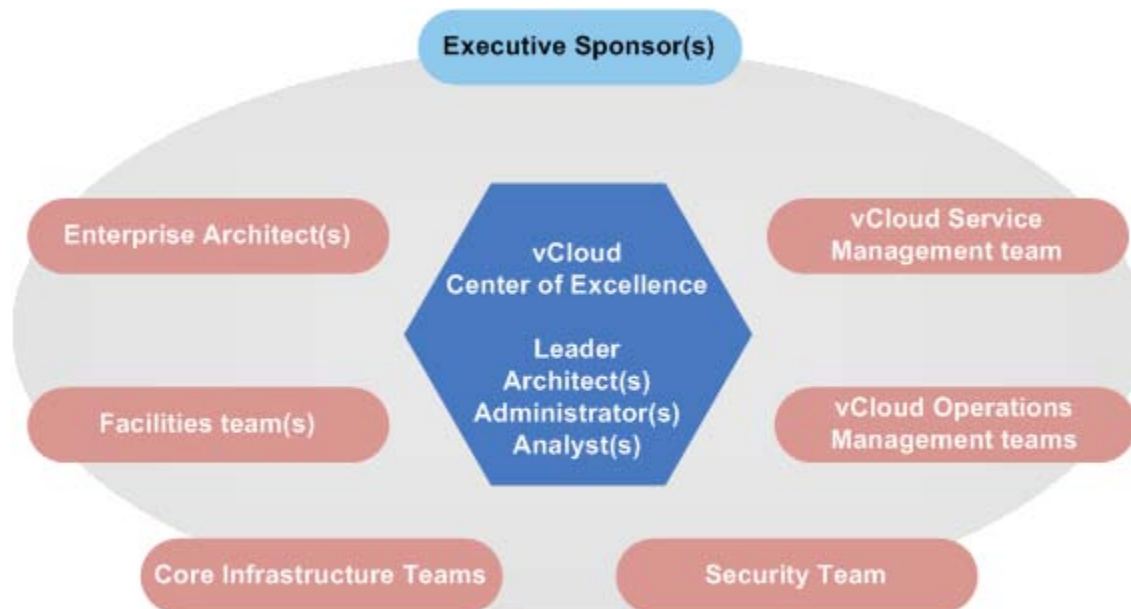
3. Organizing for vCloud Operations

One of the truly transformative aspects of vCloud computing is its impact on the IT organization. By definition, vCloud computing provides on-demand service delivery. As a result, the IT organization has to become service-driven. Delivering a vCloud-based service impacts all three layers of the VMware vCloud Operations Framework: vCloud Service Management, vCloud Operations Management, and vCloud Infrastructure Management. The VMware approach to addressing this impact is to use the vCloud Center of Excellence (vCOE) model.

3.1 vCloud Center of Excellence

The vCloud Center of Excellence model is an extension of the VMware Center of Excellence model that has been used by many organizations of various sizes to facilitate the adoption of VMware technology and to reduce the complexity of managing a VMware virtual infrastructure. The vCloud Center of Excellence model defines cross-domain vCloud Infrastructure Management accountability and responsibility within team roles across an organization. These team roles enable an organization to consistently measure, account for, and improve the effectiveness of its vCloud infrastructure management even if its IT Service Management roles and responsibilities are distributed across multiple IT functional areas. The vCloud Center of Excellence model also involves the proactive inclusion of vCloud Infrastructure Management *champions* who represent related functional teams that are critical to providing vCloud Infrastructure Management in support of vCloud service offerings.

The vCloud Center of Excellence is a focused “virtual” team of vCloud Infrastructure Management specialists and an ecosystem of related functional groups which, taken together, form a vCloud Center of Excellence extended team as shown in Figure 4. It serves as the focal point for all decisions involving vCloud Infrastructure Management, including infrastructure architecture, implementation, and management. The team is tasked with creating, reviewing, and publishing vCloud Infrastructure Management guidelines and documentation, as well as effecting policy and process change for internal use. To reach an operationally mature level, you must establish a formal vCloud Center of Excellence with aligned roles and responsibilities.

Figure 4. vCloud Center of Excellence Ecosystem

As shown in Figure 4, the vCloud Center of Excellence ecosystem includes the other two layers of the vCloud Operations Framework, vCloud Service Management and vCloud Operations Management. It also includes the other core IT teams: Enterprise Architecture, Facilities, Infrastructure (network, storage, servers), and the Security team. The organizational structure of Enterprise Architecture, Facilities, Core Infrastructure, and Security are not addressed in this document.

3.2 Role of vCloud Center of Excellence in Standardization

In a traditional organization, IT is driven by multiple business units. The business units control IT funding and each business unit (BU) can enforce separate policies and procedures for their infrastructure. This approach leads to disjointed architectures and a lack of standardization. IT groups that support such an environment struggle to achieve agreed service levels, leading to end-user frustration, IT support inefficiencies, and possibly even financial liabilities. The implementation of a vCloud changes this paradigm. A vCloud is built as a shared resource that requires enforcement of consistent standards across the entire IT organization. To define and enforce these standards, all policies and procedures associated with the vCloud should be driven by the vCloud Center of Excellence team rather than from business units. This shift poses a significant challenge for organizations who try to move into a vCloud-appropriate operating model. The vCloud Center of Excellence needs to negotiate with different business groups and rely on executive sponsorship and support during this transition. More rigorous standards need to apply across the whole organization.

One recommended approach is to align with the organizations' phased development approach, adding a new vCloud architecture review and signoff by the vCloud Center of Excellence during the analysis and design phase for all new projects. Other recommendations include running vCloud-specific assessments on applications that are being considered for migration to the vCloud. Assessments determine gaps and set expectations with business units on expected changes. The key to success is the ability to balance agility to meet business needs with stringent enforcement of defined standards within the vCloud.

3.2.1 Layers of Standardization

The vCloud is a shared resource running on infrastructure supported by the vCloud Center of Excellence and core infrastructure teams. As the vCloud Center of Excellence sets standards for the vCloud, core infrastructure teams may develop standards for the infrastructure that supports the vCloud. For example, the storage team may create standards for how new logical unit number (LUN) storage is presented for vCloud consumption. This layer of abstraction allows the storage team to have the flexibility to choose the most cost-effective SAN vendor and, if required, support a multi-vendor environment.

3.2.2 Measurement with Industry Benchmarks

Cloud technology is evolving at a rapid pace. After a vCloud is established within an organization a continuous improvement cycle needs to be set up annually to make sure that the organization's vCloud is not lacking any current industry standards or benchmarks. The vCloud Center of Excellence is responsible for running this assessment and presenting the results, including recommendations for remediation, back to the leadership team.

3.3 vCloud Service Management

vCloud Service Management is critical to providing vCloud service offerings and is the entry point into IT from vCloud Business Management. The following are roles and responsibilities that are involved with vCloud Service Management:

- vCloud Service Portfolio Management:
 - Manages the portfolio of vCloud services and works with organizational leadership to develop the vCloud service offering strategy used to determine what services should be included in the overall portfolio.
 - Proactively identifies and defines potential vCloud service offerings based on demand information gathered from vCloud Business Management or other sources such as requests coming in through the Service Desk.
- vCloud Service Owner:
 - Responsible for overall delivery of their vCloud service offering.
 - Provides the required information to Service Catalog Management to correctly set up the service catalog offering.
 - Works with Service Level Management to review Service Level Agreements and Operating Level Agreements to make sure that they are achievable. Also, negotiates updated Service Level Agreements and Operating Level Agreements as the service offering is updated.
 - Leads development and enhancement efforts and works with vCloud Service Development Management on their vCloud service offering based on new requirements from vCloud Business Management.
 - Liaises between IT Business Management and vCloud Center of Excellence.
- vCloud Service Development Management:
 - Defines a vCloud service offering based on the requirements provided by vCloud Business Management after it's determined that a particular vCloud Service Offering is to be included in the vCloud Service Portfolio. This involves translating vCloud Business Management requirements into requirements that are used by vCloud Infrastructure Management to create deployment templates.
 - Manages any additional efforts required to populate deployment templates, such as working with Application Development Managers who may provide an application for a vCloud service offering.

- Works with vCloud Business Management and Financial Management to determine a price for a vCloud service offering, and determine whether multiple prices are appropriate if the vCloud service offering is provided in multiple service tiers.
- Creates, collects, and maintains any vCloud service offering development documentation.
- vCloud Service Level Management:
 - Defines the Service Level Agreement (SLA) associated with a vCloud service offering or a tier of service such as provided by a particular provider virtual datacenter.
 - Makes sure that the service levels are met through corresponding Operating Level Agreements with vCloud Operations Management and vCloud Infrastructure Management.
 - Regularly monitors and reports on service level attainment.
- vCloud Service Catalog Management – Manages the vCloud service offering catalog and makes sure that all of the information contained in the catalog is accurate and up-to-date.

These roles and responsibilities can be satisfied by a single person or multiple people. The decision to employ one or multiple people depends on the number of vCloud service offerings. If the number of vCloud service offerings is large enough to justify multiple people, or as the number of vCloud service offerings grows and multiple people are needed, the recommended distribution is:

- A single vCloud Service Portfolio Manager.
- A single vCloud Service Catalog Manager.
- A single vCloud Service Level Manager.
- Multiple vCloud Service Owners, each responsible for vCloud service development and working with other teams to make sure that the agreed vCloud service levels for their vCloud service offering or suite of vCloud service offerings are maintained.

As key members of the vCloud Center of Excellence ecosystem, they also act as vCloud champions. In this role, they interact closely and regularly with the vCloud Center of Excellence as well as championing vCloud to the business and other teams with which they interact in the organization at large.

3.4 vCloud Operations Management

To realize the operating expense savings benefit offered by vCloud computing, vCloud Operations Management must be responsible for developing and executing standardized, automated processes and tools optimized to efficiently manage the operations and delivery of vCloud service offerings. The key terms are *standardized*, *automated*, *optimized*, and *service*. Though vCloud computing does not require a radical reorganization of how ITSM operations are provided, it does require each role to have:

- An unwavering service-oriented focus for operations resources (both people and tools), as opposed to the traditional technical infrastructure focus.
- The goal of standardizing and automating the processes required to execute their responsibilities to the highest possible degree.
- An understanding of the impact vCloud computing has on their area of responsibility, and responsibility for optimizing their area for vCloud computing.

These requirements should be directly related to each individual's performance goals.

The areas within vCloud Operations Management impacted by or impactful on vCloud include:

- Provisioning Management.
- Change Management.
- Capacity Management.
- Availability Management.
- Performance Management.
- Orchestration Management.
- Event Management.
- Incident and Problem Management.
- Configuration and Asset Management.
- Continuity Management.

Due to the increased focus on automation, in addition to the traditional IT Service Management areas, one new role and associated set of responsibilities should be included. That role is *Orchestration Management* (or Automation Management). Orchestration Management is a cross-domain role within vCloud Operations and the person in this role is responsible for:

- Being the orchestration expert within vCloud Operations.
- Working with Enterprise Architecture to develop the operation's orchestration strategy.
- Working with the other vCloud Operations roles to design, develop, test, and deploy their specific process automation workflows.
- Developing and maintaining process automation workflow documentation.
- Working with Event Management and vCloud Infrastructure Management to establish workflow monitoring and, to the extent possible, automate responses to events.
- Providing Level 3 process automation workflow incident resolution.

Like the vCloud Service Level Management team, members of the vCloud Operations Management team are key members of the vCloud Center of Excellence ecosystem and also act as vCloud champions. In this role, they interact closely and regularly with the vCloud Center of Excellence as well as championing vCloud to the other teams with which they interact in the organization.

3.5 vCloud Infrastructure Management

vCloud Infrastructure Management is responsible for architecture, deployment, and operations of the underlying vCloud infrastructure, and gains the most by reorganizing. Traditional Infrastructure Management is siloed by infrastructure domain with very little cross-domain interaction unless required for a particular project or deployment. Virtualization provided the most recent and compelling opportunity for Infrastructure Management to break from this traditional approach. If Infrastructure Management broke from the traditional approach, adapting to vCloud computing should be reasonably straightforward. If they did not, they must now break with the traditional approach to successfully support vCloud computing. This reorganization takes the form of the vCloud Center of Excellence core team. Generally, the vCloud Center of Excellence encompasses responsibility for both the VMware vSphere and vCloud layers of the infrastructure. For the purpose of the following descriptions, vCloud encompasses the vSphere infrastructure as well.

The primary roles for members of the vCloud Center of Excellence core team are described in the following sections.

3.5.1 Executive Sponsor

- Provides clear messaging, leadership, and guidance to the entire IT organization and affected organizations about the vCloud Center of Excellence.
- Drives the cross-domain alignment required for establishing a successful, functioning vCloud Center of Excellence extended team. This level of sponsorship is key to breaking down organizational barriers and mandating integrated process design and implementation across the affected organizations. Cross-domain alignment and integrated process implementation are absolutely required to sustain a vCloud infrastructure at the level required to support vCloud-based service offerings and associated service levels.

3.5.2 vCloud Center of Excellence Leader

- Provides leadership and guidance to vCloud Center of Excellence members.
- Has a direct line of communication to the executive sponsors.
- Has visibility into the planned vCloud-based service offering portfolio as well as any portfolio changes.
- Is responsible and accountable for making sure that the vCloud infrastructure can support and continue to support the vCloud-based service offerings and service levels.
- Actively promotes awareness of the impact the vCloud infrastructure has on service offering and service level support and delivery.
- Facilitates integration of the vCloud infrastructure into existing IT Service Management processes.
- Coordinates and assists with planning vCloud infrastructure initiatives.
- Provides guidance to change management for changes related to the vCloud infrastructure; may authorize low risk, low impact changes to the vCloud infrastructure.
- Facilitates development and maintenance of vCloud infrastructure capacity forecasts.
- Manages the acquisition and installation of vCloud infrastructure components.
- Maintains relationships with the vCloud Center of Excellence extended team members and provides subject matter expertise to team members as required.
- Is involved in managing vendor relationships for vCloud infrastructure components.

3.5.3 vCloud Center of Excellence Architect

- Responsible for development and maintenance of vCloud infrastructure architecture and design documents and blueprints.
- Works closely with storage and network groups to architect and design vCloud infrastructure extensions.
- Works with enterprise architects to make sure that the vCloud infrastructure architecture is aligned with company architectural standards and strategies.
- Responsible for architecting and designing the vCloud layer in support of the planned vCloud-based service offering portfolio and any portfolio changes.
- Responsible for working with the IT Security team to make sure any architecture or design decisions address security and compliance.

- Responsible for architecting or designing solutions for vCloud infrastructure integration points with extended team systems.
- Provides subject matter expertise to support design, build, configuration, and validation processes.
- Maintains awareness of VMware software patches and their impact on the environment.
- Develops and maintains operational guidelines for the maintenance and support of the vCloud infrastructure.
- Mentors vCloud Center of Excellence core and extended team members.
- Assists with the incident and problem management processes to resolve issues related to vCloud infrastructure.
- Develops software and hardware upgrade plans.
- Develops and maintains the availability policy for the vCloud infrastructure in coordination with Availability Management and Service Level Management.

3.5.4 vCloud Center of Excellence Analyst

- Responsible for the development and maintenance of the vCloud infrastructure capacity forecast.
- Responsible for the day-to-day capacity and resource management of the vCloud infrastructure.
- Works with the IT Security team to make sure that the vCloud infrastructure aligns with IT security and compliance policies.
- Initiates requests for new vCloud infrastructure components.
- Assists with Incident and Problem Management processes for issues related to vCloud infrastructure capacity and performance.
- Assists with Change Management process as applied to the vCloud infrastructure.
- Responsible for maintaining the vCloud infrastructure Configuration Management data.
- Responsible for validating billing metering data collected for the vCloud-based service offerings.

3.5.5 vCloud Center of Excellence Administrator

- Installs and configures vCloud infrastructure components.
- Executes the validation plan when deploying new infrastructure components.
- Works with vCloud Center of Excellence extended team members to configure vCloud infrastructure components.
- Responsible for auditing vCloud infrastructure component configuration consistency.
- Develops and maintains vSphere and vCloud internal user access roles.
- Creates, configures, and administers vCloud provider-related objects.
- Works with the IT Security team to implement vCloud-related security and compliance policies.
- Works with Service Level Management to determine maintenance windows for the vCloud infrastructure.
- Provides Tier 3 support of the vCloud infrastructure.

- Tests and installs vCloud infrastructure patches.
- Confirms that the vCloud infrastructure is correctly instrumented for monitoring and logging purposes.
- Responsible for working with other teams to implement any required vCloud integration with their systems.
- Works with Orchestration Management to implement the vCloud infrastructure-impacting workflows.

These roles and responsibilities require a unique set of skills so, at a minimum, each role should be filled by a different person. The number of people in each role, with the exception of the vCloud Center of Excellence Leader, depends on the scale and scope of the vCloud infrastructure.

4. vCloud Service Management

4.1 Service Catalog Management

The purpose of a *service catalog* is to provide a clearly defined set of services available to customers for consumption in a vCloud environment. Ideally, the service catalog is offered from a “one stop shop” where a customer can select the services they require with minimal intervention or manual activity. An initial vCloud service catalog should align with the vCloud provider’s goals, but should aim for simplicity while at the same time integrating with capacity planning and cost transparency. Regular reviews of the service catalog should be performed with adjustments made in line with any increased functionality provided by future releases of VMware vCloud® Director™ (VCD), vSphere, or any additional supporting products.

4.1.1 vCloud Service Catalog Components

The service catalog for the vCloud supported by VCD offers a number of different service components to the end customer. The combination of all of these components creates the service as a whole. At a minimum, the vCloud provider must offer:

- Organization – This is the *container* for the customer’s IaaS with attributes that hold basic, default service configuration information. Typically, only one organization container is purchased per customer.
- Organization virtual datacenters – These are the boundaries for running the virtual machines within the IaaS service. They are configured with sizing information depending on the customers’ requirements, and have an appropriate SLA assigned to them. A minimum of one organization virtual datacenter is required for a customer to offer a service, but additional organization virtual datacenters can be requested if required.

In addition to these core vCloud components, it is possible for the vCloud provider to establish a standard set of offerings within the vCloud service catalog to provide vApps (standardized groupings of preconfigured virtual machines) and media (installable software packages) to end customers. These offerings are grouped into the following types of VCD catalogs:

- Public catalogs – These contain vApps and media (for installation of software) that are offered to the end customer by the vCloud provider.
- Organization catalogs – These also contain vApps and media, but are only available within an individual organization, and can only be shared with individuals within that particular customer organization. Organization catalogs are created, controlled, and owned by individuals within the customer organization.

The benefits associated with offering vApps and media via public catalogs as part of vCloud provider's overall service catalog are listed in Table 1.

