VMware vCloud® Architecture Toolkit™
for Service Providers

Architecting a VMware vCloud Director® Solution for VMware Cloud Providers™

Version 2.9

January 2018

Tomas Fojta

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

Contents

List of Figures

List of Tables

Introduction

VMware Cloud Providers™ with the VMware Hybrid Cloud Powered Services badge provide their enterprise customers with a true hybrid cloud experience—full compatibility for their existing virtual applications running on an internal VMware vSphere® environment using the same set of tools to manage their internal and cloud virtual data centers.

The service provider must be validated by VMware to verify that their public cloud fulfills the following hybrid requirements:

- Cloud service is built on vSphere and VMware vCloud Director®
- VMware vCloud user API is exposed to cloud tenants
- Cloud supports Open Virtualization Format (OVF) for bidirectional workload movement

This document provides guidance on how to design vSphere and vCloud Director and other supporting technologies to enable any service provider to obtain the VMware Hybrid Cloud Powered Service badge.

Technology Mapping

To build a hybrid cloud, a prescriptive set of technologies must be used to secure required compatibility and mobility of workloads with standardized API access.

vSphere together with VMware NSX® is the basis for the underlying compute and networking virtualization platform, which can use any compatible physical resources (servers, storage, network switches, and routers).

VMware Virtual SAN™ can optionally extend platform virtualization services as a hypervisor-converged, scale out alternative to traditional SAN or NAS storage systems.

vCloud Director is used for the cloud management plane. It natively supports, pools, and further abstracts the virtualization platform in terms of virtual data centers. It provides multitenancy features and self-service access for tenants through a native graphical user interface, or through the vCloud API, which allows programmable access both for the tenants (for consumption) and for the provider (cloud management). The vCloud API also provides the framework for extension services which enable VMware Cloud Providers or 3rd Independent Software Vendors to add additional services onto existing or new API objects. The vCloud API allows service providers to differentiate from others and to build their own or use a third-party user interface. The user side of vCloud API that is exposed to the tenants allows usage of the same VMware or third-party tools for management. As of vCloud Director 8.20, a separate vCloud Director API for VMware NSX® is used to expose networking services provided by the VMware NSX platform. It shares the authentication mechanism with vCloud API and acts as a proxy API for multitenant safe access to NSX APIs.

VMware vCloud Usage Meter collects consumption data of various VMware software components to provide per-usage licensing of VMware Cloud Provider Program bundles.

Additional VMware components can be optionally used for operations and business support systems integration:

- VMware vCloud Connector® for simplifying virtual machine, template, and ISO image migrations to and from the public cloud for customers as well as for the provider (public catalog management).
- VMware vCenter® Chargeback™ is the legacy tool for metering and usage reporting. The vCenter Chargeback API integrates with existing billing services.
- VMware vRealize® Business™ for Cloud is replacing vCenter Chargeback as the consumption metering, costing and reporting tool. It also provides an API for integration with existing billing services.
- VMware vRealize Operations Manager™ for performance and capacity monitoring and analysis with additional Management Packs that extend its monitoring capabilities.
- VMware vRealize Log Insight™ for centralized log management and analytics.
- VMware vRealize Orchestrator™ for extension services or for automation recurring tasks (tenant onboarding and lifecycle).
- VMware Site Recovery Manager™ for disaster recovery protection of cloud management components.

2.1 Glossary of Terms

Table 1. Glossary

Term	Definition
Allocation Pool	A pool of allocated resources for which a certain percentage of compute resources is guaranteed.
Availability Zone	Single failure domain of resources.
Catalog	A repository of vApp templates and media available to users for deployment. Catalogs can be published and shared between organizations in the same vCloud environment.
Dedicated Cloud	Cloud resources dedicated to one tenant on dedicated physical infrastructure.
Edge Gateway	Virtual appliances that provide network edge security. Edge gateways connect the isolated, private networks of cloud tenants to the public side of the provider network through services such as routing, firewall, NAT, DNS relay, DHCP, site-to-site IPsec VPN, and load balancing.
External Networks	External networks provide internet connectivity to organization networks and are backed by port groups configured for Internet accessibility.
Management Cluster	Physical and virtual resources dedicated for management purposes.
Network Pools	Collections of isolated Layer 2 virtual networks available to vCloud Director for the automated deployment of organization and vApp networks.
On-Demand Cloud	Cloud resources that are committed to tenant and billed only when used.
Organization	The unit of multitenancy representing a single logical security boundary. An organization contains users, virtual data centers and catalogs.
Organization Administrator	Administrator for a vCloud Director organization responsible for managing provided resources, network services, users, and vApp policies.
Organization VDC Networks	Organization VDC networks are instantiated through network pools and bound to a single organization VDC or shared across Organization. Organization VDC networks can be isolated, routed, or directly connected to an external network.

Term	Definition
Organization Virtual Data Center	A subgrouping of compute, network, and storage resources allocated from a provider virtual data center and assigned to a single organization. A virtual data center is a deployment environment where vApps can be instantiated, deployed, and powered on. Organization virtual data centers cannot span multiple organizations.
Pay-As-You-Go	Provides the illusion of an unlimited resource pool. Resources are committed only when vApps are created in the organization virtual data center.
Provider Virtual Data Center	A grouping of compute and storage resources from a single VMware vCenter Server®. A provider virtual data center consists of one or more resource pools and one or more datastores. Provider virtual data center resources can be shared with multiple organizations.
Reservation Pool	The compute resources allocated to the organization virtual data center are completely reserved and dedicated.
Resource Group	Compute, storage and networks for tenant workloads managed by one vCenter Server.
vApp	Container for software solution in the cloud. A vApp is the standard unit of deployment for vCloud Director. It contains one or more virtual machines, networks, and network services, has power-on operations, and can be imported or exported.
vApp Edge	Virtual router instantiated inside a vApp to provide routing, NAT, firewalling and DHCP services for vApp networks.
vApp Network	A network that connects virtual machines within a vApp, deployed by a consumer from a network pool.
vCenter Chargeback	A metering and cost calculation solution that enables accurate cost measurement, configuration, analysis, and reporting for virtualized environments. vCenter Chargeback provides the ability to map IT costs to business units or external customers.
vCloud API	An open, RESTful API for providing and consuming virtual resources from the cloud. It enables deployment and management of virtualized workloads in internal and external clouds.
vCloud Director	A software solution providing the interface, automation, and management feature set to allow service providers to supply vSphere resources as a Web-based service.

Term	Definition
vCloud Director cell	A cell is runtime of vCloud Director services on a physical or virtual machine. Multiple cells within one vCloud Director instance can be grouped together connecting to one vCloud Director database for load balancing and high availability.
Virtual Private Cloud	Cloud resources dedicated to one tenant within a shared infrastructure.

Deployment Model Considerations

3.1 Service Offerings

The actual deployment model depends on the particular service provider offering. This section discusses the most typical configurations and their deployment considerations.

Note It is possible to combine service offerings.