Table 1. Public Catalog Benefits

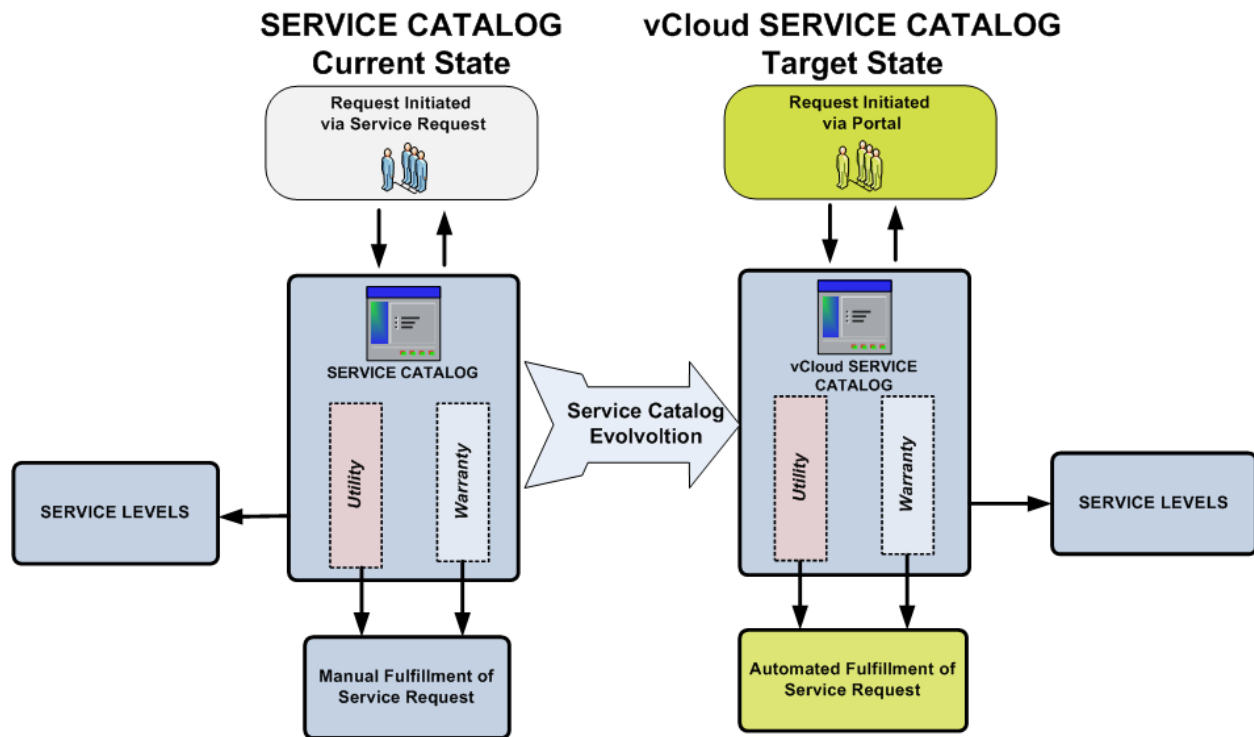
Qualities	Description
Supportability	By offering a discrete set of services, it is much simpler to provide a reliable demand pipeline and, in turn, provide the capacity to support that pipeline. By providing standardized vApps, it becomes simpler to manage the overall environment.
Agility	Standard offerings provide for the simple selection of virtual machine configurations (vApps) and enable quick provisioning using vCloud Director.

4.1.2 vCloud Service Catalog Evolution

To improve the vCloud service catalog process and to help realize vCloud benefits as many service offerings as possible should be made available directly to the end user with automated provisioning.

Typically, in the virtualization world, the initial process for procurement of virtual machines follows the model that is applied to physical infrastructure. Though this works, it is not the most efficient mechanism for providing services, and vCloud benefits cannot be fully realized unless this process is changed. A logical representation of the evolution of the vCloud service catalog from this current state to the desired end state is illustrated in Figure 5.

Figure 5. Service Catalog Evolution



In the Service Catalog current state, when a new service is requested, a Service Request is raised that is used to select and provision an offering from the service catalog. This request includes not only the utility (vApp or organization virtual datacenter) to be provided to the customer, but also includes the required service level (this is provided by the virtual datacenter in which the vApp is to be provisioned as well as any built-in availability features within the vApp itself). After the service has been ordered, the end customer must wait for staff to fulfill the Service Request for the virtual machines that will provide the service to be provisioned.

To satisfy the self-service, on-demand attribute of vCloud computing, the customer should be able to connect to a portal, select the required service offering, and have it automatically provisioned. This removes the manual task of selecting from the service catalog, and also removes the delay in the provisioning processes. This process is shown as the vCloud Service Catalog target state in Figure 5.

VCD provides the ability to manage these requests from the service catalog. For vApps, an organization administrator can determine who within the organization has rights to request and provision vApps, thus providing end-to-end self-service. With VCD, not only can the end user select and provision the vApp, but the user can also specify in which organization virtual datacenter it is deployed. Because organization virtual datacenters are associated with provider virtual datacenters, the end user is effectively selecting the service level they require.

The evolution to the target vCloud service catalog can be accomplished as follows:

1. Continue with the Service Request process until the vCloud service catalog is available on the portal.
2. Enable appropriate IT staff to perform vCloud service catalog requests with automated provisioning, including required approvals, on behalf of the end user.
3. Add the ability for end users to access the vCloud service catalog and request services that result in automated provisioning, including required approvals, of the corresponding vApps.

4.1.3 Standardization of vCloud Offerings into the Service Catalog

Standardization of the service offerings is essential to achieving a scalable, cost efficient vCloud environment. Typically, compute resource-based service offerings (CPU, memory, and storage) are a baseline for vCloud consumption and should be standardized as much as possible regardless of whether they apply to organization virtual datacenters or vApps (and their associated virtual machines).

Compute resources for organization virtual datacenters available in the service catalog should be standardized into various sizes. Additionally, the required compute resource configurations vary depending on the selected VCD allocation model (Allocation Pool, Pay-as-You-Go, or Reservation Pool), because attributes such as CPU speed and CPU/memory guarantee vary. Combining these two components means the service catalog could offer a number of differently sized organization virtual datacenters for each type of allocation model.

Similarly, to create a vApp catalog item (public or organization), there should be as much standardization as possible. Initially, from a compute resource point of view, standard sized virtual machines should be created to enable a *pick list* of machines for vApp creation. These standardized virtual machines could vary in resource size for CPU, memory and storage; for example, Standard, Standard Plus, Advanced, Premium, and Premium Plus. As a vApp comprises a number of individual virtual machines, the appropriately sized virtual machines can be selected from the pick list during the vApp catalog creation process.

In addition to the basic compute offerings of the virtual machines within the vApps, it is necessary to develop the service catalog to include vApp software configurations. These could be basic groupings of compute resources or could be expanded over time to offer more advanced services. Sample vApp offerings are shown in Table 2.

Table 2. Sample vApp Offerings

vApp	Configuration
2-Tier Standard Compute	1 x Standard RHEL Web virtual machine 1 x Standard Windows Server 2008 Application virtual machine
3-Tier Standard Compute, Advanced Database	1 x Standard RHEL Web virtual machine 1 x Standard RHEL Application virtual machine 1 x Advanced MySQL Database virtual machine
3-Server Standard Plus Compute (not necessarily tiered)	3 x Standard Plus Windows Server 2008 Application virtual machine

4.1.4 Establish Service Levels for vCloud Services in the Service Catalog

To provide an appropriate level of service depending on the vCloud customers' requirements, services should be further differentiated by their corresponding service levels. These service levels can be defined by offering availability and recoverability attributes such as Recovery Time Objective (RTO), Recovery Point Objective (RPO), and incident response times. These attributes can be applied to the different components within the service catalog.

Within a vApp, it is possible to design for different service levels via the virtual machines contained in the vApp. For example, a vApp could contain multiple Web servers to provide resilience in the event of server failure, and thus a lower RTO for the service.

As virtual datacenters provide abstracted physical and virtual resources, different service levels can be defined by using (or not using) the underlying hardware technology (server capabilities, storage array technologies, storage protocols, replication, and so on) and virtualization technology (HA, DRS, VMware vSphere® vMotion®, and others).

Taken together, the vApps offered and the capabilities of the virtual datacenters on which they can be deployed make for a powerful and complete vCloud service catalog.

4.2 Service Level Management

Service Level Management defines the SLA associated with a vCloud service offering or a tier of service, makes sure that the service levels are met through corresponding OLAs, and regularly monitors and reports on service level attainment.

4.2.1 Definition of Service

IT services can be defined as a set of related activities or workflows that serve a defined business purpose, supported by a combination of people, process, and technology components. Generally, IT services are offered to users through a service catalog.

4.2.2 Service Types

Service types include business user services and technology services.

4.2.2.1. Business User Services

Services are generally directly consumed by end users and are available as part of the organization's enterprise service catalog.

4.2.2.2. Technology Services

Technology services are not consumed directly by users, but enable infrastructure automation that enhances an IT organization's ability to better support business needs.

4.2.3 Service Interrelationships

To optimally provide vCloud business user services, all types of technology services need to be seamlessly integrated. Generally, this is accomplished using a workflow engine called the *orchestration layer*. Invoking a business user service may automatically trigger one or a combination of technology services. The rules governing these workflows need to be preconfigured and preapproved for control. They are also needed to provide an agreed to level of service to the business user. This agreed to level of service is known as a *Service Level Agreement* (SLA).

4.2.4 Definition of Service Level Agreement

A service level is a predetermined agreement between the service consumer and the service provider that measures the quality and performance of the available services. SLAs can be of multiple types, from measuring pure server uptime to measuring response time for technology components, process workflows, users, and so on.

There are services running at every layer of the vCloud stack, so service consumers may be business users or internal IT groups who access the vCloud primarily for technology and infrastructure services. In cases where SLAs are established for base technology services that are not consumed directly by business users, but are needed to make sure that downstream operations and infrastructure components support the business users' SLAs, these agreements are referred to as *Operational Level Agreements* (OLAs).

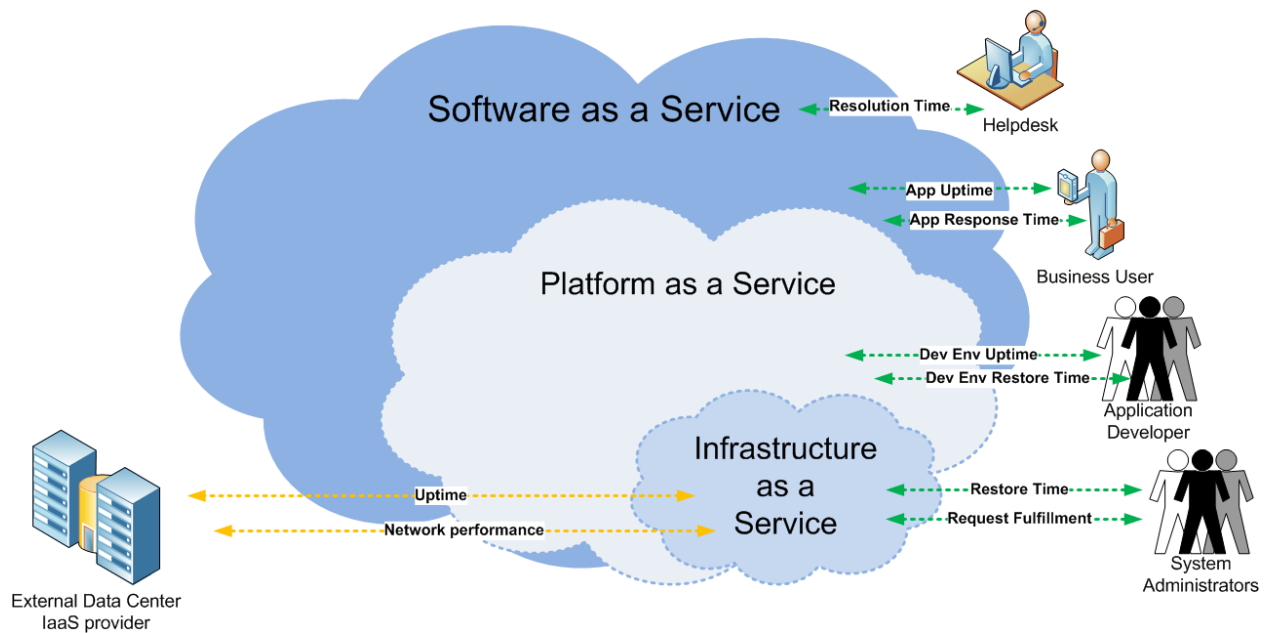
4.2.5 vCloud Layers and SLAs

A typical vCloud computing environment consists of multiple layers (IaaS, PaaS, SaaS, and so on.). Based on business requirements, every customer chooses how to implement the vCloud stack. Options include creating a private vCloud, using a public vCloud provider, or creating a hybrid vCloud model in which both private and public vCloud resources are used. The enabler for this flexibility is the ability of an organization to guarantee the availability and performance at every vCloud layer. This is achieved by signing SLAs externally with service providers, and for a private vCloud, creating SLAs and OLAs internally.

4.2.5.1. Example

The following is a use case example for an organization with an IaaS layer hosted by a public vCloud provider, but PaaS and SaaS layers are maintained internally.

Figure 6. Example Organization with Public vCloud IaaS and Private vCloud PaaS/SaaS Layers



Note The SLAs listed are for illustration purposes only and are merely a subset of the total number of SLAs created within an organization in such a case.

SLAs for this example are:

- IaaS Layer:
 - Uptime/Availability SLA signed with the external vCloud service provider.
 - Network performance SLA signed with the external service provider.
 - Request Fulfillment SLA – Measure of response time for provisioning and access configuration requests.
 - Restore time SLA.
- PaaS Layer:
 - Uptime/Availability SLA for development environment.
 - Uptime/Availability SLA for critical development environment components.
 - Restore time SLA for development environment.
- SaaS Layer:
 - Uptime/Availability SLA specific to an application.
 - Application response time SLA – Measure of how the application is performing for the business users.
 - Time to resolution SLA – Time to recover an application in case of a failure.

Given this example, the following are some key takeaways:

- SLAs are relevant at all levels within a vCloud stack. SLAs are required to provide efficiency and accountability at every layer, for both external providers and internal IT groups. Managing SLAs within every layer also helps isolate systemic problems and eliminates delays.
- SLAs can be between external vendors or providers of vCloud services or between internal IT groups. An organization can choose whether to implement a private, public, or hybrid cloud. At every layer, SLAs give organizations this flexibility by guaranteeing availability and quality of service.
- There are interrelationships between SLAs set up at different vCloud layers. A change in quality of service or breach of an SLA at a lower vCloud layer may impact multiple SLAs in a higher vCloud layer. In this example, if there is a breach of a performance SLA that results in the external vCloud provider's inability to support OS performance needs, this breach has a ripple effect at the SaaS layer, decreasing application performance and response time for business users.
- SLAs need to be continuously managed and evaluated to maintain quality of service within a vCloud. Business needs are continuously evolving, resulting in changing vCloud business requirements. SLAs must be continuously updated to reflect current business requirements.

Consider the impact of adding another 1000 users to a particular application. Given this new demand the application criticality is increased and the application is classified as mission critical. This business change means that SLAs supporting the application may need to be updated to provide increased uptime and availability. This may lead to increased demands at the IaaS layer, so SLAs with the external IaaS provider may have to be expanded.

4.2.6 vCloud SLA Considerations

Some example vCloud SLA considerations are:

- Uptime/Availability SLA:
 - Business hours – For what timeframe does the SLA pertain? These are generally divided into tiers depending on business criticality (9 to 5, 24 by 7).
 - Are maintenance windows (configuration changes, capacity changes, OS and application patch management) included or excluded from availability SLAs?
 - Single versus multi-virtual machine vApps – Do multi-virtual machine vApps need to be treated as a single entity from a SLA perspective?
- End User Response Time SLA – This is generally focused on end-to-end response time as perceived by the business user. This may require implementing remote simulators to measure and monitor response time.
- Recovery (system, data) SLA – What Recovery Time Objectives and Recovery Point Objectives need to be met?
 - Are backups required?
 - Is high availability required?
 - Is fault tolerance required within the management cluster?
 - Is automated disaster recovery failover required within certain time parameters?
- Privacy SLA (data security, access and control):
 - Do data privacy requirements (encryption, others) exist?
 - Are there regulatory requirements?
 - Are specific roles and permission groups required?
- Provisioning SLA – Are there provisioning time requirements?

5. vCloud Operations Management

vCloud Operations Management includes Configuration Management, Orchestration Management, Availability Management, and Continuity Management.

5.1 Configuration Management

Configuration Management is a key process for realizing vCloud operational benefits. Configuration Managers track configuration items using a configuration management system.

5.1.1 Configuration Management Definition

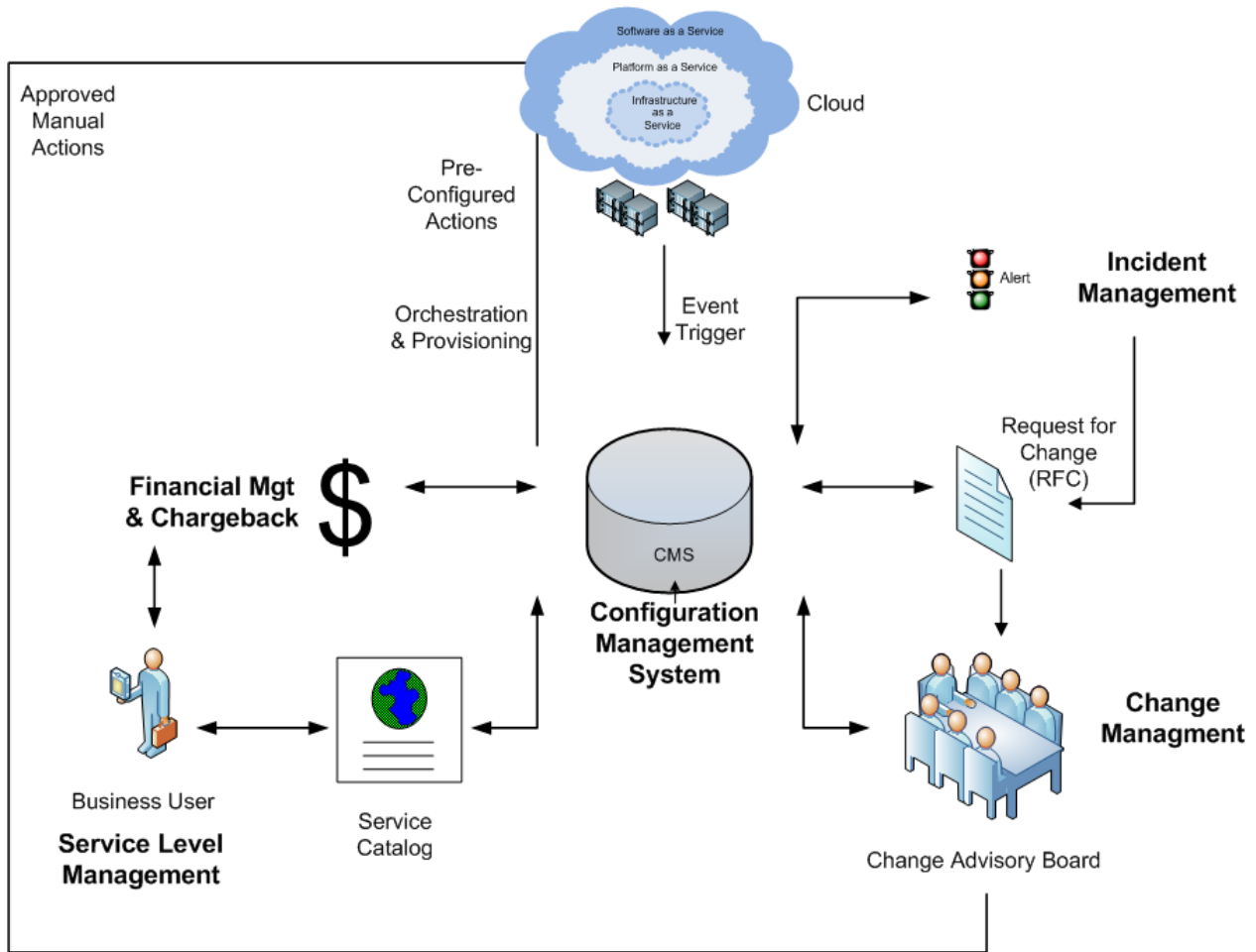
Configuration Management is the process responsible for defining and maintaining information about components of an IT service—these components are referred to as *configuration items* (CIs) and are managed end-to-end throughout their lifecycle. The goal of configuration management is to understand the historical, planned, and current state of configuration items, their interrelationships, and their impact on an IT service.

5.1.2 Value of Configuration Management in a vCloud

To fully realize the benefits of vCloud computing from an operations perspective, configuration management is the key. The configuration management process is administered in an organization through a set of tools and databases collectively known as a *Configuration Management System* (CMS). The CMS acts as the definitive source of record for all components and their interrelationships within a vCloud environment. The CMS provides the visibility IT needs to manage the multilayered vCloud environments. Configuration management is the enabler for critical vCloud functions, from automated provisioning, discovery, and maintenance, to helping effectively plan and implement changes in the environment.

Figure 7 shows interrelationships between a Configuration Management System and other critical IT Service Management processes.