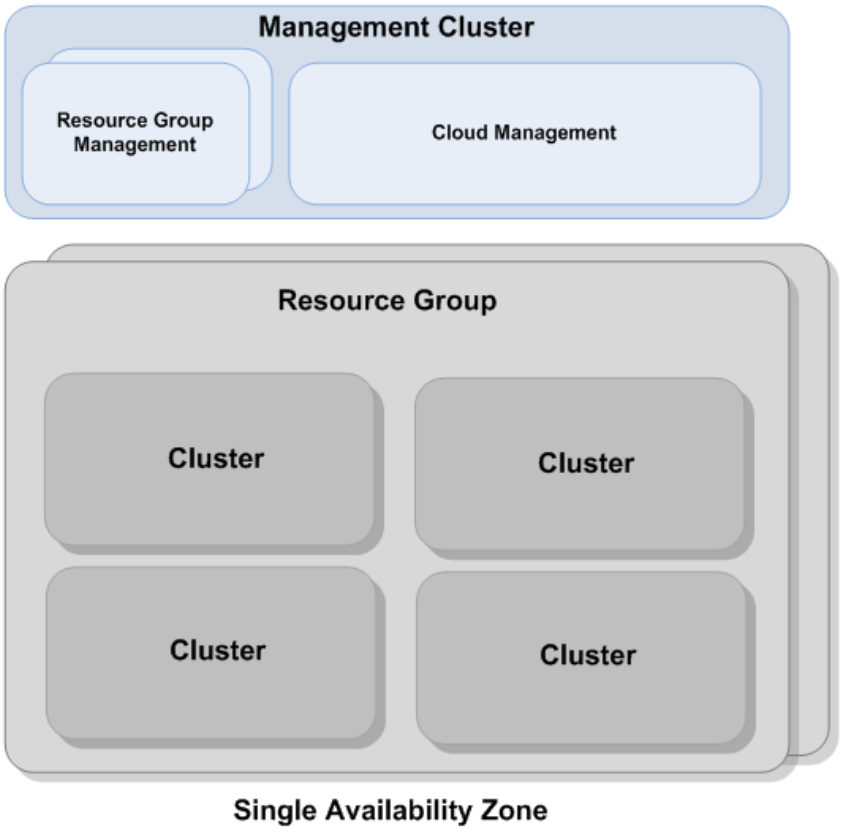
3.1.1 IaaS Single Availability Zone

Infrastructure as a Service (IaaS) is defined as service that provides compute, storage, and networking resources to end users as building blocks for deploying operating systems and applications (workloads).

The availability zone is generally defined as a location (data center) which, while offering an availability SLA (for example, 99.9 percent uptime) to customer workloads, is a single fault domain. Major outage of such a location results in total disruption of the services running there.

The following figure represents a typical deployment model of the IaaS single availability zone offering.

Figure 1. Single Availability Zone



A single management cluster is running management components of the cloud (vCloud Director) as well as management components of resource groups (vCenter Server instances and supporting systems). Each resource group is represented by a vSphere environment consisting of multiple clusters managed by a single vCenter Server. There can be more than one resource group to scale beyond the limits of one vCenter Server, but they are all located within the same availability zone.

3.1.2 IaaS Multiple Availability Zones

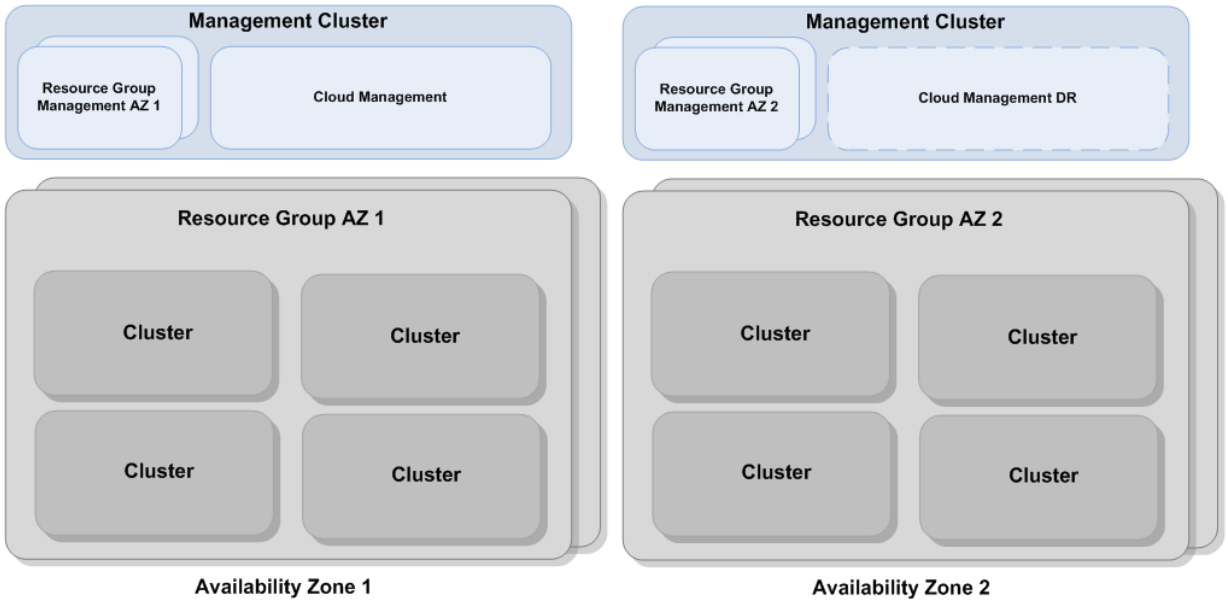
With the multiple availability zone concept, end users can deploy their applications in a distributed model, so when one availability zone fails, the applications continue running. The end user is responsible for maintaining the application state between zones. The provider must have two sites (data centers) distant enough to decorrelate their probability of failure, but close enough to have them under the same cloud management. The provider must also secure disaster recovery of the cloud management components.

There are two options to design multiple availability zones, and they differ as to how resource group management components are deployed – in distributed or stretched configurations.

3.1.2.1 Distributed Resource Groups

With a distributed resource group, each set of resource groups is in an availability zone together with its management components, as shown in the following figure.

Figure 2. Distributed Resource Groups



Cloud management components are located in a primary site. In the event of a disaster or outage, the management components can be failed over to the secondary site to reduce downtime. Resource groups and their management components are isolated per availability zone with no failover. Downtime in an availably zone would mean complete downtime for the resource group, but not for the cloud management components.

Advantages of this design:

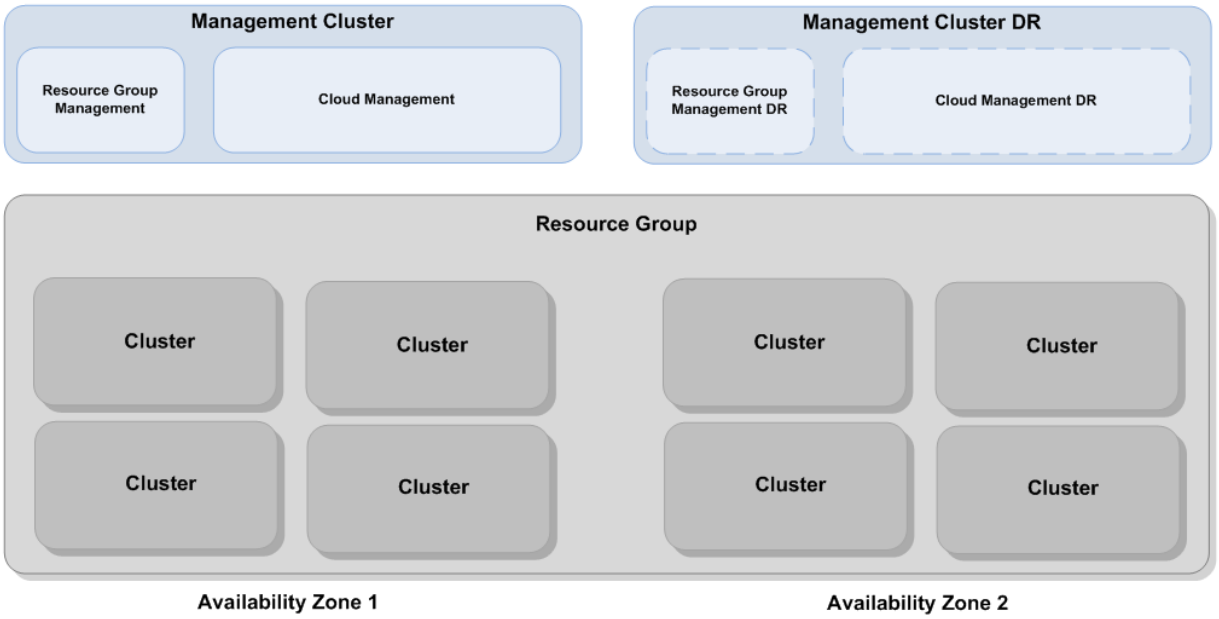
- Access to the cloud management components is provided even in the event of failure.
- With proper application design, the tenant can avoid application downtime.
- The service provider can offer a disaster recovery service.

3.1.2.2 Stretched Resource Group

This less frequently used design stretches a resource group across both availability zones, with some clusters in the first site and others in the second. Resource group management is located in the primary site with failover to the second site.

The main advantage of this design is that it is possible to stretch organization networks across both sites and achieve Layer 2 adjacency for customer workloads. Disadvantages are lack of scalability, the need for disaster recovery of resource group management components, and the possibility of logical corruption of the management components that can down both availability zones.