Figure 7. Configuration Management Interrelationships



The following sections provide a detailed review and explain how configuration management supports a vCloud in an IT organization.

5.1.2.1. Configuration Management in Relation to Orchestration and Provisioning in a vCloud

In the vCloud environment users request new resources (virtual machines, vApps) that are provisioned directly from the service catalog. The enabler for this automated provisioning is the CMS. After a request is received from a user via the service catalog the orchestration engine interacts with the CMS. For example, to confirm available capacity, it determines configuration data for provisioning as well as access and security restrictions, and commissions the new virtual machine. The CMS is then informed of this new virtual machine and begins its lifecycle management.

5.1.2.2. Configuration Management in a Multilayer Environment

A vCloud environment, as opposed to a traditional IT environment, is built of independent vCloud layers. These layers can be built internally or created by the use of internal and external service providers. To manage this new dynamic environment, the CMS role becomes even more critical. The CMS needs to have capabilities to expand and understand all of the internal and external components of an IT service. Therefore, a CMS needs integration and reconciliation abilities with multiple sources and needs to be service-centric to show dependencies for how different components (internal or external) can impact an IT service.

5.1.2.3. Configuration Management Needs to Be Self-Aware

A vCloud environment is dynamic. The mobility of workloads and resources within the vCloud makes tracking components a challenge. A CMS platform for a vCloud needs automated discovery and mapping functions to make it self-aware. Tighter integration with virtual vCloud infrastructures allows for real time views of a service to be created. This allows the IT organization to organize their vCloud infrastructures better with more predictive planning capabilities.

5.1.2.4. Configuration Management in relation to Change Management

A vCloud has dynamic functions that require their own set of change considerations. These changes need to be preapproved by change management, and stored as configuration data within the CMS to be used by the orchestration layer to execute.

CMS also remains central to the change management process itself, providing the visibility required by the change advisory board to access the impact of a change in the environment.

5.1.2.5. Configuration Management in Relation to a Service

A CMS needs to be service-centric, with graphical views showing IT service dependencies. The CMS should also be used to store service-related information such as expected service levels and key stakeholder information. As IT organizations evolve to be more service oriented the CMS role in relation to service-related data increases. It will include not only component information, but information about service dependencies and tools that enable outage simulations to understand impacts. The CMS is responsible for predictive and trending analysis of not only technical components, but for an IT service as a unit.

5.1.2.6. Configuration Management in Relation to Physical vCloud Infrastructure

A vCloud is comprised of multiple service layers (SaaS, PaaS, IaaS, and possibly others). All of these vCloud layers are supported by a set of core physical infrastructure components (server, storage, network, and so on). As the vCloud automatically allocates resources, the CMS needs to understand the impact of the changes to the physical core infrastructure. The role of the CMS within the capacity management function is to protect against over-allocation of resources in order to maintain stability and quality of service within a vCloud.

5.1.3 vCloud Configuration Management Considerations

As the vCloud continues to evolve so does configuration management. Newer configuration management systems are emerging in the market. Organizations that are moving to a vCloud model need to continuously reassess the capabilities of existing configuration management systems in relation to the vCloud. Opportunities for process improvement and automation need to be reviewed during every assessment iteration. The following are some considerations to help evaluate CMS capabilities:

- vCloud-aware considerations for a CMS:
 - Is it capable of modeling logical business services within a vCloud?
 - Does it understand both physical and virtual components?
 - Can it map dependencies between components?
 - Can it auto-discover applications?
 - Does it have an understanding of vCloud data models? For example, provider virtual data centers, organization networks, vApps, and so on.
- Flexibility and configuration considerations within a CMS:
 - Does it have a dashboard that displays IT services and their components?
 - Can it integrate with multiple data sources through pluggable connectors?
 - Does it have a reconciliation capability?
 - Does it allow for CI mapping?
 - Does it allow for dependency mapping?
 - Can it manage the various lifecycle states of vCloud components? For example, Dev, UAT, Prod, EOL.
- End-to-end automation:
 - Is it capable of integrating with orchestration engines?
 - Is it capable of integrating with vCloud financial management and chargeback systems?
 - Does it have built in commission and decommission processes?
 - Is it capable of integrating with Asset Management for the allocation and recovery of licenses?

5.1.4 Typical vCloud Configuration Items (CIs)

The following list is starting point for recommended configuration items that should be represented in a CMS for an Infrastructure as a Service vCloud environment.

- VMware vCenter™ Server.
- vCenter Database.
- Datacenter.
- Cluster.
- Resource pool/provider virtual datacenter.
- Organization virtual datacenter.
- Organization network.
- External network.
- VMware ESXi™ host.
- Virtual machine.
- Datastore.

- Virtual distributed switch.
- Port group.
- vApp.

For each configuration item there is an attribute or relationship entry. Each CI has a CI ID, a CI type field, and a CI status field.

5.2 Orchestration Management

Orchestration Management is responsible for gathering and understanding service orchestration workflow requirements, managing their development, testing, and release, and interacting with the vCoE to integrate infrastructure-related automation workflows

5.2.1 Orchestration Management Definition

Orchestration Management is the process responsible for governance and control over the orchestration workflows and the resulting automation within the vCloud. The goal of Orchestration Management is to understand the impact of orchestration workflows on an organization's vCloud, the approvers of and those who benefit from the orchestration, and the interrelations between orchestration and traditional IT service management processes.

5.2.2 Value of Orchestration Management in a vCloud

Orchestration abilities make a vCloud dynamic. This key feature adds to vCloud agility, elasticity, and self-healing properties. Along with the benefits, elasticity also raises some risks. A successful vCloud implementation must focus on delivering consistent quality of services. Orchestration Management adds the layer of control required to achieve this consistency within a vCloud. Another aspect of control is the ability to protect and secure the vCloud. Unwarranted actions within a vCloud cannot be tolerated, so orchestration workflows and actions need to be tightly controlled to protect the vCloud.

The following sections provide information about how orchestration should be controlled in a vCloud. Orchestration is a relatively new feature, and as organizations mature in their management of vCloud environments the role of orchestration management becomes more and more relevant.

5.2.2.1. Orchestration Workflow Creation Control in a vCloud

Before implementing orchestration workflows within a vCloud environment some basic questions need to be answered:

- Who approved the orchestration workflow?
- Why do we need it?
- What impact does this orchestration workflow have on the vCloud environment?
- Who needs to be informed when this workflow is executed?

These questions need to be answered for all orchestration workflows that are built into the vCloud. VMware recommends that two separate teams be involved during development of orchestration workflows. The first team is the Orchestration Management team that focuses on business requirements gathering and business unit negotiations. The second team is the vCloud Center of Excellence team that focuses on technical development of workflows. This provides for the implementation of consistent standards across all orchestration workflows within an organization.

Development of orchestration workflows is very complex. Orchestration engages with multiple internal and external systems in a vCloud environment, so a complete development lifecycle must be followed with dedicated support from the application and business teams.

VMware recommends that appropriate testing be completed at every stage of development, including unit, system, and integration testing before moving orchestration workflows into production. As part of development testing, operational testing that includes performance and scalability scenarios for end-to-end automation processes must also be completed. In many cases, orchestration workflows themselves may be able to withstand new loads, but external or downstream systems may experience a performance impact. A clear roll-back procedure must be established for exceptions to protect against impacting production functions.

5.2.2.2. Orchestration Workflow Execution Control in a vCloud

A vCloud is a dynamic environment where continuous changes are made to improve the quality of services that run on it. Orchestration plays a key part in this agility, allowing for automated actions to be performed as required by vCloud. Orchestration management focuses on vCloud impacts, and avoids adding inflexibility in the environment. VMware recommends that there be control on the execution of orchestration workflows developed for vCloud, with error handling built into the workflows. If there are workflow execution issues, notifications need to be sent to the operations team with appropriate escalations and tiering for alerts.

5.2.2.3. Orchestration Management in Relation to Change Management

Orchestration leads to change in a vCloud environment. As orchestration becomes more mature complex manual tasks are automated. Workflows that lead to changes in business services that directly impact users need to be analyzed in detail before implementation. The Change Advisory Board (CAB) needs to preapprove actions on production applications. Additional controls may also be set, allowing for notification back to the CAB on execution of critical business that impacts orchestration workflows. This must be done in accordance with an organization's change control policies. Business impact should be the main driver for discussion between the orchestration team and CAB. Simple orchestration actions that impact vCloud internal background operations (for example, capacity-related actions), but which do not directly impact a business application or service, should be allowed more flexibility by the CAB and may not need approval.

5.2.2.4. Orchestration Management in Relation to Configuration Management

Orchestration can be used to provision new vApps within a vCloud. Orchestration needs to integrate with and provide status on new or updated configuration items (CIs) to the Configuration Management System (CMS) to provide consistency. Also, the CMS can trigger auto scaling actions for vApps executed by an orchestration workflow to provide quality of service.

Another aspect of the relationship between orchestration and configuration management is the understanding of the physical layer that supports the vCloud environment. In mature implementations, orchestration can interact with the configuration management layer to understand gaps within the physical layer and remediate as needed to maintain environment stability (for example, adding new storage capacity).

5.2.2.5. Orchestration Management in Relation to Security

vCloud-based services are focused on business users, allowing them to request new services directly via the service catalog. Orchestration is critical to such automation, and should have an API to communicate with external systems. Orchestration adds flexibility within a vCloud. With flexibility comes a requirement to add controls such that there are no security risks or exposure for the organization. Because the orchestration workflows have access rights to multiple systems, the orchestration workflow code needs to be protected. Encryption controls such as Set Digital Rights management need to be enabled while moving workflow code packages within servers. Also, access to the orchestration servers must be limited. VMware recommends that the vCloud Center of Excellence exclusively control and manage access on these servers.

5.2.2.6. Orchestration Management in Relation to Audit and Compliance

As noted, orchestration workflows enable vCloud to be more dynamic. Automated actions enhance key vCloud functions such as provisioning and self-service. Though enhanced automation is very beneficial, it poses a challenge to an organization that is bound by tight audit, regulatory, and compliance rules. VMware recommends that orchestration engines running the orchestration workflows are centralized within an organization. Centralized error handling and logging is recommended for all workflows. Reporting features that checkpoint all workflow actions must be enabled for audit compliance. Centralized orchestration engines also enhance an organization's problem management and root-cause analysis capabilities.

Some of the recommended orchestration management principles cannot currently be fully automated and require manual configuration actions based on individual client needs. VMware continues to improve existing libraries and as vCloud implementations mature, more packaged orchestrations with control and governance features should be available for clients to download.

5.3 Availability Management

Availability Management focuses on making sure that the level of availability provided for all vCloud service offerings meets or exceeds the agreed service level requirements in a cost-effective manner. Managing availability within a vCloud environment depends on VMware vCloud Director component availability as well as the resilience of the underlying infrastructure. VCD works transparently with VMware vCenter Server to provision and deploy virtual machines on hosts. Therefore, it is imperative to architect redundancy and protect the infrastructure components. Provisioned virtual machines can be protected by VMware vSphere High Availability (HA). Virtual machines can also be protected using backup tools within the guest OS or vStorage API (vStorage APIs for Data Protection (VADP)-based) applications. See Section 5.4.2, Backup and Restore of vApps, for additional information.

At this time, virtual machines provisioned by VCD cannot be protected by VMware Fault Tolerance (FT) or VMware vCenter Site Recovery Manager™ (SRM).

5.3.1 Uptime SLAs

VMware vCloud components support a 99.9% uptime SLA “out-of-the-box.” This may be sufficient for noncritical applications or applications that are inherently highly available. For vCloud, uptime SLAs typically require verification that:

- End customer workloads are running.
- End customer workloads are accessible (via the vCloud portal and API, as well as through remote access protocols).

In some cases a provider (either an external service provider or internal IT) may want to increase the vCloud uptime SLA. VMware can only control the resiliency of its vCloud platform components and provide recommendations to mitigate single points of failure (SPOF) in the underlying infrastructure. A provider can eliminate SPOF by providing redundancy. For example:

- Redundant power sourced from multiple feeds, with multiple whips to racks, as well as sufficient backup battery and generator capacity.
- Redundant network components.
- Redundant storage components:
 - Storage design needs to be able to handle the I/O load as well. Customer workloads may not be accessible under high disk latency, file locks, and so forth.
 - Storage design should also be tied to business continuity and disaster recovery plans, possibly including array-level backups.
- Redundant server components (multiple independent power supplies, network interface cards (NICs) and, if appropriate, host bus adaptors (HBAs)).
- Sufficient compute resources for a minimum of N+1 redundancy within a vSphere high availability cluster including sufficient capacity for timely recovery.
- Redundant databases and management.

Appropriate change, incident, problem and capacity management processes must also be well defined and enforced to make sure that poor operational processes do not result in unnecessary downtime. In addition to a redundant infrastructure, employees or contractors responsible for operating and maintaining the environment and the supporting infrastructure must be adequately trained and skilled.

For more detailed information about increasing vCloud component resiliency, refer to the “vCloud Availability Considerations” section of *Architecting a VMware vCloud*.

5.4 Continuity Management

Continuity Management for vCloud focuses on making sure that the vCloud-based service offerings, as well as the infrastructure upon which they are hosted, can be resumed within an agreed timeframe in the case of a disruption of service—regardless of whether the outage is at the vApp level or an entire vCloud environment instance. In this context, VMware defines two components to Continuity Management: Disaster Recovery (strategic), and vApp Backup and Restore (tactical).

5.4.1 Disaster Recovery

Disaster Recovery (DR) focuses on the recovery of systems and infrastructure after an incident that interrupts normal operations. A disaster can be defined as partial or complete unavailability of resources and services, including software, the virtualization layer, the vCloud layer, and the workloads running in the resource groups. Different approaches and technologies are supported, but there are at least two areas that require disaster recovery: the management cluster and consumer resources. Different approaches and technologies are supported.

5.4.1.1. Management Cluster Disaster Recovery

Good practices at the infrastructure level lead to easier disaster recovery of the management cluster. This includes technologies such as HA and DRS for reactive and proactive protection at the primary site. VMware vCenter Heartbeat™ can also be used to protect vCenter Server, specifically, at the primary site. For multi-site protection of virtual machines, VMware vCenter Site Recovery Manager is a VMware solution that works well for this use case, because the management virtual machines are not part of a vCloud instance of any type (they run the vCloud instances).

5.4.1.2. vCloud Consumer Resources Disaster Recovery

The vCloud infrastructure can be failed over to an alternate site, but vCenter Site Recovery Manager is not supported. Manual procedures can be applied as long as vApp metadata is saved, configuration information is matched between the primary site and the recovery site, and the documented steps are validated.

Though SRM is vCenter Server-aware, SRM is not vCloud Director-aware. Without the collaboration between vCloud Director and SRM, the underlying mechanisms that work to synchronize virtual machines cannot work to keep vCloud Director in sync as well—thus, the recovery of vCloud Director can be problematic. Though it is possible to architect a solution where one site's total environment (100% of the operational parameters of that site including IP addressing, start-up order of dependent systems and the like) can be duplicated to another site, it would be very difficult to implement and maintain.

5.4.2 Backup and Restore of vApps

This section focuses on handling of backup and restore procedures for the vApps that are deployed into the vCloud. Traditional backup tools do not capture the required metadata associated with a vApp, such as owner, network, and organization. This results in recovery and restoration issues. Without this data, recovery must include manual steps and configuration attributes to be manually reentered.

Within a vCloud environment, a vApp can be a single virtual machine or group of virtual machines, treated as one object. Backup of vApps on isolated networks must be supported. Identifying inventories of individual organizations becomes challenging based on current methods that enumerate the backup items using vSphere, which uses Universally Unique Identifiers to differentiate objects. vCloud Director uses object identifiers.

For backing up and restoring vApps, VMware recommends the use of VMware vSphere® Storage APIs – Data Protection based backup technologies. This technology has no agents on guest operating systems, is centralized for improved manageability, and has a reduced dependency on backup windows.

Guest-based backup solutions may not work in a vCloud because not all virtual machines are accessible by network. Also, virtual machines may have identical IP addresses and that can cause problems. Therefore, backups of vCloud vApps require a virtual machine-level approach.

Use the full name and computer name fields to specify realistic names that will help describe the virtual machines when deploying virtual machines (as part of a vApp). If this is not done, the generic information in these fields can make it difficult to specify individual virtual machines. vApps and virtual machines that are provisioned by vCloud Director have a large GUID template_name—so that many virtual machines could appear to be very similar, making it difficult for a user or administrator to identify and ask for a specific virtual machine to be restored.

5.4.2.1. VMware Solutions

VMware Data Recovery is a vStorage APIs for Data Protection-based solution. Other vStorage APIs for Data Protection-based backup technologies are available from third-party backup vendors. Currently, due to the Universally Unique Identifier versus object identifier issue, VMware Data Recovery cannot be used with VMware vCloud Director.

There are a few requirements to address for backup of vCloud workloads. VMware recommends that clients validate the level of support provided by the vendor to make sure client requirements are supported. Table 3 provides a list of vCloud vApp requirements to ask your vendor about.

Table 3. vCloud vApp Requirements Checklist

vApp Requirement	Detail
vStorage API Data Protection integration	<input type="checkbox"/> vStorage API Data Protection provides change-block tracking capability to reduce backup windows. <input type="checkbox"/> Integration to enable backup of isolated virtual machines and vApps. <input type="checkbox"/> Integration with vStorage API Data Protection to provide LAN-free and server-free backups to support better consolidation ratios for vCloud and the underlying vSphere infrastructure. <input type="checkbox"/> Use of the virtual machine Universally Unique Identifier versus virtual machine name will support multitenancy and avoid potential name space conflicts.

vApp Requirement	Detail
vCloud Director integration	<input type="checkbox"/> Interface support for vCloud provider administrator teams. In the future, consumer (organization administrator and users) access may be provided by some vendors. <input type="checkbox"/> Include vCloud metadata for the vApps. This includes temporary and permanent metadata per virtual machine or vApp. This is required to make sure that recovery of the virtual machine or vApp has all data required to support resource requirements and SLAs.
vApp requirements	<input type="checkbox"/> Provide vApp granularity for backups. Support backup of multitiered vApps (for example, a Microsoft Exchange vApp that has multiple virtual machines included. Backup selection of the Exchange vApp would pick up all the underlying virtual machines that are part of the main vApp). This capability is not available today, but is being developed by vendors.

5.4.2.2. Challenges

Challenges associated with backing up and restoring a vCloud are:

- vApp naming poses conflict issues between tenants.
- vApp metadata required for recovery.
- Multi-object vApp backup (protection groups for multitiered vApps).
- Manual recovery steps in the vCloud.
- Support for backup of vApps on isolated networks or with no network connectivity.
- Enumeration of vApps by organization for use by the organization administrator.
- Enumeration of vApps by organization and provider for use by the organization provider.
- User initiated backup/recovery.
- Support of provider (provider administrator) and consumer (organization administrator and user).

For a more detailed treatment of vCloud Business Continuity, see Appendix F: Business Continuity.

6. vCloud Infrastructure Management

vCloud Infrastructure Management includes Security and Compliance Management, Capacity Management, Performance Management, and monitoring.

6.1 Security and Compliance Management

The following sections describe Security Management for vCloud access management and logging.

6.1.1 Access Management – User Access Security

The service provider has to configure a directory service for vCloud Director. This is also true for a private vCloud.

Authentication and authorization mechanisms built into vCloud Director provide user access security for vCloud resources. vCloud Director can be configured to integrate with a directory service (LDAPv3) such as Active Directory, OpenLDAP, or Kerberos v5. Refer to the *VMware vCloud Director Administration Guide* (http://www.vmware.com/support/pubs/vcd_pubs.html) for more information about how to set up the naming services (LDAPv3), Active Directory, or OpenLDAP and Kerberos v5 integration.

User authorization within vCloud Director is controlled through role-based access control (RBAC). Also refer to the latest VMware *vCloud Director Administrator's Guide* for additional information on permissions, roles, and default settings.

6.1.2 Log Management

Logs should be available for customers and their providers in a vCloud for numerous reasons, including:

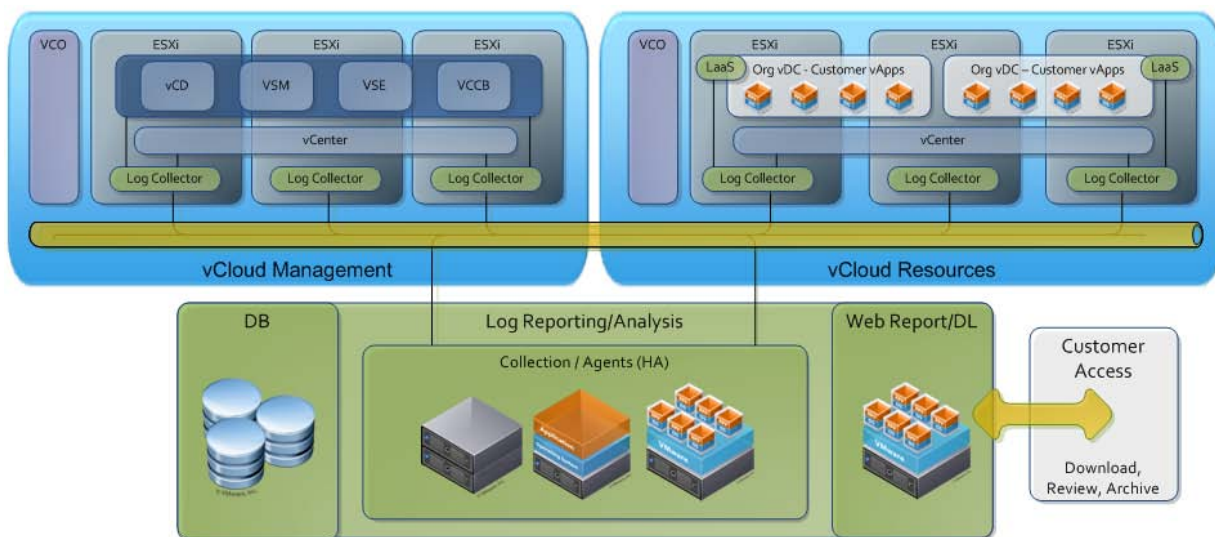
- **Regulatory Compliance** – Collect logs to make them available for analysis, security review, and compliance requirements as described in Appendix B: Compliance Considerations. Individual logs can be used to satisfy specific compliance controls; for example, a user access log can be used to show an audit trail for user access success and failure.
- **Customer Requirements** – End customers (tenants) can retrieve logs that pertain to their environment to satisfy their own requirements, many of which, such as compliance, will probably be similar to provider requirements.
- **Operational Integrity** – Operational alerts should be defined so that specific logs trigger notifications for further remediation. This is typically a backup alert, and is secondary to monitoring.
- **Troubleshooting** – Closely related to operational integrity, troubleshooting can be performed using logs. For example, VMware vShield Edge™ logs can show whether or not a specific external connection request is being passed through the firewall or via Network Address Translation (NAT) by the firewall.

6.1.2.1. Logging Architectural Considerations

- Redundancy – Many components rely on syslog for logging events. Syslog is a UDP-based protocol that lacks delivery guarantees. To facilitate delivery:
 - Verify that infrastructure components have physically and logically redundant network interfaces.
 - Send logs to more than one syslog target.
 - If only one syslog target is possible, VMware recommends logging to a local syslog daemon that is configured to retransmit to two remote syslog targets. VMware vCloud Director 1.5 supports only a single syslog target for its activity logs.
 - Place log receivers on DRS-enabled hosts if possible so that vCenter can restart them in case of failure.
- Scalability – vCloud infrastructure components generate a relatively low level of logs for provider infrastructure. Customer components, especially vShield Edge firewalls, can generate a very high volume of logs. Collecting logs on IOPS performance is critical. Collecting logs on CPU performance is negligible, but they are needed for analysis. It is strongly recommended that logs be collected to dedicated log partitions on collection servers.
- Reporting:
 - Logs need to be available to customers. Customers should be able to download in raw format all vCloud Director and vShield Edge logs that pertain to their organizations and networks. Logs with customer identifiers should be flagged or indexed for retrieval.
 - Customer activity in vCloud Director generates logs that are flagged with their organization identifier.
 - vShield Edge devices can be uniquely identified. vCloud Director 1.5 can deploy vShield Edges with descriptive, unique names, such that Security Event and Incident Management (SEIM) products can definitively correlate log messages from vShield Edge appliances to the organization that generated them.

Figure 8 shows an example logging architecture.

Figure 8. Architectural Example Drawing



6.1.2.2. Logging as a Service

Logging as a service can be done with customer collection and forwarding to provider servers for analysis and reporting, or with customer collection, reporting, and analysis in the customer environment and provider logs forwarded to the customer environment.

- Customer collection forwarding to provider:
 - Pros:
 - Logs can be sent directly to collector even on customer private IP space.
 - Resources can be allocated at the customer level for collection, allowing more granular scaling of collection.
 - Cons:
 - More difficult to scale analysis; challenges correlating customer activity to storage consumption.
 - Collection nodes still required even though utilization will be low.
 - Most of the resource consumption is on the storage and analysis side, so the resources billed using the IaaS model will be minimal.
- Customer collection, with provider logs forwarded into customer environment:
 - Pros:
 - Distributed analysis relies on general vCloud resources and can scale.
 - Customer can employ their own analysis tools to organize and report on the data, or use a provider-supported package or appliance.
 - Cons:
 - Provider needs duplicate copy of infrastructure logs for provider purposes.
 - Transmission of logs to the customer environment requires connectivity; either Internet or a provider service network and inbound traffic through a firewall into the customer environment, adding risks.

6.2 Capacity Management

Capacity Management focuses on providing vCloud capacity to meet both existing and future needs in support of vCloud service offerings.

Management cluster sizing is fairly predictable, with the main variables being the number of vCloud Director cells and the size of the vCloud Director database. Initial sizing guidelines for the management cluster are provided in *Architecting a VMware vCloud*.

The vCloud consumer resources have unpredictable usage, and thus should be sized by making an estimate of the initial capacity required, and by employing capacity management techniques. Capacity management techniques predict future usage needs based upon past usage trends.

6.2.1 Ongoing Capacity Management

One of the key benefits of implementing a vCloud is the ability for service provider customers (equivalent to public vCloud providers or private vCloud internal IT) to rapidly provision vApps into the vCloud environment. The goal of capacity management is to make sure that sufficient capacity exists within the vCloud infrastructure to meet the current and future needs of the service provider customers under normal circumstances. Sufficient reserve capacity must be maintained within the vCloud infrastructure to prevent vApps from contending for resources, and thus potentially breaching agreed services levels.

As vApps are provisioned and consumed within the vCloud infrastructure, available capacity is reduced and additional capacity must be procured and provisioned. Capacity Management processes should be instituted to make sure appropriate resources are available to support the service level requirements associated with vApp provisioning and performance. Proper capacity management also prevents costly over-provisioning of hardware resources by balancing high resource utilization with agreed-upon levels of performance.

As the vCloud is consumed, additional capacity must be added to the vCloud consumer resources to allow for anticipated future demand while preserving sufficient headroom. To predict future capacity needs, analyze current capacity usage and trends to determine growth rates as well as estimate future needs, largely coming from new consumers and projects.

VMware vCenter CapacityIQ™ is a tool that can be used to monitor and predict capacity usage and requirements. In a vCloud environment, CapacityIQ can provide details on capacity at the virtual machine and host levels, as well as the provider virtual datacenter-level (because, following VMware best practices, a provider virtual datacenter is equivalent to a vSphere cluster), but it currently does not provide insight at the organization virtual datacenter level.

For further information on capacity planning and management, refer to the latest *VMware vCenter CapacityIQ Installation Guide* (http://www.vmware.com/support/pubs/ciq_pubs.html). Also see Appendix C: Capacity Planning for guidance on how to manually calculate capacity requirements and forecast capacity.

6.3 Performance Management

A key IT paradigm shift enabled by vCloud computing is that of delivering services to end users. Ultimately these are business user services; service offerings directly consumed by end users that also include technology services. Technology services are not directly consumed by end users, but represent the underlying infrastructure components which, when considered together, enable the consumption of business user services. One of the key challenges at the vCloud Operations Management layer concerns reorienting operations processes around sustaining the delivery of a business user service as a whole, as opposed to focusing on operating the underlying infrastructure components as discreet entities. This challenge manifests itself across several key operations processes, not the least of which is Performance Management.

It is no longer sufficient to monitor and manage the performance of individual IT infrastructure components. With the adoption of vCloud computing, IT must monitor and manage the performance of the business user service and, more precisely, the business user's service from the end-user's perspective. The focus is on how the service is performing for the end user, as opposed to considering individually the network performance or server performance upon which a particular application is hosted.

Performance Management from the service end-user perspective can have far reaching ramifications. These might range from deploying remote probes that can simulate an end user's interaction with a service while tracking performance metrics, to more integrated performance monitoring, filtering, and analytics solutions that "understand" what comprises a business user service, monitors the performance of the components in the context of the complete service, and can take automated action based on early warning (predictive) "smart alerts." Along with these more sophisticated tools come the people, process, and process interrelationship modifications that are required to efficiently and effectively complement them.

Based on this service orientation, VMware provides the following considerations and guidance for vCloud Performance Management from a people, process, and tools perspective.

6.3.1 People Perspective

- Within the Operations Management layer, assign a Performance Management owner who has the following responsibilities:
 - Overall responsibility for the performance management process.
 - Regular interaction with other, related process owners.
 - Establishing agreed to Performance Management Key Performance Indicators (KPIs).
 - Tracking and reporting on KPIs.
 - Working with the vCloud Center of Excellence architect to design a Performance Management monitoring solution that is business user service-oriented.
- Establish a training plan for the tools to be used by the Performance Management owner and other designated individuals.

6.3.2 Process Perspective

- Define a business user service-focused performance management process.
- Include an interface with Service Level Management for performance reporting in support of Service Level Agreements.
- Include an interface with Capacity Management for forward-looking capacity requirements.
- Include an interface with event, incident, and problem management to provide Tier 3 incident resolution support.

6.3.3 Tool Perspective

VMware recommends the following for a vCloud Performance Management tool:

- Integrates with vCloud Director, the underlying vSphere infrastructure, and third-party monitoring tools for a holistic performance view.
- Aggregates, correlates, and presents performance data in the context of business user service health and performance.
- Simplifies management of alarms by using a combination of static and dynamic thresholds based on self-learning.
- Capable of predictive analysis resulting in alerts of impending performance degradation.
- Correlates cause and effect for performance problem resolution.
- Provides visibility and drills down from business user service to component level.

Though no tool is currently available in the industry that provides all of these capabilities out-of-the-box, the VMware vCenter Operations™ Enterprise management solution, in combination with the vCloud Director adapter, provides capabilities to assess the health of your vCloud infrastructure from a performance perspective. This is visualized using *health scores*, *heat maps*, and *health trees* at various aggregation levels—from provider virtual datacenters and organizations to virtual machines.

6.4 Monitoring

Monitoring the components of a vCloud Director implementation is essential to the health of a vCloud environment, and is necessary to maintain capacity and meet service level agreements. This section provides recommendations on what systems and associated objects to monitor, and readily available tools that can be used to extract health-related metrics. Details of specific limits or thresholds are not identified here as they are available in the product documentation. This document does not attempt to provide specifics for setting up a monitoring solution as various service providers and enterprises may have very different monitoring solutions in place to be integrated.

6.4.1 Management Cluster

The best practices for monitoring the management cluster components are the same as the best practices for monitoring vSphere components. As part of this, a centralized monitoring tool such as VMware vFabric™ Hyperic® HQ Enterprise can be used to monitor the core objects (Oracle Server, SQL Server, Active Directory Server, DNS Server, Red Hat Enterprise Linux Server, and Windows Server) that are needed to run a vCloud environment. A customer can use SNMP and SMASH to monitor the hosts on which the vCloud Director cells are installed and running, but the vCloud Director application itself cannot be monitored by SNMP or SMASH. However, SNMP can be integrated from vCenter. Alternatively, cells can be monitored through integration with a third-party monitoring platform via JMX Beans. Note that JMX Beans monitoring is only the start. The vCloud and vSphere APIs provide a significant amount of component, resource, and activity metrics that can be used for health and capacity management.

6.4.2 Cloud Consumer Resources and Workloads

The best practices for monitoring the vCloud consumer resources and workloads are the same as for monitoring vSphere. However, there are additional vCloud-specific considerations for VMware vShield Edge and vCloud consumer workloads.

6.4.2.1. vShield Edge

vShield Edge appliances are self-contained environments that are stateless in nature. There is a “health check” API call that can be made to a vShield Edge appliance to determine if it is functioning correctly. If the API returns negative, initiate a reboot of the vShield Edge device. At the time of reboot, configuration information is updated from the VMware vShield Manager™ and the vShield Edge device continues to function properly.

6.4.2.2. vCloud Consumer Workloads

It may be desirable to monitor workloads provisioned by vCloud consumers. vCloud Director does not provide any built-in monitoring of workloads for availability or performance. Several third-party solutions are available to monitor vSphere resources and workloads running on vSphere; however, not all of these solutions may work all of the time when vCloud Director is in use. Isolated networking in vApps may prevent monitoring tools from acquiring the performance or availability information of a vApp. Furthermore, vApps may be provisioned and de-provisioned or power-cycled at any time by a vCloud consumer and these actions may create false positives in the monitoring environment. Until there are solutions in the market that are fully integrated with vCloud Director, it may be difficult to provide detailed monitoring for vCloud consumer workloads.

Appendix A: vCloud Director Cell Monitoring

The following table represents a subset of MBeans that can be used for improving the monitoring performance of a vCloud instance.

Table 4. MBeans Used to Monitor vCloud Cells

Local user sessions	
Mbean	com.vmware.vcloud.diagnostics.UserSessions
Description	Local (cell) user session statistics
Cardinality	1
Instance ID	n/a
Attribute	Description
totalSessions	Total number of sessions created on this cell
successfulLogins	Total number of successful logins to this cell
failedLogins	Total number of failed login requests to this cell
Global user sessions	
Mbean	com.vmware.vcloud.GlobalUserSessionStatistics
Description	List of active user sessions by organization.
Cardinality	1
Instance ID	n/a
Attribute	Description
organization	Database ID of the organization
active	Number of active sessions
Open_Session	Number of open sessions
Data access diagnostics	
Mbean	com.vmware.vcloud.diagnostics.DataAccess
Description	Local (cell) user session statistics
Cardinality	1
Instance ID	Conversation
Attribute	Description
lastAccessInfo.objectType	Object type of the last database object accessed
lastAccessInfo.accessTime	Time taken to access the last database object accessed
worstAccessInfo.objectType	Object type of the worst (slowest) database object access
worstAccessInfo.accessTime	Time taken by the worst (slowest) database object access
Database Connection Pool	

Mbean	com.vmware.vcloud.datasources.globalDataSource
Description	Statistics and configuration information about the database connection pool. This information is currently specific to the database JDBC driver being used (Oracle).
Cardinality	1
Instance ID	
Attribute	Description
abandonedConnectionTimeout	
availableConnectionsCount	
borrowedConnectionsCount	
connectionHarvestMaxCount	
connectionHarvestTriggerCount	
connectionPoolName	
connectionWaitTimeout	
databaseName	Database connection database name (SID)
dataSourceName	
fastConnectionFailoverEnabled	
inactiveConnectionTimeout	
initialPoolSize	
loginTimeout	
maxConnectionReuseCount	
maxIdleTime	
maxPoolSize	Maximum number of connections allowed in the pool
maxStatements	
minPoolSize	Minimum number of connections that will exist in the pool
networkProtocol	Network protocol used by JDBC driver
ONSConfiguration	
portNumber	Database connection port number
SQLForValidateConnection	
timeoutCheckInterval	
timeToLiveConnectionTimeout	
URL	Database connection URL
user	Database connection username
validateConnectionOnBorrow	

VIM Operations	
Mbean	com.vmware.vcloud.diagnostics.VlsiOperations
Description	Local (cell) user session statistics
Cardinality	1 per VIM end-point (VC or host agent)
Instance ID	VIM end-point URL
Attribute	Description
ObjectType.MethodName.httpTime	The total network round-trip time taken to make the "MethodName" call on object of type "ObjectType" in the VIM endpoint.
Presentation API Methods	
Mbean	com.vmware.vcloud.diagnostics.VlsiOperations
Description	Local (cell) user session statistics
Cardinality	1 per presentation layer method
Instance ID	method name
Attribute	Description
currentInvocations	Currently active invocations
totalFailed	Total number of failed executions
totalInvocations	Total number of invocations over time
executionTime	Total time taken to execute
Jetty	
Mbean	com.vmware.vcloud.diagnostics.Jetty
Description	Web server request statistics
Cardinality	2:1 for REST API and 1 for UI
Instance ID	"UI Requests" for UI, "REST API Requests" for REST API
Attribute	Description
Active	Number of Web requests currently being handled
REST API	

Mbean	com.vmware.vcloud.diagnostics.VlsiOperations
Description	Local (cell) user session statistics
Cardinality	1 per operation stage/granularity: RoundTrip, BasicLogin, Logout, Authentication, SecurityFilter, ConversationFilter, JAXRSServlet. RoundTrip is the most interesting, as it represents the overall REST API performance.
Instance ID	One of: RoundTrip, BasicLogin, Logout, Authentication, SecurityFilter, ConversationFilter, JAXRSServlet
Attribute	Description
currentInvocations	Currently active invocations
totalFailed	Total number of failed executions
totalInvocations	Total number of invocations over time
executionTime	Total time taken to execute
Task Execution	
Mbean	com.vmware.vcloud.diagnostics.TaskExecutionJobs
Description	Statistics about long running tasks
Cardinality	1 per task
Instance ID	Name of task
Attribute	Description
currentInvocations	Currently active invocations
totalFailed	Total number of failed executions
totalInvocations	Total number of invocations over time
executionTime	Total time taken to execute
Query Service (UI)	
Mbean	com.vmware.vcloud.diagnostics.QueryService
Description	Presentation layer query service statistics
Cardinality	1 per query
Instance ID	query name
Attribute	Description
currentInvocations	Currently active invocations
totalFailed	Total number of failed executions
totalInvocations	Total number of invocations over time
executionTime	Total time taken to execute
returnedItems	Number of items returned by successful query executions
VC Task Manager	

Mbean	com.vmware.vcloud.diagnostics.VcTasks
Description	VC task management statistics
Cardinality	1
Instance ID	
Attribute	Description
successfulTasksCount	total successful tasks
failedTasksCount	total failed tasks
waitForTaskInvocationsCount	total invocations of VIM "wait for task"
completedWaitForTasksCount	total completed task waits
historicalTasksCount	total historical task updates received
vcRetrievedTaskCompletionsCount	total task completions received
taskCompletionMessagesPublishedCount	total task completion messages published on message bus
taskCompletionMessagesReceivedCount	total task completion messages received on message bus
success_elapsedTaskWaitTime	time elapsed for successful tasks
failed_elapsedTaskWaitTime	time elapsed for failed tasks
VIM Inventory Update Processing – Object Update Statistics	
Mbean	com.vmware.vcloud.diagnostics.VimInventoryUpdates
Description	Inventory processing statistics
Cardinality	3: one for ObjectUpdate, PropertyCollector and UpdateSets respectively
Instance ID	ObjectUpdate
Attribute	Description
totalUpdates	Total number of object updates received
totalFailed	Total number of object updates failed to be processed
executionTime	Time taken for updates