Figure 3. Stretched Resource Group



Note None of these multisite designs requires stretched storage (storage metro cluster solution). However, stretched storage can be used for disaster recovery protection of the management layer. In such a case, distance between sites is limited by the supported round-trip latency of the storage network (usually 5–10 ms).

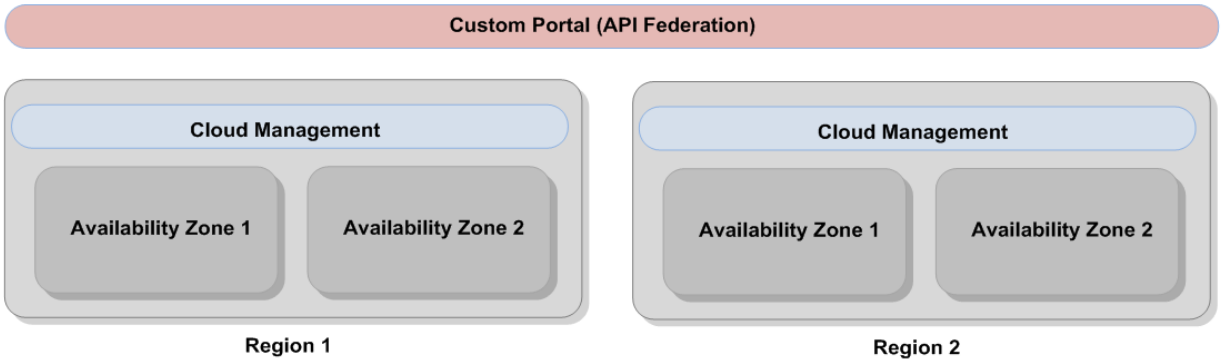
vCloud Director 8.20 and its components support a maximum 40 ms of network round-trip latency between sites (approximately 3300 km). However, the user experience diminishes with larger latencies and slower bandwidth.

3.1.3 IaaS Multiple Regions

Regions are completely isolated cloud instances with large distances between them. Tenants can place their workloads closer to the end users, and protect from a disaster that affects the whole region. The large distance means that tenant applications must keep state asynchronously.

vCloud Director 8.20 does not yet provide native support for multiple instances (federation). Therefore, the provider must use its own or a third-party portal on top of vCloud Director.

Figure 4. Regions



3.1.4 IaaS with Disaster Recovery SLA

None of the previous deployment options provides guaranteed disaster recovery SLAs to tenant workloads. It is the responsibility of the tenant to secure application availability by means of application-level replication. However, legacy applications do not support replication of state at the application level and must rely on replication provided by the infrastructure.

When this is the case, the provider can leverage storage replication. In the case of loss of one availability zone, the provider can restore and power on all the workloads in another availability zone.