VIM Inventory Events

Mbean	com.vmware.vcloud.diagnostics.VimInventoryEvents
Description	VIM inventory event manager statistics. Tracks the frequency of common vCenter events.
Cardinality	1 per folder per VC URL, 1MBean per event name
Instance ID	Event name
Attribute	Description
totalInvocations	Total number of VIM inventory events dispatched since that VCD cell started
totalFailed	Total number of VIM inventory events that were failed to be handled
executionTime	Total time to handle VIM inventory events
VC Object Validations	
Mbean	com.vmware.vcloud.diagnostics.VcValidation
Description	VC object validation statistics
Cardinality	1 global plus 1 per validator
Instance ID	null = global, validator name = per validator
Attribute	Description
totalInvocations	Total number of validation executions
executionTime	Total time spent in validator
totalItemsInQueue	Total items currently queued for validation (global)
objectsInQueue	Total items currently queued for validation (per validator)
objectBusyRequeueCount	Total number of objects re-queued for validation due to object being busy
loadValidationObjectTime	Time taken to load validation object
duplicatesDiscarded	Total number of discarded duplicate validations
VC Object Validation Reactions	
Mbean	com.vmware.vcloud.diagnostics.Reactions
Description	validation reaction statistics
Cardinality	1 global plus 1 per reaction
Instance ID	null = global, reaction name = per reaction
Attribute	Description
totalReactionsFired	Total number of reaction executions
requeueCount	Total number of reactions re-queued due to objects being busy
totalInvocations	Total number of executions of this reaction
executionTime	Total time spent in reaction

failedReactions	Total number of failed reactions
objectRequeueCount	Number of times this reaction was re-queued due to objects being busy
VC connections	
Mbean	com.vmware.vcloud.diagnostics.VimConnection
Description	Local (cell) user session statistics
Cardinality	1 per VC
Instance ID	"VC-VcInstanceId" where VcInstanceId is an integer identifying the vCenter instance
Attribute	Description
Connected Count	Total successful connections
Disconnected Count	Total disconnections
Start Count	Total number of times the VC listener was started
UI Vim Reconnect Count	Total number of times the VC was reconnected through the UI
ActiveMQ	
Mbean	com.vmware.vcloud.diagnostics.ActiveMQ
Description	Active MQ (message bus) statistics
Cardinality	1 global and 1 per peer vCloud Director cell (each cell other than the current one)
Instance ID	"Global" = global statistics "to_cellName_cellPrimaryIp_cellUUID"=per cell
Attribute	Description
lastHealthCheckDate	Last time health check was performed (date/time)
messageRoundTripDurationMs	Time taken for an echo message to be sent and returned (ms)
isHealthy	Health of connection to peer cell in the case of the per-cell Mbean, overall message bus connection health in the case of the global Mbean (true/false)
timedOutMessages	Total number of echo messages for which no reply was received within the timeout (controlled by the activeMonitorCheckDelayMs config parameter, default 10 minutes)
sendErrors	Total number of failed echo message sends (messages)
corruptedOrBadEchoMessages	Total number of corrupted/bad echo messages received (starts)_ (messages)
Transfer Server	
Mbean	com.vmware.vcloud.diagnostics.VlsiOperations

Description	Transfer server statistics
Cardinality	1
Instance ID	
Attribute	Description
successfulPuts	Number of items successfully transferred (transfer items)
failedPuts	Number of items that were failed to be transferred (transfer items)
successfulUploads	Number of successful upload operations (uploads)
acceptedQuarantinedTransferSessions	Number of quarantined transfers which were accepted (quarantined items)
rejectedQuarantinedTransferSessions	Number of quarantined transfers which were rejected (quarantined items)
expiredTransferSessions	Number of transfer sessions which timed out (transfer sessions)

Appendix B: Compliance Considerations

Compliance

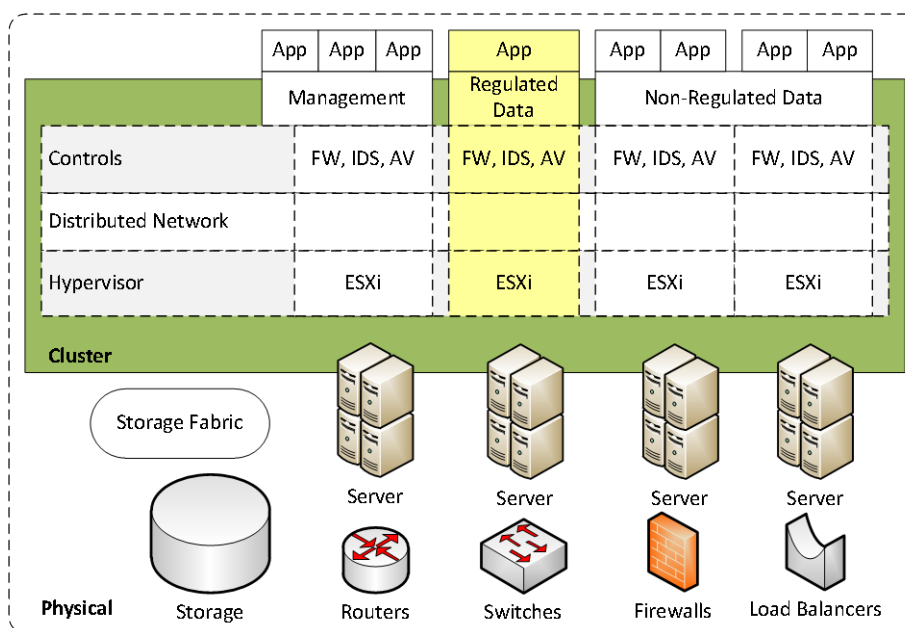
Audit concepts such as segmentation and monitoring applied to a vCloud environment reveal new challenges. Elasticity may break old segmentation controls and the ability to isolate sensitive data within a rapidly growing environment. Role-based access controls and virtual firewalls must also demonstrate compatibility with audit requirements for segmentation, including detailed audit trails and logs. Can a provider guarantee that an offline image with sensitive data in memory is accessible only by authorized users, and can a log tell who accessed it and when? Multiple admin-level roles are necessary for vCloud resource management.

The complexity of cloud environments, coupled with new and different technology, requires careful audits to document and detail compliance. Table 5 lists common audit concerns within the vCloud.

Table 5. Audit Concerns Within the vCloud

Concern	Detail
Hypervisor	An additional layer of technology is present in every vCloud and may present an attack surface. It introduces a layer between the traditional processing environment and the physical layer, which brings a new level of communication with layers above and below it.
Segmentation and isolation	Any environment may expose sensitive data when not configured and monitored properly; physical and logical isolation has always been an audit concern. The ease and speed of change to a virtualized environment within cloud computing, often called <i>elasticity</i> , makes the setup and review of segmentation controls even more relevant to compliance through isolation.
Different/multiple primary functions per host	The vCloud environment can make more efficient use of hardware, but it increases the proximity of information in transit and at rest. Some compliance standards explicitly require one primary function per server (or virtual server), as illustrated in the following figure.

Figure 9. One Primary Function per Server



Enforcement of least privilege	In a vCloud environment, remote network access is the only available path offered to customers to manage their environment. Instead of physical access audits for equipment installation and modification, virtual system management software must be audited.
Machine state and migration	The ability of systems to quickly change and move within a vCloud environment gives auditors a need to track authorization and related change controls. Separate and isolated networks should be used for data migration that is in the clear to avoid exposure of sensitive information.
Data is much less permanent	Cloud environments make extensive use of short-lived instances. Virtual machines may have a lifecycle far shorter than physical systems as they are easy to provision and repurpose. Systems also share data across large arrays in swap space. Permanence of data is also affected by environments that push as much storage as possible through high-speed memory to avoid the latency of spinning disks.
Immaturity of monitoring solutions in vCloud environments	Customers need audit trails and views unique to their own use of the vCloud environment, which also supports incident response and investigations. Providers have to extend and develop log management and monitoring solutions to meet regulatory and client requirements for the vCloud environment.

Use Cases: Why Logs Should be Available

It is important to monitor and record events in order to mitigate damage and prevent future attacks. An audit log enables an organization to verify compliance, detect violations, and initiate remediation activities. It can help detect attempts, whether successful or not, for unauthorized access, information probes, or disruption.

Log Purposes

Logs are a foundation of many controls used to achieve internal requirements as well as regulatory compliance. They are the technical solution to track and record changes and incidents as they form an audit trail. Logs offer the following benefits:

- **Compliance requirements** – Logs are required for all compliance regulations to assist with control auditing as well as breach review, analysis, and response. Specific types of logs often can be matched with specific compliance controls. For example, the authentication log can demonstrate access controls allowed to only authorized users.
- **Customer requirements** – End customers can retrieve logs that pertain to their environment in order to meet their own requirements.
- **Operational integrity** – Operational alerts should be defined for logs to trigger notifications for remediation. This is frequently set up as a backup alert, secondary to monitoring. A storage array that goes offline generates error messages in the logs, which can be used to alert administrators.
- **Troubleshooting** – Closely related to operational integrity, logs are essential for troubleshooting. For example, the use of vShield Edge logs can show whether a specific external connection request is being passed through or NATted by the firewall.

Frequency of Review

Logs should be reviewed daily for unauthorized or unusual and suspicious activity on all systems and especially those that handle intrusion detection, authentication and authorization. This requires review and verification of logs to establish baselines of normal operations, such as monitoring access and authorization (every login and logout) from the console, network, and remote access points. More frequent and routine log analysis for security often helps give early identification of system configuration errors, failures, and issues that can impact SLAs.

Minimum Data Types

The following are the minimum set of data types required to adequately log vCloud environment activity for regulatory compliance:

- User (including system account) access.
- Action taken.
- Use of identification and authentication mechanisms.
- Start and stop of audit logs.
- Creation or deletion of system-level objects.

The audit trail entries recorded for each event must include the following details:

- Identification (ID).
- Type of event.
- Date and time.
- Success or failure.
- Origination of event.
- ID of affected data or component.

Retention

Daily review of logs alone may not be sufficient to detect incidents—they also must be retained for a period consistent with “effective use” and legal regulations. The laws for log retention range from one year to more than twenty. Therefore, log archives should always be able to provide at least one year of history, typically scheduled to match financial calendar cycles, and a minimum of three months available for immediate response and review in case of an incident.

Example Compliance Use Cases for Logs

The following use cases are a sample of events that benefit from careful logging and monitoring in the vCloud environment. Other examples may include unauthorized services or protocols, remote login success, and certificate changes.

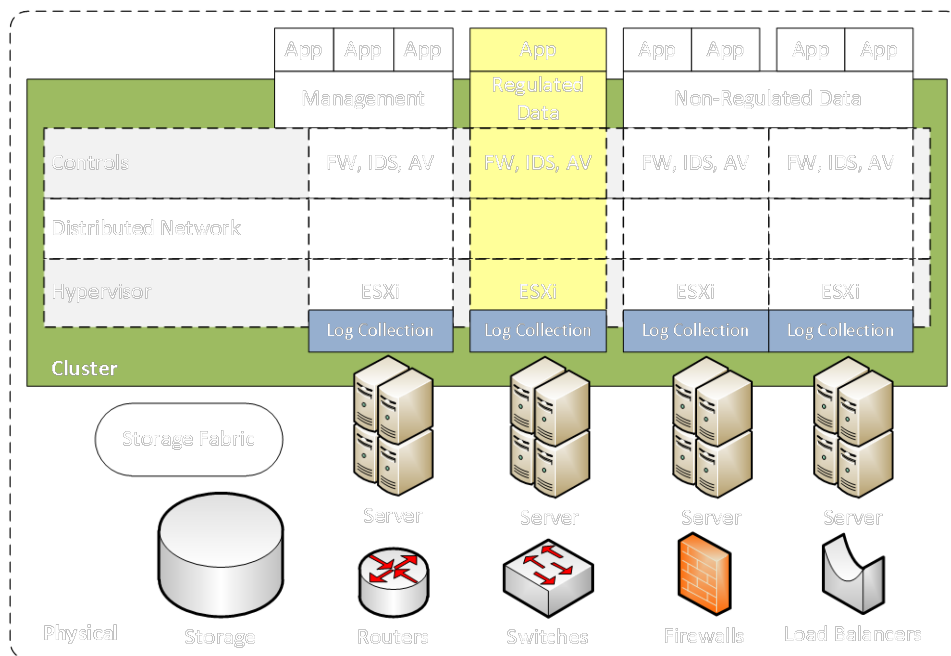
- Shared accounts – An investigation is initiated to review network outages and finds multiple instances of an Administrator account had logged into critical servers before failure. Shared accounts make it very difficult to trace fault to one individual; it is impossible to determine from the logs on that system which person was logged into the user account that made the error. Therefore, usage must be tied to an individual user ID and unique password with correct time to aid in investigations. Systems also should be configured to detect any and all use of generic IDs such as an administrator or root account and trace them to unique identities.

- User account changes – A malicious user finds an un-patched flaw in an environment that allows elevation of privileges. That user then uses system-level privileges to create a new bogus user object from which to launch further attacks. A user object is, for example, a Microsoft Widows Domain or local user account. User object logs can be used to figure out when a name was changed or an account added. This assists in detection of actions without authorization or users trying to hide attacks.
- Unauthorized software – Malware or a new virtual machine instance in the vCloud can be found in system object logs. A system must track system objects that are added, removed or modified. This can be very helpful during installation to monitor system changes caused by software.

VMware vCloud Log Sources for Compliance

Customers should be able to retrieve logs from all areas that are relevant and unique to their organization. Programmatic retrieval should be possible, such as an API to allow for automated queries. Log collection nodes must be added to a vCloud environment, as illustrated in Figure 10.

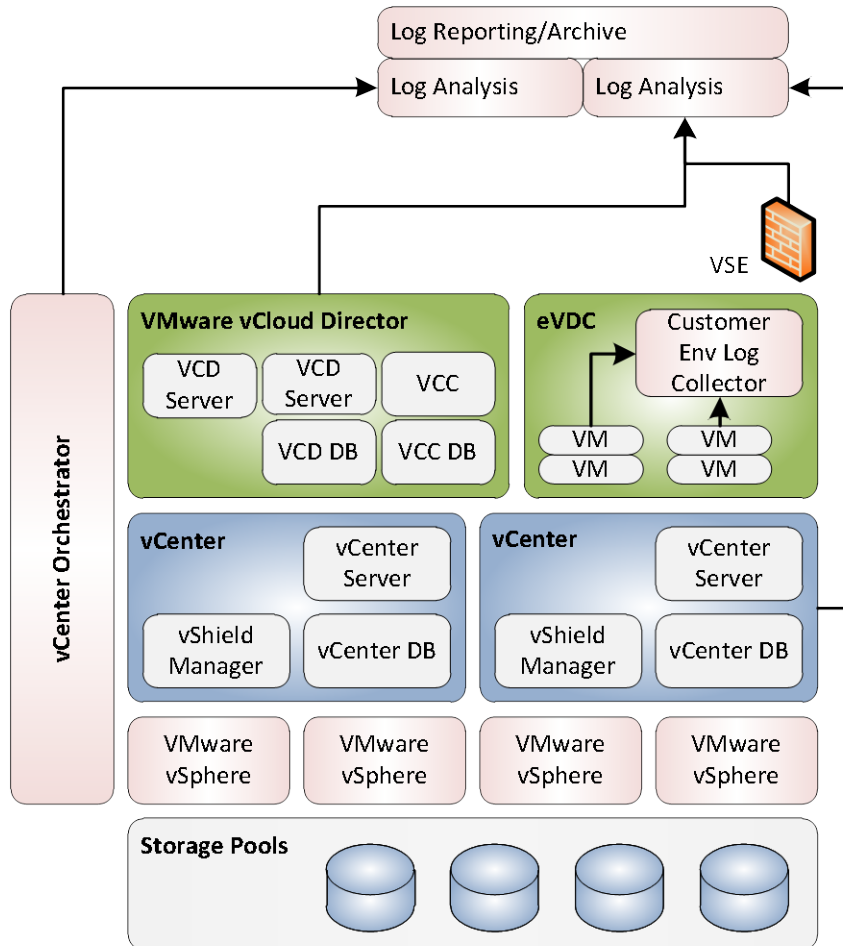
Figure 10. Log Collection in the vCloud Environment



Logs generated by VMware components must be maintained by the provider, but also must be available to tenants. Tenants should be able to download in raw format all vCloud Director and vShield Edge logs that pertain to their organizations and networks. Logs with customer identifiers should be flagged or indexed for retrieval.

Figure 11 illustrates architecture of vCloud components and log collection.

Figure 11. Architecture of vCloud Components and Log Collection



The following table lists the logs to which the vCloud tenant must have access.

Table 6. vCloud Component Logs

VMware Component	Provider Logs	Tenant Logs
VMware vCloud Director	✓	✓
vCenter Server	✓	
vSphere Server (ESXi)	✓	
Chargeback Manager	✓	
vCenter Orchestrator	✓	
vShield Manager	✓	
vShield Edge	✓	✓

Other components also generate logs in the vCloud environment that must be maintained by the provider, but direct tenant access is not required.

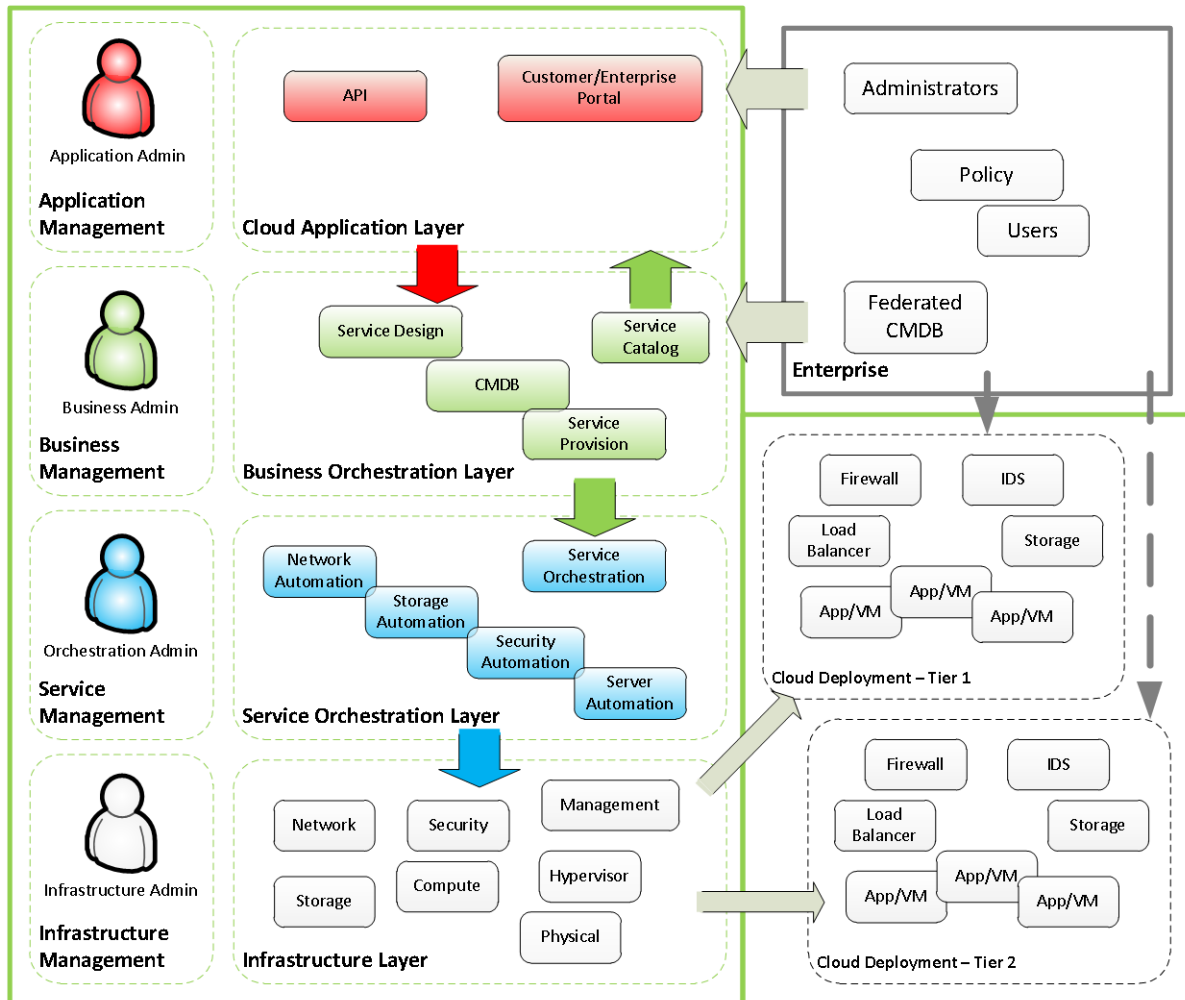
Table 7. Other Component Logs

Other Component	Provider Logs	Tenant Logs
vCloud Director DB (Oracle)	✓	
vCenter Database	✓	
vCenter Chargeback Database	✓	
Microsoft SQL Server	✓	
Linux (VCD)	✓	
Windows System Logs (CBM, vCO, vCenter Server)	✓	

Logs in the vCloud datacenter environment can further be categorized into the following logical business layers:

- vCloud Application – Represents the external interface with which the enterprise administrators of the vCloud interact. These administrators are authenticated and authorized at this layer, and have no (direct or indirect) access to the underlying infrastructure. They interact only with the Business Orchestration Layer.
- Business Orchestration – Represents both vCloud configuration entities and the governance policies that control the vCloud deployment:
 - Service catalog – Presents the different service levels available and their configuration elements.
 - Service design – Represents the service level and specific configuration elements along with any defined policies.
 - Configuration Management Database (CMDB) – Represents the system of record, which may be federated with an enterprise CMDB.
 - Service provision– Represents the final configuration specification.
- Service Orchestration – Represents the provisioning logic for the vCloud infrastructure. This layer consists of an orchestration director system, and automation elements for network, storage, security, and server/compute—vCenter Server, VMware vCloud Director (vCloud Director), vCenter Orchestrator.
- Infrastructure Layer – Represents the physical and virtual compute, network, storage, hypervisor, security, and management components – vSphere Server (ESXi), vShield Manager, and vShield Edge.

Figure 12. Infrastructure Layers



The abstraction of these four layers and their security controls helps illustrate audit and compliance requirements for proper authentication and segregation.

For example, vCloud provider administrator accounts should be maintained in a central repository integrated with two-factor authentication. Different tiers of vCloud deployments (provider virtual datacenters) would be made available to enterprise users.

vCloud Director Diagnostic and Audit Logs

VMware vCloud Director includes the following types of logs:

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

The VMware vCloud Director system audit log is maintained in the Oracle database and can be monitored through the Web UI. Each organization administrator and the system administrator have a view into the log scoped to their specific area of control. A more comprehensive view of the audit log (and long-term persistence) is achieved through the use of remote syslog (described below). Log management products are available from a variety of vendors and open source projects.

Audit events are not the only event types. Diagnostic logs contain information about system operation events and are stored as files in the local file system of each cell's operating system.

Diagnostic logs can be useful for problem resolution, but are not intended to preserve a trail of system interactions for audit. Each VMware vCloud Director cell creates several diagnostic log files, as described in the "Viewing the vCloud Director Logs" section of the *VMware vCloud Director's Administration Guide* for the latest version of VCD (http://www.vmware.com/support/pubs/vcd_pubs.html).

Audit logs record significant actions, including login and logout. A syslog server can be set up during installation as detailed in the *vCloud Director Installation and Configuration Guide* (http://www.vmware.com/support/pubs/vcd_pubs.html). Exporting the logs to a syslog server is required for compliance due to multiple reasons:

- Database logs are not retained after 90 days, but logs transmitted via syslog can be retained as long as desired.
- It allows audit logs from all cells to be viewed together in a central location at the same time.
- It protects the audit logs from loss on the local system due to failure, a lack of disk space, compromise, and so on.
- It supports forensics operations in the face of problems like those listed above.
- It is the method by which many log management and Security Information and Event Management (SIEM) systems will integrate with vCloud Director. This enables:
 - Correlation of events and activities across vCloud Director, vShield, vSphere, and even the physical hardware layers of the stack.
 - Integration of vCloud security operations with the rest of the vCloud provider's or enterprise's security operations, cutting across physical, virtual, and vCloud infrastructures.
- Logging to a remote system, instead of the system the cell is deployed on, provides data integrity by inhibiting tampering. Even if the cell is compromised it does not necessarily enable access to or alteration of the audit log.

Appendix C: Capacity Planning

Capacity forecasting provides an efficient way to acquire the appropriate amount of physical resources to support the increased demand for the vCloud. This allows for the growth of vCloud to be planned and included in the service providers' budgetary process, and reduces the likelihood of "panic buying," which generally increases costs dramatically and undermines standardization efforts. Capacity Planning also reduces the likelihood of last minute surprises, such as a lack of available space or power to support the new vCloud infrastructure components.

From a vCloud perspective, capacity management is simplified by the existence of the provider virtual datacenter and organization virtual datacenter constructs, but potentially more complicated by the addition of three models of consumption: Pay-As-You-Go, Allocation Pool (committed), and Reservation Pool (dedicated). Finally, all of these capacity management aspects, within a vCloud context, must address both the vCloud (service provider) administrator and the end-customer (organization) administrator perspectives.

Sizing for the workload resource group clusters can be difficult to predict because the provider is not in charge of what the consumer may run. The provider is also not aware of existing usage statistics for virtual machines that are run in the vCloud. The following information should assist in initial sizing of the vCloud environment and is based on information from the *Service Definition for a Private VMware vCloud*. This information is provided in the form of examples. VMware recommends that you engage with your local VMware representative for detailed sizing of your environment.

vCloud Administrator (Service Provider) Perspective

The primary capacity management concerns of the vCloud administrator are:

- Capacity management of provider virtual datacenters and the service offerings backed by each provider virtual datacenter.
- Network capacity management (network bandwidth capacity management is beyond the scope of this document).
- Capacity forecasting.
- Capacity monitoring and establishing triggers.

The VMware vCloud solution makes extensive use of reservations. As such, previous approaches to capacity management used in vSphere are not as applicable to a vCloud. For example, CPU and memory over-commitment cannot be applied as extensively as it was in a multitenant environment.

Unlike managing capacity for vSphere, in a vCloud, the virtual machine is no longer the basis for resource consumption from a service provider perspective. The organization virtual datacenter is the basis for resource consumption in a vCloud.

Capacity management is further impacted by the introduction of multiple consumption models in the vCloud model. Each model requires its own capacity management approach. As a result, this appendix provides guidance for capacity management from a service provider vCloud administrator perspective as it applies to each of the consumption models: Pay-As-You-Go, Allocation Pool, and Reservation Pool.

Regardless of the particular consumption model applied in a provider virtual datacenter, the common starting point of vCloud capacity management is to calculate the total amount of CPU and memory resources available for consumption. Because the underlying infrastructure provisioning unit of a provider virtual datacenter is a ESXi host, the first step is to determine the total CPU and memory at the vSphere host level. The following table shows the key vSphere host variables needed to calculate capacity, along with example values.

Table 8. vSphere Host Variables

Item	Variable	Value	Units
Processor Sockets	$N_{socket,1}$	2	integer
Processor Cores	$N_{cores,1}$	4	integer
Processor Speed	$S_{proc,1}$	2.4	GHz
Host Memory	$M_{host,1}$	64	GB

Calculating the total memory available is straightforward. It is the total amount of RAM for the vSphere host. Total CPU resources are calculated using the following formula:

$$P_{host} = N_{socket}N_{cores}S_{proc}$$

Using the example values from the table, the total CPU resource is equal to 19.2 GHz.

After the vSphere host capacity model has been defined, the next step is to determine the provider virtual datacenter (vSphere cluster) capacity. Determining the provider virtual datacenter capacity is critical as vCloud capacity management should be performed at the provider virtual datacenter level, not the vSphere host level.

When considering vCloud provider virtual datacenter capacity, an additional step is required to make sure that redundancy has been accounted for. The provider virtual datacenter cluster redundancy may vary depending upon service levels offered. For the following example, we assume N+2 cluster redundancy. This means that the provider virtual datacenter can absorb up to two vSphere host failures and continue to support all hosted virtual machines at the same level of performance. To accomplish this, there must be capacity available on the remaining vSphere hosts to take over all workloads.

Based on a requirement for provider virtual datacenter cluster redundancy, the overall number of memory and CPU consumption units for the provider virtual datacenter (cluster) must be reduced. To determine the redundancy overhead, the number of vSphere hosts in the cluster and the desired number of redundant vSphere hosts need to be considered. This is described in Table 9.

Table 9. Determining Redundancy Overhead

Redundancy Variables	Description
N_{nodes}	Represents the number of nodes in a cluster.
$N_{redundant}$	Represents the minimum number of redundant nodes.
$R_{redundancy, HA}$	<p>Represents a targeted ratio of redundancy as indicated by a real number greater than one. This ratio (such as 1.10) indicates that there is a ten percent overhead committed to availability. For example, a 10 node provider virtual datacenter with a 1.10 redundancy ratio would require 11 nodes to deliver the appropriate capacity. Note that this level of redundancy may vary depending on the class of service offering being delivered on that provider virtual datacenter.</p> <p>Redundancy variables can be determined with the equation below.</p>

Calculating Redundancy Ratio from Minimal Level of Redundancy

$$\left(\frac{N_{nodes} + N_{redundant}}{N_{nodes}} \right) = R_{redundancy}$$

For example, the level of redundancy is calculated below for a cluster size of ten nodes containing two redundant nodes.

$$\left(\frac{N_{nodes} + N_{redundant}}{N_{nodes}} \right) = \left(\frac{8 + 2}{8} \right) = 1.25 = R_{redundancy}$$

After the ratio of redundancy is calculated, the number of units of consumption per provider virtual datacenter can be determined using the following equation:

CPU resources per Cluster

$$N_{CPU, cluster} = \frac{N_{hosts, cluster} P_{CPU, host}}{R_{redundancy, HA}}$$

For our example where:

$$P_{CPU, host} = 19.2GHz$$

This results in:

$$N_{CPU, cluster} = \frac{8 \times 19.2}{1.25} = 122.88GHz$$

The number of memory units of consumption is calculated in the following equation.

For our example where:

$$N_{mem,host} = 64GB$$

This results in:

$$N_{mem,cluster} = \frac{N_{hosts,cluster} M_{mem,host}}{1.25} = \frac{8 \times 64}{1.25} = 409.6GB$$

We have now established that our example provider virtual datacenter has 122.88GHz of available CPU and 409.6GB of available memory, taking a vSphere cluster redundancy of N+2 into account. Next we look at some guidance for capacity management as it applies to each of the consumption models.

Pay-As-You-Go Model

When an organization virtual datacenter is created in the Pay-As-You-Go model, a resource pool is instantiated with expandable reservations. As such, the customer organization virtual datacenters contained on that provider virtual datacenter can grow to consume all of the available provider virtual datacenter resources. While this could be true in any vSphere environment, the added challenge in a vCloud is the use of reservations at the vApp level. When an organization virtual datacenter is created out of a provider virtual datacenter using the Pay-As-You-Go consumption model, a %guarantee is configured for CPU and memory. This is applied to each vApp or virtual machine within a vApp. For example, if the service provider configures the organization virtual datacenter with a 50% guarantee for CPU and 75% guarantee for memory, then the customer creates a virtual machine consuming 1 vCPU of 1GHz and 1GB of memory, a reservation for that virtual machine will be set at 50% of 1GHz, or 0.5 GHz and 75% of 1GB, or 0.75GB of memory.

Because there is no way of knowing how a customer will define their virtual machine templates in their private customer catalogs, coupled with the fact that organization virtual datacenters can expand on demand, VMware recommends the following:

- Calculate the total available CPU and memory resources (less an amount reserved for global catalog templates), adjusted by the cluster redundancy ratio, at the provider virtual datacenter level.
- Establish a CPU and Memory %RESERVED threshold at the provider virtual datacenter level.
- Establish the %RESERVED for the provider virtual datacenter at a number in the 60% range initially.
- As the total amount of reserved CPU or reserved memory approaches the %RESERVED threshold, do not deploy new organization virtual datacenters in that provider virtual datacenter without adding additional resources. If the corresponding vSphere cluster has reached its maximum point of expansion, a new provider virtual datacenter should be deployed and any new organization virtual datacenter s should be assigned to the new provider virtual datacenter. In this way there is 40% of expansion capacity for the existing organization virtual datacenters in the case where the provider virtual datacenter has reached its maximum point of expansion.
- CPU and memory over-commitment can be applied, and if so, the %RESERVED value should be set lower than if no over-commitment is applied due to the unpredictability of the virtual machine sizes being deployed (and hence reservations being established).
- Monitor the %RESERVED on a regular basis and adjust the value according to historical usage as well as project demand.

Allocation Pool Model

When an organization virtual datacenter is created in the Allocation Pool model, a non-expandable resource pool is instantiated with a %guaranteed value for CPU and memory that was specified. Using a %guaranteed value of 75%, this means if an organization virtual datacenter is created specifying 100GHz of CPU and 100GB of memory, a resource pool is created for that organization virtual datacenter with a reservation of 75GHz and limit of 100GHz for CPU and a reservation of 75GB with a limit of 100GB for memory. The additional 25%, in this example, is not guaranteed and can be accessed only if it's available across the provider virtual datacenter. In other words, the 25% can be over-committed by the provider at the provider virtual datacenter level and therefore may not be available depending on how *all* of the organization virtual datacenters in that provider virtual datacenter are using it.

At the virtual machine level, when a virtual machine is deployed, it is instantiated with no CPU reservation but with a memory reservation equal to the virtual machine's memory allocation multiplied by the %guaranteed. Despite the fact that no CPU reservation is set at the virtual machine level, the total amount of CPU allocated across all virtual machines in that organization virtual datacenter is still subject to the overall CPU reservation of the organization virtual datacenter established by the %guarantee value.

Based on this use of reservations in the Allocation Pool model, VMware recommends the following:

- Calculate the total available CPU and memory resources (less an amount reserved for global catalog templates), adjusted by the cluster redundancy ratio, at the provider virtual datacenter level.
- Determine how much resource, at the provider virtual datacenter level, you want to make available for expanding organization virtual datacenters that are deployed to that provider virtual datacenter.
- Establish a CPU and Memory %RESERVED (guaranteed, not allocated) threshold at the provider virtual datacenter level based on the %guaranteed less the amount reserved for growth. The remaining unreserved resources are available to all organization virtual datacenters for bursting.
- As the total amount of reserved CPU or reserved memory approaches the %RESERVED threshold, do not deploy new organization virtual datacenters in that provider virtual datacenter without adding additional resources. If the corresponding vSphere cluster has reached its maximum point of expansion, a new provider virtual datacenter should be deployed and any new organization virtual datacenters should be assigned to the new provider virtual datacenter. This gives some predetermined amount of capacity available for expanding the existing organization virtual datacenters in the case where the provider virtual datacenter has reached its maximum point of expansion.
- CPU and memory over-commitment can be applied, but it should be based only on the amount of unreserved resources at the provider virtual datacenter level, allowing for over-committing the resources available for organization virtual datacenter bursting.
- Monitor the %RESERVED on a regular basis and adjust the value according to historical usage as well as project demand.

Reservation Pool Model

When an organization virtual datacenter is created in the Reservation Pool model, a non-expandable resource pool is instantiated with the reservation and limit values equivalent to the amount of resources allocated. This means if an organization virtual datacenter is created allocating 100GHz of CPU and 100GB of memory, a reservation pool is created for that organization virtual datacenter with a reservation and limit of 100GHz for CPU and a reservation and limit of 100GB for memory.

At the virtual machine level, when a virtual machine is deployed, it is instantiated with no reservation or limit for either CPU or memory.

Based on this use of reservations in the Reservation Pool model, VMware recommends the following:

- Calculate the total available CPU and memory resources (less an amount reserved for global catalog templates), adjusted by the cluster redundancy ratio, at the provider virtual datacenter level.
- Determine how much resource, at the provider virtual datacenter level, you want to make available for expanding organization virtual datacenters that are deployed to that provider virtual datacenter.
- Establish a CPU and Memory %RESERVED threshold at the provider virtual datacenter level equivalent to the capacity of the underlying vSphere cluster, taking into account HA redundancy.
- As the total amount of reserved CPU or reserved memory approaches the %RESERVED threshold, do not deploy new organization virtual datacenters in that provider virtual datacenter without adding additional resources. If the corresponding vSphere cluster has reached its maximum point of expansion, a new provider virtual datacenter should be deployed and any new organization virtual datacenters should be assigned to the new provider virtual datacenter. In this way there is some predetermined amount of capacity available for expanding the existing organization virtual datacenters in the case where the provider virtual datacenter has reached its maximum point of expansion.
- No over-commitment can be applied to the provider virtual datacenter in the Reservation Pool model due to the reservation being at the resource pool level.
- Monitor the %RESERVED on a regular basis and adjust the value according to historical usage as well as project demand.

Storage

VMware vCloud Director uses a largest available capacity algorithm for deploying virtual machines to datastores. Storage capacity must be managed on both an individual datastore basis as well as in the aggregate for a provider virtual datacenter.

In addition to considering VMware storage allocation best practices, manage capacity at the datastore level using the largest virtual machine storage configuration, in terms of units of consumption, offered in the service catalog when determining the amount of spare capacity to reserve. For example, if using 1TB datastores (100 storage units of consumption based on a 10GB unit of consumption) and the largest virtual machine storage configuration is 6 storage units of consumption (60GB), then applying the VMware best practice of approximately 80% datastore utilization would imply managing to 82 storage units of consumption. This would result in 82% datastore utilization and reserve capacity equivalent to three of the largest virtual machines offered in the service catalog in terms of storage.

Network Capacity Planning

A vCloud also brings network capacity planning to the forefront. Providers must consider IP address, VLAN, and ephemeral port capacity. The following table describes what must be managed from a capacity perspective and its impact.

Table 10. Network Capacity Planning Items

Item to Manage	Impact
IP addresses	<ul style="list-style-type: none">• Available IP addresses to be assigned in support of a dedicated external network for an organization, such as for Internet access or hardware-based firewall rules.• Need to track IP addresses assigned to specific organizations to determine what is available for a shared external organization network.
VLANs	<ul style="list-style-type: none">• VLANs available for VLAN-backed pool assignment, if required.• VLANs available for vCloud Director Network Isolation transport networks, one per vCloud Director Network Isolation pool.
Expandable static port bindings	<ul style="list-style-type: none">• Default vCloud Director network pool type.• Overall number of static ports expands in increments of ten as needed. Note that unused but allocated static port bindings do not increase the total number of static port bindings available.

Appendix D: Capacity Management

vCloud-Specific Capacity Forecasting (Demand Management)

Capacity forecasting consists of determining how many organization virtual datacenters are expected to be provisioned during a specific time period. Capacity provisioning is concerned with determining when vCloud infrastructure components must be purchased in order to maintain capacity. From a financial budget perspective, the procurement of the vCloud infrastructure requires more planning and understanding of customer future requirements.

VMware recommends performing two forecasting functions over time.

- **Capacity Trending** – Using historical organization virtual datacenter capacity and utilization data, it is possible to predict future capacity requirements.
- **Demand Pipeline** – Understanding future customer requirements via the sales pipeline provides the necessary information to understand future capacity requirements, as well as knowledge of marketing/business development functions bringing new service offerings to market.

Initially, no historical utilization metrics are available, and thus it is not possible to perform capacity trending for some period of time. During this initial period, a good understanding of the customer demand pipeline needs to be established. Over time, this pipeline can be combined with trending analysis to more accurately predict capacity requirements.

The customer demand pipeline must be established in conjunction with the service provider's sales teams, or lines of business (LOB) if a private cloud, so future vCloud capacity requirements can be determined. This demand pipeline must contain information of all known new customers, expansion of existing customer organization virtual datacenters, projected sizing metrics, plus any new service offerings that are in development. The forecasting plan must fit both the budgetary cycle and the procurement and provisioning timeframes. For example, if a quarterly budgetary cycle exists, and the procurement and provisioning timeframe is one month, it is necessary to have a pipeline of *at least four months* to make sure all requests in the pipeline can be fulfilled.

Over time, capacity trending can be used to assist with the forecasting of organization virtual datacenter provisioning needs. It uses historical information to determine trends and validates the organization virtual datacenter forecast based on demand pipeline data.

Capacity Monitoring and Establishing Triggers

The metrics listed in Table 11 should be carefully monitored to warn of approaching or exceeding consumption thresholds. These metrics should be measured against each vCloud provider virtual datacenter and for each organization virtual datacenter within each provider virtual datacenter. To monitor for threshold breaches, and possible subsequent violation of service level commitments to the vCloud consumer, the appropriate tools and triggers are needed for proper notification.

Table 11. Capacity Monitoring Metrics

Attribute	Monitored per
%RESERVED CPU	Provider virtual datacenter, organization virtual datacenter Note For the Pay-As-You-Go allocation model this is the aggregation of reservations values for the contained virtual machines.
%RESERVED Memory	Provider virtual datacenter, organization virtual datacenter
CPU utilization	Provider virtual datacenter, organization virtual datacenter
Memory utilization	Provider virtual datacenter, organization virtual datacenter
Datastore utilization	Provider virtual datacenter
Transfer store utilization	vCloud
Network IP addresses available	vCloud
Network IP addresses consumed	Organization
Network VLANs available	vCloud
Network ephemeral ports consumed	vNetwork Distributed Switch

If thresholds are exceeded, the group responsible for capacity management of the vCloud should be notified to add additional capacity. Take into account the time required to add the physical components necessary to increase the capacity of a provider virtual datacenter. A vCloud-aware capacity management tool should be deployed. Whichever tool is chosen, the capacity model can be used to forecast new provider virtual datacenter capacity utilization as well as ongoing capacity management of existing provider virtual datacenters. It should also account for expansion triggers based on provisioning timeframes.

After the total amount of available resources has been calculated for a provider virtual datacenter, no adjustments to that provider virtual datacenter (such as adding or removing hosts) should be made without updating the calculated value. This model may be altered if long-term CPU and memory reservations are not at the levels for which they were designed. An increase in the resources allocated to an organization virtual datacenter can affect the remaining capacity of a “full” provider virtual datacenter. Full provider virtual datacenters should be monitored on a weekly basis. The resource consumption of virtual machines within an organization virtual datacenter should be reviewed for trends that indicate the resources purchased for that organization virtual datacenter are insufficient.

vCenter CapacityIQ, though not vCloud Director aware, can be used to provide insight into provider virtual datacenter utilization and trends.

Capacity Management Manual Processes – Provider Virtual Datacenter

The following vCloud administrator capacity management activities include periodic planning activities supported by day-to-day operational activities. Periodic continuous improvement activities are critical to extracting the most value from your vCloud infrastructure.

Planning activities (initially monthly, then quarterly):

- Determining usable capacity by provider virtual datacenter and organization virtual datacenter (taking into account vSphere overhead).
- Reviewing current utilization.
- Reviewing provisioning timeframes for new provider virtual datacenter components (hosts, network, storage).
- Forecasting utilization growth over the coming period (preferably based on the actual pipeline, validated with historical trending).
- Planning for procurement and implementation of additional capacity over the coming period, including bills of materials and budgets.
- Reviewing capacity alert threshold levels and setting alerts for capacity warnings.

Operational activities (daily):

- Monitoring for alerts.
- Investigating performance issues to determine whether capacity is the root cause.
- Initiating and managing the procurement and provisioning of additional provider virtual datacenter capacity.

Continuous improvement activities (quarterly/yearly):

- Comparing capacity model utilization levels to observed levels and tuning model to drive greater utilization without sacrificing reliability.
- Optimizing provisioning timeframes (shortening them and making them more predictable).

End-Customer (Organization) Administrator Perspective

The primary capacity management concern of the organization administrator is capacity management of the organization's organization virtual datacenters.

VMware recommends that all organizations establish a capacity management process based on a standard unit of consumption. The recommended base Unit of Consumption for each resource important to capacity management from an organization administrator perspective is shown in Table 12.

Table 12. Organization Virtual Datacenter Units of Consumption

Attribute	Variable	Value
vCPU	P_{vc}	1 GHz
Memory	M_{vc}	1GB
Storage	D_{vc}	10GB

Taking this approach enables more efficient capacity management because the vApp component virtual machine resource allocations are predefined in the service catalog, resulting in vCloud infrastructure resource consumption being more accurately predicted.

Each organization will be provided with a finite quantity of resources (in the cases of the Allocation Pool and Reservation Pool consumption models) from one or more provider virtual datacenters in the form of organization virtual datacenters. This means that as the organization consumes the organization virtual datacenter resources, a trigger point needs to be defined to prompt actions to be taken to expand the organization virtual datacenter.

First, the resource consumption limits for an organization's organization virtual datacenters need to be defined, with these limits defining when action needs to be taken to remove the potential capacity issue.

Table 13. Recommended Organization Virtual Datacenter Capacity Thresholds

Attribute	Variable	Limit	Description
organization virtual datacenter CPU Peak Utilization	$C_{CPU\text{Limit}}$	80%	The limit for allocating CPU resources within the organization virtual datacenter before expansion is required. This value varies depending on the consumption model used. From an organization virtual datacenter perspective, reservation values should be considered equal to the amount of CPU allocated as reservation values are not available to the organization administrator.
organization virtual datacenter Memory Allocation Limit	$C_{mem\text{Limit}}$	80%	The limit for allocating memory resources within the organization virtual datacenter before expansion is required. This value varies depending on the consumption model used. From an organization virtual datacenter perspective, reservation values should be considered equal to the amount of memory allocated as reservation values are not available to the organization administrator.

The CPU and memory resources vary depending on the size of the contracted organization virtual datacenter. Table 14 provides an example of the resources needed to calculate the organization virtual datacenter's capacity.

Table 14. Sample Organization Virtual Datacenter Resource Allocation

Item	Variable	Value	Units
Total organization virtual datacenter vCPU Units of Consumption	$S_{org\text{virtual datacenter}}$	50	GHz
organization virtual datacenter Memory Allocation in Units of Consumption	$M_{org\text{virtual datacenter}}$	64	GB

The number of capacity units available within this organization virtual datacenter is found using the following equations.

Determining organization virtual datacenter memory units of consumption

$$M_{UC,orgVDC} = \left(\frac{C_{memLimit} M_{orgVDC}}{M_{UC}} \right)$$

Based on the information from the above tables, the total memory unit of consumption for the organization virtual datacenter is calculated as shown below.

$$M_{UC,orgVDC} = \left(\frac{C_{memLimit} M_{orgVDC}}{M_{UC}} \right) = \left(\frac{0.8 \times 64}{1} \right) = 51.2GB$$

This results in 51.2 memory units of consumption for the sample organization virtual datacenter.

Determining organization virtual datacenter CPU units of consumption

$$P_{UC,orgVDC} = \left(\frac{S_{orgVDC} C_{CPULimit}}{P_{UC}} \right)$$

Based on the information from the above tables, the CPU units of consumption per organization virtual datacenter are calculated as shown below.

$$P_{UC,orgVDC} = \left(\frac{S_{orgVDC} S_{CPULimit}}{P_{UC}} \right) = \left(\frac{50 \times 0.8}{1} \right) = 40GHz$$

This results in 40 CPU units of consumption for this sample organization virtual datacenter.

Organization Virtual Datacenter-Specific Capacity Forecasting

Capacity forecasting consists of determining how many virtual machines are expected to be deployed during a specific time period of the organization's choosing. The time period used for the virtual machine forecast should correspond to the budgetary process. Capacity provisioning is concerned with determining when an organization virtual datacenter must be expanded in order to maintain capacity.

VMware recommends that organizations perform two forecasting functions over time.

- Capacity Trending – Using historical virtual machine capacity and utilization data, it is possible to predict future capacity requirements.
- Capacity Pipeline – Understanding future end-user virtual machine resource requirements, via IT and LOB projects, provides the necessary information to understand future capacity requirements.

Over time, capacity trending can be used to assist with the forecasting of virtual machine provisioning needs. It uses historical information to determine trends and validates the virtual machine forecast based on pipeline data.

Capacity provisioning depends on determining the point of expansion for the organization virtual datacenter. This is based on determining a point of resource consumption at which the process of procuring and expanding the organization virtual datacenter must begin so that reserve capacity is not exhausted before the additional capacity is available. In the vCloud context, this can be considered to be dependent upon the time it takes to process the purchase request for additional organization virtual datacenter resources. Provisioning time can be assumed to be zero but depends upon specific contractual agreements with the service provider.

The following are recommended steps to perform capacity trending and to determine a point of organization virtual datacenter expansion.

Regularly Collect Organization Virtual Datacenter Consumption Information

The primary issue with the trending of organization virtual datacenter consumption is identifying the point of record for all new virtual machines. This can then be used to determine the capacity trends and therefore determine the overall need for purchasing additional organization virtual datacenter capacity. To establish the point of record for new virtual machines, the items listed in Table 15 should be tracked, ideally in a Configuration Management or Capacity Planning Database as virtual machine attributes.

Table 15. Organization Virtual Datacenter Trending Information

Variable	Name	Description	Units
$orgvirtual\ datacenter$	Organization virtual datacenter	This is the organization virtual datacenter in which the virtual machine resides.	Identifier
D_{build}	Build Date	This is the date the virtual machine is built.	Date
$N_{UC,cpu}$	CPU Units of Consumption	This is the number of CPU units of consumption allocated to the virtual machine.	CPU Units of Consumption
$N_{UC,mem}$	Memory Units of Consumption	This is the number of memory units of consumption allocated to the virtual machine.	Memory Units of Consumption
N_{VGB}	Storage	This is the amount of storage (GB) allocated to the virtual machine.	GB

Determine Trending Variables

With the information recorded as described in Table 15 it is possible to determine the rate of organization virtual datacenter consumption.

Table 16. Organization Virtual Datacenter Capacity Trending Variables

Variable	Name	Description	Units
T	Time	This is the time between points of observation.	Weeks
N_{cpuUC}	New CPU Units	This is the total number of CPU units of consumption required for the forecasted virtual machines.	CPU Units of Consumption
N_{memUC}	New Memory Units	This is the total number of memory units of consumption required for the forecasted virtual machines.	Memory Units of Consumption
N_{VGB}	New Storage (GB)	This is the total amount of storage required for the forecasted virtual machines.	GB
$T_{purchase}$	Organization Virtual Datacenter Expansion Purchase Time	The amount of time to procure additional organization virtual datacenter resources.	Weeks

Determining the Trended Growth Rate

$$\Delta N_{cpuUC} = \frac{N_{cpuUC}}{\Delta T}$$

$$\Delta N_{memUC} = \frac{N_{memUC}}{\Delta T}$$

$$\Delta N_{VGB} = \frac{N_{VGB}}{\Delta T}$$

Determining the Trend

It is important to understand that the rate of increase dictates how far in advance additional organization virtual datacenter resources need to be purchased. The following table presents a sample virtual machine forecast for a quarter along with sample “time to purchase” value.

Table 17. Sample Organization Virtual Datacenter Trending Information

Attribute	Value
ΔN_{cpuUC}	12
ΔN_{memUC}	12
ΔN_{VGB}	360GB
$T_{purchase}$	2 weeks
$N_{cpuUC, cluster}$	320
$N_{memUC, cluster}$	717

In this example, $N_{cpuUC, free}$ and $N_{memUC, free}$ represents the number of free resources within an organization virtual datacenter at which point additional organization virtual datacenter resources should be ordered. To determine the trigger point for ordering use the following equation if no pipeline data exists.

Determining Trigger Point for Ordering Capacity Using Trends

$$N_{UC, free} = \Delta N_{CU} \times T_{purchase}$$

For example, from the data provided below, one would calculate the needed free consumption units as listed in the following equation, or 24 units.

$$N_{cpuUC, free} = \Delta N_{cpuUC} \times T_{purchase} = 12 \times 2 = 24GHz$$

$$N_{memUC, free} = \Delta N_{memUC} \times T_{purchase} = 12 \times 2 = 24GB$$

For storage, in this example, the trigger point is calculated at 720GB:

$$N_{VGB, free} = \Delta N_{VGB} \times (T_{purchase}) = 360 \times 2 = 720GB$$

Determine the Automatic Point of Expansion

Based on the example above, additional organization virtual datacenter resources would need to be ordered when the available units of CPU or memory fall to 24GHz or 24GB respectively, or when storage capacity falls to 720GB. The additional capacity needs to be on order when described or the capacity will not be available in time to meet demand.

Currently there are no tools available to assist in organization virtual datacenter capacity management. However, it is possible to develop scripts to gather pertinent information using languages such as PowerCLI.

Capacity Management Manual Processes – Organization Virtual Datacenter

The following organization administrator capacity management activities include periodic planning activities supported by day-to-day operational activities. Periodic continuous improvement activities are critical to extracting the most value from your vCloud.

- Planning activities (initially monthly, then quarterly):
 - Determining usable capacity by organization virtual datacenter.
 - Reviewing current utilization (and performance, where possible).
 - Reviewing purchasing timeframes for expanding an organization virtual datacenter.
 - Forecasting utilization growth over the coming period (preferably based on actual pipeline validated by historical trending).
 - Reviewing capacity alert threshold levels and setting alerts for capacity warnings.
- Operational activities (daily):
 - Monitoring for alerts.
 - Investigating performance issues to determine whether capacity is the root cause.
 - Initiating and managing the procurement and provisioning of additional capacity.
- Continuous improvement activities (quarterly/yearly): Comparing capacity model utilization levels to observed levels and tuning model to drive greater utilization without sacrificing reliability.

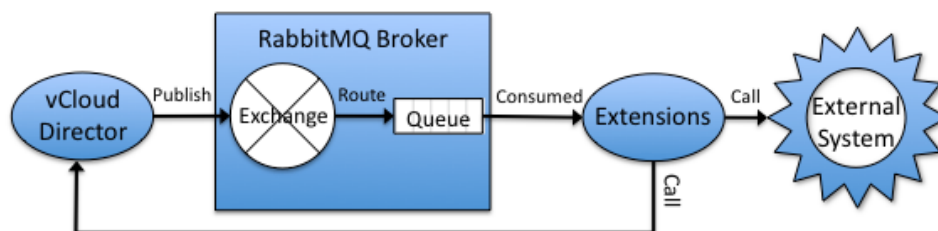
Appendix E: Integrating with Existing Enterprise System Management

There are several mechanisms available for integrating vCloud with existing enterprise system management tools. These range from the VCD notification capabilities introduced in VCD 1.5 to vCenter Orchestrator, the vCloud API, and, for providers, the VIX API. This appendix addresses the VCD notification capability, using vCenter Orchestrator, and the VIX API. For more information about the vCloud API, see the *vCloud API Specification* (https://www.vmware.com/pdf/vcd_10_api_spec.pdf) and the *vCloud API Programming Guide* (https://www.vmware.com/support/pubs/vcd_pubs.html).

vCloud Director Notifications and Blocking Tasks Messages

vCloud Director 1.5 supports notifications and blocking tasks features that allow it to interoperate with applications, extending its capabilities.

Figure 13. vCloud Director Extension Overview



Message Publication

The system administrator can configure vCloud Director to enable the publication of messages for all event notifications and/or for specific blocking tasks:

- The notifications are published upon user-initiated events (for example, creation, deployment and deletion of a vApp) as well as system-initiated events (for example, vApp lease expiration) containing the new state of the corresponding vCloud Director entity.
- The blocking tasks suspend long running operations started as a task before publishing messages and wait until a system administrator takes action.

The message publication is enabled both for operations started in the vCloud Director GUI or vCloud API and can be acted upon by using either interface.

The notification messages are published to an Advanced Message Queuing Protocol (AMQP) exchange (AMQP version 0.9.1 supported by RabbitMQ version 2.0 and above).

Routing

The AMQP broker uses routing as an effective way to filter vCloud director notification messages and dispatch them to different queues for one or multiple extensions.

The exchange routes notifications to its bound queues according to their queue routing key and exchange type. The vCloud notification messages routing key has the following syntax format:

```
<operationSuccess>.<entityUUID>.<orgUUID>.<userUUID>.<subType1>.<subType2>...
<subTypeN>.[taskName]
```

Extension

An extension is a script or an application with the following capabilities:

- Subscribe to an AMQP queue for receiving new messages.
- Triage the received messages.
- Process messages into operations (internal or external calls).
- Call vCloud Director API back for getting more information on the objects involved in an operation and taking action on blocked task.

Subscribe to an AMQP queue

Subscribing to queues involves declaring a queue, binding with a routing key, and then subscribing to the declared queue.

The queue routing key supports the “*” and “#” wildcard characters to match a single segment and zero or more segments. For example `true.*.*.*.com.vmware.vcloud.event.vm.create` or `true.#.com.vmware.vcloud.event.vm.create` will route a notification to the queue with this binding key every time any user from any organization successfully creates a virtual machine).

Declaring asserts the existence of the object. If it does not exist, it will create it.

Triage the consumed messages

When a message is consumed, the extension can use the message header that contains all the routing components to further filter and act upon. For example, some notifications may be ignored.

Separate the notifications messages from the blocking tasks because the blocking tasks must be handled differently.

Handling the notification messages

The notification messages contain the operation triggering the event; the object type, and identifiers and names for organization, user, and object.

These can be used as markers for applications such as audit logging, Change Management, and Incident Management. If the application cannot correlate the IDs to present the objects properties in an end-user consumable form, the extension application has to call back the vCloud API to extract these.

Use notification messages to start an operation that must follow another one. For example, enabling the public IPs of a vApp in a load balancer.

Handling blocking tasks messages

Blocking tasks messages have similar identifier with the object being the blocking task. The blocking task references:

- Its parent task – The suspended task referencing the object and the task parameters attributes it was set with in the original request.
- TaskOwner – The object on which the task operates.
- The actions that can be taken on this blocking task (resume, abort fail, updateProgress).

Receiving and acting upon on the blockings task is accomplished with the vCloud director API callbacks. System admin privileges are required to perform these operations.

Aborting a task returns a success status. It should be done only:

- If the requested vApp went through automatic approval logic and was disapproved.
- To replace an operation to be carried out by another one. For example, start a pre-provisioned vApp instead of provisioning a vApp.
- When it is required that parameters for a requested task be replaced. For example, when determining a specific virtual datacenter for a vApp based on placement logic

Note When calling the same operation as the one that triggered the notification routing and filtering must be properly configured to avoid creating a loop.

A task should be failed when the operation occurring before the task is determined to fail. An example is an operation required before running the task failed. For example, CMDB was not reachable.

The task must be resumed for operations that must complete before the next task starts. Examples include:

- OVF user information must be added to a vApp before adding a vApp to catalog.
- Requested vApp goes through automatic approval logic and was approved before being added to vCloud.
- Change request must record the object state in CMDB system before making change.

Task progress should be updated to avoid having the task time out, or to log a status message to the end user.

Blocking Tasks and Notifications Use Case

This section covers the messages published during the use case: App Author adds a vApp from catalog. Notifications and blocking task for “Instantiate vApp from vApp Template” are enabled.

- Notification message: vApp creation requested.
(true.#.com.vmware.vcloud.event.vapp.create_request - # is used as a placeholder)
- Notification message: VM creation requested – a scaffold object is created and resources are locked
(true.#.com.vmware.vcloud.event.vm.create_request)
- Notification message: A task to instantiate a vApp is created.
(true.#.com.vmware.vcloud.event.task.create.vdcInstantiateVapp)
- Blocking tasks message: vApp instantiation has been blocked.
(true.#.com.vmware.vcloud.event.blockingtask.create.vdcInstantiateVapp)
- vCloud Director User Interface shows the task as “Pending processing ...”

Case 1: System admin calls abort on the blocked task.

- Blocking tasks message: The blocking task has been aborted.
(true.#.com.vmware.vcloud.event.blockingtask.abort.vdcInstantiateVapp)
- Notification message: The vApp is modified as per the next operation.
(true.#.com.vmware.vcloud.event.vapp.modify)
- Notification message: The scaffold object is deleted. Resources are unlocked.
(true.#.com.vmware.vcloud.event.vm.delete)
- Notification message: The vApp instantiation is aborted.
(true.#.com.vmware.vcloud.event.task.abort.vdcInstantiateVapp)

- The newly created object is no longer displayed from vCloud Director user interface. The task can be seen in Logs/Tasks.

Case 2: System admin fails the blocked task.

- Blocking tasks message: The blocking task has been failed.
(true.#.com.vmware.vcloud.event.blockingtask.fail.vdcInstantiateVapp)
- Notification message: The VM is not created (false.#.com.vmware.vcloud.event.vm.create)
- Notification message: The vApp instantiation task has been failed
(true.#.com.vmware.vcloud.event.task.fail.vdcInstantiateVapp)
- vCloud Director User Interface shows the task is having an error on object grid and in Logs/Tasks.

Case 3: System admin resumes the task.

- Blocking tasks message: The blocking task has been resumed.
(true.#.com.vmware.vcloud.event.blockingtask.resume.vdcInstantiateVapp)
- Notification message: The vApp is instantiated.
(true.#.com.vmware.vcloud.event.task.start.vdcInstantiateVapp)
- Notification message: The vApp is created. (true.#.com.vmware.vcloud.event.vapp.create)
- Notification message: The VM is created. (true.#.com.vmware.vcloud.event.vm.create)

Case 3a: The vApp instantiation is successful or aborted.

(true.#.com.vmware.vcloud.event.task.complete.vdcInstantiateVapp)

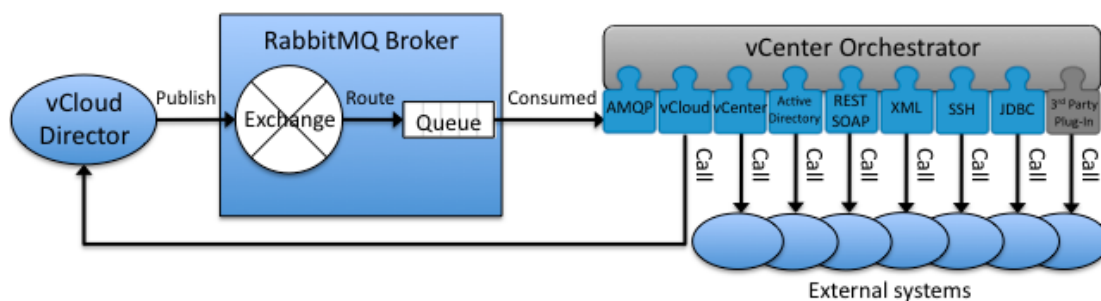
Case 3b: The vApp instantiation fails.

(false.#.com.vmware.vcloud.event.task.complete.vdcInstantiateVapp)

Using vCenter Orchestrator as a vCloud Director Extension

VMware vCenter Orchestrator fully supports consumption of blocked tasks and notifications messages, callbacks, and calls to external systems via the vCloud Director, AMQP, and other product plug-ins.

Figure 14. vCenter Orchestrator as a vCloud Director Extension



The AMQP plug-ins comes with workflows, and requires a one time setup.

1. Add a broker – Add an AMQP broker with providing hostname and credentials.
2. Declare an exchange – Declare an exchange for the configured broker.
3. Declare a queue – Declare a queue.
4. Bind – Bind a queue to an exchange by providing a routing key.
5. Subscribe to queues – Enables message updates on new messages.

This configuration is saved and reloaded automatically when the vCenter Orchestrator server is restarted.

The plug-in supports adding a policy element of type subscription having an onMessage trigger event. A policy can be set up to start a workflow that processes new messages.

Workflows are provided to triage and process the message to output vCloud Director objects. These can provide all of the information necessary for audit purposes and for designing custom logic before calling external systems. External systems are called using specific vCenter Orchestrator plug-in adapters such as vCloud Director, vCenter, Update Manager, Active Directory or generic plug-ins adapters such as REST, SOAP, XML, SSH, and JDBC. Blocked tasks objects can then be aborted, resumed, or failed by calling vCloud Director Workflows.

vCenter Orchestrator as an Extension Example

This section shows a simple example leveraging the blocked tasks as a trigger mechanism for starting extension workflows using different vCenter Orchestrator plug-ins.

As a prerequisite, a subscription to an AMQP queue, bound to the exchange used by vCloud director, was created using the workflows listed in the previous section. As part of this, the routing key is set to filter on vApp creation (#.blockingtask.create.vdcInstantiateVapp).

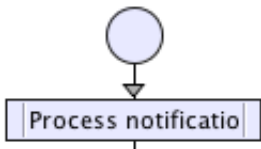
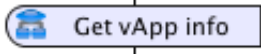
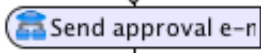
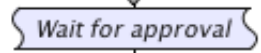
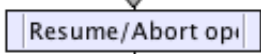
Next an “Approve new vApp” policy is created to listen on new messages. It is set to start the “Approve a vApp” workflow.

Figure 15. vCenter Orchestrator AMQP Subscription Policy

Local Parameter	Source parameter	Type
subscription	self [in-parameter]	AMQP:Subscription

The “Approve a vApp” workflow is designed as shown in Table 18.

Table 18. Approve a vApp workflow

Workflow	Description	Plug-in in use
	Important information is extracted from the subscription message such as the name of the vApp requester and the scaffold object of the vApp being requested.	AMQP
	The detailed properties of the requested vApp are gathered.	vCloud Director
	The vApp requester’s manager name and email is found in Active Directory, an email is sent to approve the vApp. It contains all the details gathered before.	Active Directory and Mail
	The workflow is stopped until the approver follows the link in his email, authenticates using his Active Directory credential, and approves or rejects the vApp.	
	Depending on if the vApp was approved or not, the aborted task is resumed or aborted. An email message is sent to the requester.	vCloud Director and Mail

VIX API

The VMware® VIX API enables automation of virtual machine operations, and libraries are available for C, Perl, and COM. Programs or scripts making use of the VIX API are referred to as VIX clients. Common use cases for VIX API virtual machine operations include:

- Performing power operations (start, stop, suspend, resume) on a virtual machine.
- Performing VMware Tools installation (some manual intervention may be required).
- Resetting passwords.
- Killing system processes.
- Cleaning up temporary log files.
- Installing/configuring software inside the guest operating system.
- Copying files to or from the guest operating system.

If performing operations that can affect the file system or execute programs within a guest operating system, the VIX client must authenticate with the guest operating system. The VIX client provides a username and password that can be authenticated as a valid user account by the guest operating system.

VIX clients may run programs or scripts within a guest operating system. This capability can be used to install software, run maintenance tasks, and trigger actions based on complex event processing. When installing software using the VIX API, having the ability to install the software in an unattended and/or scripted fashion will simplify the process.

Because VIX API virtual machine operations use VMware Tools as the communication path to the guest operating system, an available network connection is *not* required. This allows VIX clients to run programs or scripts and perform other configuration tasks before a network connection is made available by the guest operating system.

Private vCloud and managed services providers often require agent-based software to be installed and configured in the guest operating system of virtual machines. Public vCloud providers having additional value-add capabilities may also require agent-based software and/or the ability to perform customization of virtual machine guest operating system configuration elements.

Agent-based software examples include:

- Backup/restore.
- Performance monitoring.
- Virus scanners.

Customization of software within a virtual machine may be possible through scripts or programs executed within the guest operating system. A VIX client can execute these scripts or programs using command line arguments to pass values to the script or program. As an example, consider the public vCloud provider that:

- Provides NAS storage as a value-add service.
- Has a portal that allows configuration and provisioning of the storage for consumption by client virtual machines running in the vCloud.
- Automatically configures the guest operating system to mount the storage and makes the mount consistent across reboots.

In large environments where complex events are occurring in systems linked by infrastructure services or application components, it may be necessary to have a centralized workflow system that can trigger tasks within virtual machine guest operating systems. vCenter Orchestrator has a VIX plugin that extends the workflow capabilities in vCenter Orchestrator all the way to the guest operating system within a virtual machine.

Appendix F: Business Continuity

Backup and restore of the entire vCloud infrastructure involves the coordination of numerous components. Consider what is necessary to recover from a service disruption. What components are most critical and complex to restore? What types of failures would be the most catastrophic? The biggest threat to data loss is not hardware failure, but people accidentally deleting or incorrectly configuring their vApps.

vApp Backup/Restore

Currently, most backup products lack integration with vCloud Director. Without visibility into the vApp metadata stored in the vCloud Director database, recovery involves manual steps to restore data and re-establish configuration attributes. Some of the configuration attributes include the owner, network, and organization, and can be manually configured or re-assigned through the vCloud API.

The following sections walk through how a vCloud backup product would backup and restore a vApp in the vCloud environment.

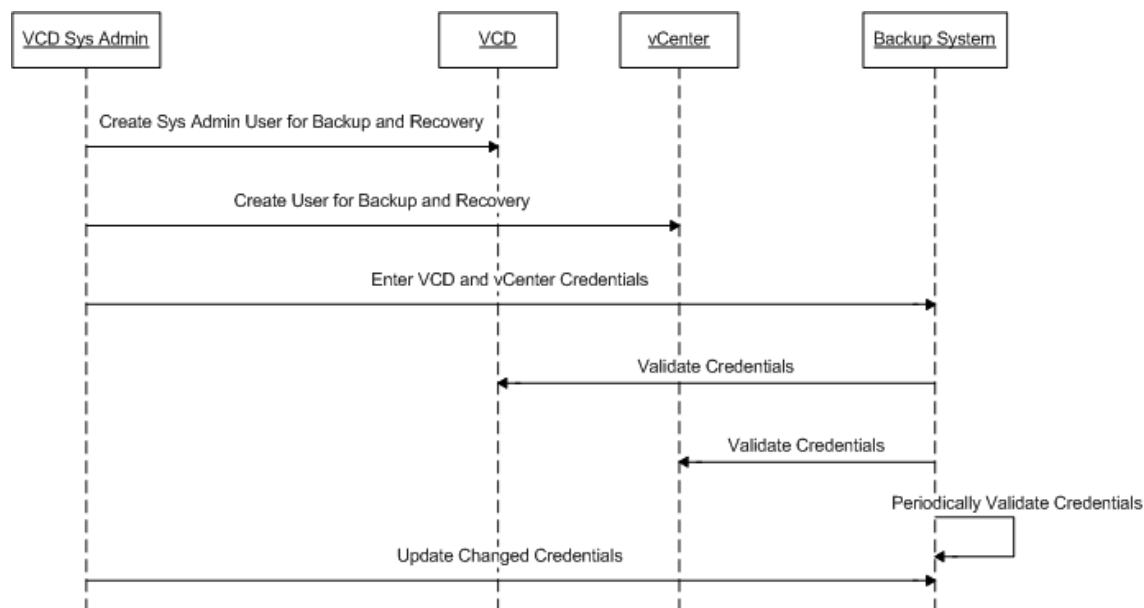
Use the following high-level procedure to back up and restore a vCloud vApp:

1. Manage credentials.
2. Protect vApps and create backup jobs.
3. Execute backup job.
4. Recover vApp to new or overwrite existing.

Manage Credentials

Without credentials, no systems are accessible. Because vCloud Director and vSphere components have separate sets of credentials, the backup product either requests the user to enter both sets of credentials at runtime or harvests the credentials for later use.

Figure 16. Credential Management Workflow



Protect vApps and Create Backup Jobs

With valid credentials, the backup product can connect to vCloud Director and vSphere components, extract the data hierarchy, and list the UUIDs of the vApps available for backup. Use the vCloud API or the vCloud Director Web console to perform this task. Then, find the location of the virtual machines to backup.

If using REST code, the logic is to:

1. Start at the top level of the inventory by getting a list of the vCenter Servers that are attached to vCloud Director and all of the organizations.
2. Build a map of the vCenter Servers keyed on their ID for easy lookup later.
3. Browse to the appropriate level. When browsing to an organization virtual datacenter, all the vApps in that organization virtual datacenter are visible, as well as all the datastores accessible to the organization virtual datacenter (through the parent provider virtual datacenter). When browsing to a vApp, all virtual machines in that vApp are visible.
4. The data captured by the end of the process should be:
 - Organization.
 - Organization virtual datacenters.
 - Datastores.
 - vApp network configuration (vApp networks, organization networks, and NAT, firewall, and DHCP settings).
 - Virtual machines belonging to that vApp. For each virtual machine, retrieve the same virtual machine properties needed to perform vSphere backups (managed object reference, network, description, others).

Execute Backup Jobs

After locating the virtual machines to back up, the backup product can execute the backup job using the appropriate information. The APIs used are the vCloud API, vSphere API, and the VMware Disk Development Kit (VDDK). VDDK is a subset of the VMware APIs for Data Protection (VADP).

Most customers are no longer using agent-based backups, opting for more efficient and tightly integrated products that leverage VADP. Agent-based backups can be used in a vCloud environment to overcome some of the challenges posed by vApp networks.

Recovery

Prior to recovery, place the vApp in maintenance mode to prohibit end users from performing operations that change the state of the vApp. After recovering the vApp, make the vApp available by exiting maintenance mode.

To restore vApps to a previous state, shut down the vApp and use the backup product to overwrite existing virtual disk files in the vApp.

Recovery of a deleted vApp requires re-importing virtual machines into vCloud Director as follows:

1. Import the first virtual machine into a new vApp, thereby creating the vApp.
2. Import the rest of the virtual machines belonging to the vApp.
3. Configure each virtual machine with the appropriate properties (organization virtual datacenter, the newly restored name, vApp network, and so on.)
4. After all virtual machines have been imported, validate that the correct properties are in place (network connections, ownership).

Infrastructure Backup/Restore

Synchronize the backup of all vCloud infrastructure components. There are multiple ways of achieving this by using snapshots, VADP, or other backup tools. Quiesce all databases at the same time before taking snapshots or creating backups. A database out of sync can cause a recovery nightmare.

Table 19 covers the recommended protection policies.

Table 19. Recommended Protection Policies

	Type	Description	Data Protection Policy
vCloud Director installation files	Infrastructure	Static information consists of product binaries for each cell.	<ul style="list-style-type: none"> • VM snapshot • Frequency – Once
vCloud Director log files	Infrastructure	Dynamic information generated by each cell. Located in <code>\$VCLLOUD_HOME/logs</code> . Multicell installations would use a syslog server to centralize log files.	<ul style="list-style-type: none"> • File level backup • Frequency – periodic
vCloud Director configuration file	Infrastructure	Dynamic information for each cell. File is <code>\$VCLLOUD_HOME/etc/global.properties</code> .	<ul style="list-style-type: none"> • File level backup • Frequency – on change, periodic
vCloud Director VC Proxy	Infrastructure	Stateless	None
vCloud Director Console Proxy	Infrastructure	Stateless	None
vCloud Director Database Server	Infrastructure	Dynamic information shared by all cells. The database instance may be shared with other applications.	<ul style="list-style-type: none"> • vCloud database schema level backup • Frequency – periodic
vCenter Server installation files	Infrastructure	Static information consists of product binaries, and configuration files. See <i>Backup vCenter Chargeback database and configuration files</i> (http://kb.vmware.com/kb/1026796).	<ul style="list-style-type: none"> • VM snapshot • Frequency – Once

vCenter Server Log Files	Infrastructure	Dynamic generation generated by each vCenter Server.	<ul style="list-style-type: none"> • File level backup • Frequency – periodic
vCenter Database Server	Infrastructure	Dynamic information shared by all cells. There may be multiple database servers in a multi-VC configuration.	<ul style="list-style-type: none"> • vCenter database schema level backup • Frequency – periodic
vCloud Organizations	Content	Dynamic information virtual datacenter, networks, vApps, virtual machines, users, catalogs.	<ul style="list-style-type: none"> • vCloud REST API • Frequency – periodic
vCloud Provider Resources	Content	Provider virtual datacenters, provider networks, network pools.	<ul style="list-style-type: none"> • vCloud REST API • Frequency – periodic
Orchestrator Application database	Orchestration	Contains the workflow engine library (workflows, actions, policy templates, configuration elements, resource elements, web views) and the workflow engine current state (workflows status, events).	Very frequently
Orchestrator Plug-ins databases	Orchestration	Contains plug-ins database objects.	Very frequently
Orchestrator Application and plug-ins configuration	Orchestration	Contains the configuration.	Upon configuration change
Orchestrator Application and plug-ins	Orchestration	Contains the vCenter Orchestrator Server application.	Upon application or plug-ins upgrade
Orchestrator Application logs	Orchestration	Contains the vCenter Orchestrator Server logs.	Very frequently

Appendix G: Upgrade Checklists

These checklists cover the upgrade of vCloud Director and associated components. Review all applicable product documentation for a detailed upgrade process.

Phase 1

Upgrade vCloud Director Cells

- Verify operating system, database, and other component compatibility with target vCloud Director version. See the online *VMware Compatibility Guide* (<http://www.vmware.com/resources/compatibility/search.php>).
- Obtain updated vCloud Director installation package.
- Backup vCloud Director configuration and response files.
- Perform backup of vCloud Director database and vCenter database(s).
- If multiple cells exist, use cell management tool to quiesce and shutdown services on each server (see the *vCloud Director Installation and Configuration Guide*).
- Upgrade vCloud Director software on all servers, but do not start the services yet. See the *vCloud Director Installation and Configuration Guide* for recommendations on minimizing the interruption of vCloud Director portal service.
- Upgrade the vCloud Director database with scripts included in vCloud Director 1.5 installation.
- Restart the vCloud Director services on upgraded vCloud Director servers.

Caution If Chargeback is in use, upgrade to Chargeback 1.6.2 or later before continuing in order to minimize disruption of metering service. Versions prior to Chargeback 1.6.2 cannot collect data from vCloud Director 1.5.

Note For details, refer to the *vCloud Director Installation and Configuration Guide* (https://www.vmware.com/support/pubs/vcd_pubs.html).

Upgrade vShield Manager and Edge Devices

- Obtain vShield Manager update package. Do *not* deploy a new appliance.
- Perform upgrade of vShield Manager servers.
- Update vShield Manager authentication settings within the vCloud Director portal for each configured vCenter and vShield Manager to utilize directory-based service accounts with appropriate permissions within vCenter
- Reset organization and vApp networks within the vCloud Director portal to redeploy the updated vShield Edge devices.

Note For details, refer to the *vShield Administration Guide* (https://www.vmware.com/support/pubs/vshield_pubs.html).

Upgrade Validation

- Verify vCloud Director version on each cell.
- Within vCloud Director portal, confirm that vCenter and hosts are available.
- Verify version of vShield Manager.
- Verify version of each deployed vShield Edge device.
- If in use, verify that load balancer accurately detects status of all cells.
- Validate service availability through access to vCloud Director organization portals.
- Validate usage metering collection within Chargeback.

Note Refer to the *vShield Administration Guide* for more details

Phase 2

Upgrade vCenter Server

- Verify operating system, database, and other component compatibility with target vCenter version.
- Perform backup of vCenter Server configuration files.
- Backup vCenter database using a method appropriate for configured databases.
- Disable the vCenter server within the vCloud Director system portal.
- Perform upgrade installation of vCenter Server.
- Enable the vCenter server within vCloud Director system portal.
- Install VMware Update Manager and register with vCenter Server.

vCenter Upgrade Validation

- Validate vCenter version and availability status within the vCloud Director system portal.
- Validate usage metering collection within Chargeback.

Phase 3

Upgrade Hosts

- Backup host configurations.
- Place host in maintenance mode, and confirm that vCloud Director detects that host is unavailable.
- Perform upgrade to ESXi 5, removing any incompatible third-party packages that may be installed.
- Reconnect upgraded host within vCenter to upgrade vCenter agents.
- Disable maintenance mode.

Host Upgrade Validation

- Within the vCloud Director portal refresh status to verify that new agents are installed and hosts are listed as available.
- Verify detected ESXi version within vCloud Director system portal.

Phase 4**Additional Upgrades**

- Upgrade all hosts that are connected to datastores and vSphere Distributed Switches.
- Upgrade VMFS Datastores to VMFS-5.
- Upgrade vSphere Distributed Switches.
- Modify provider virtual datacenters to support virtual hardware version 8, if desired.
- Modify organization virtual datacenters to enable fast provisioning, if desired.