Figure 5. DR with Storage Replication

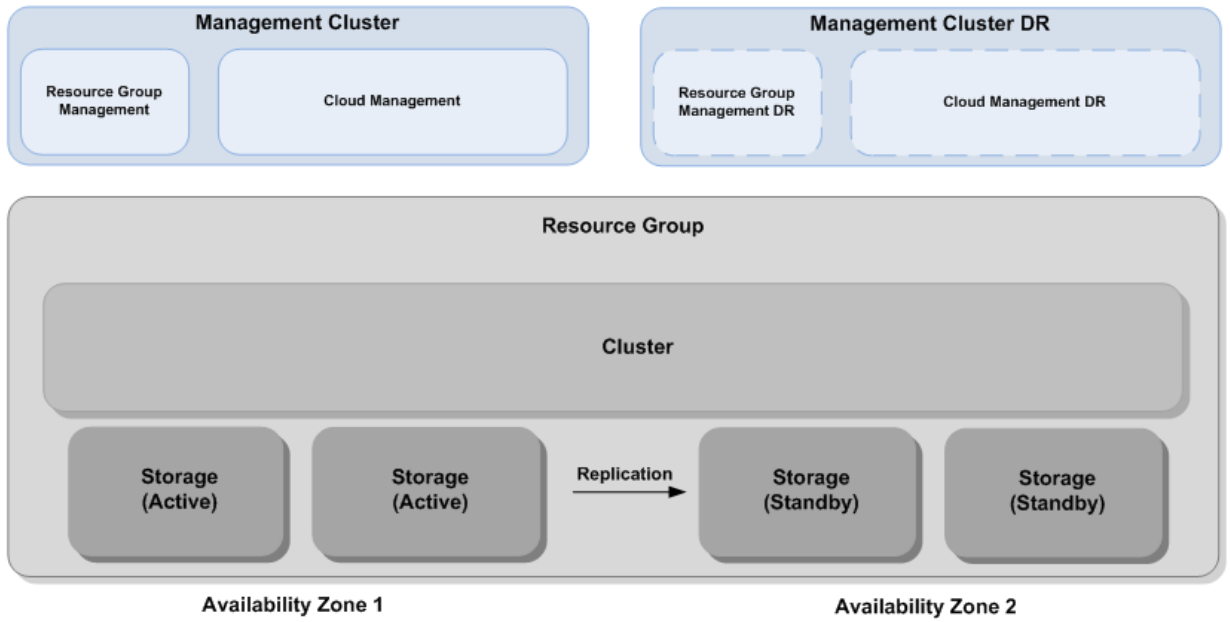
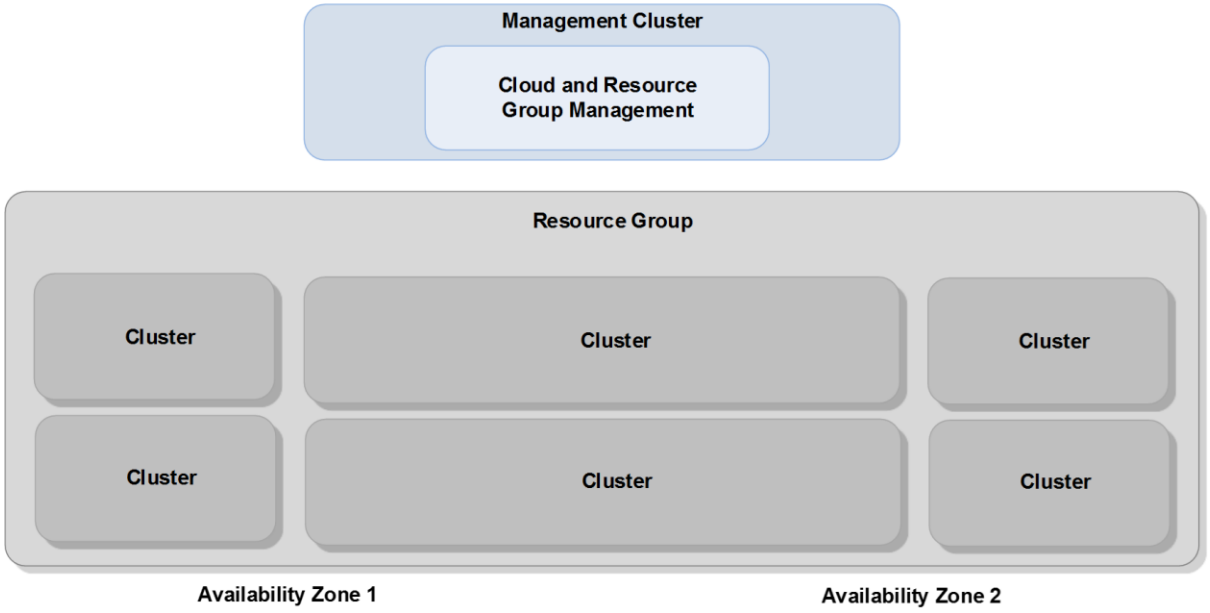


Figure 6. DR with vSphere Metro Storage Cluster



There are two options for designing disaster recovery with storage replication:

- Generic synchronous or asynchronous storage replication with manual or scripted failover to the other site. For more details, see the *VMware vCloud Director Infrastructure Resiliency Case Study* at http://www.vmware.com/files/pdf/techpaper/VMware-vCloud-Directore-Infrastructure-resiliency-whitepaper.pdf.
- Storage solution that supports vSphere Metro Storage Cluster with automated failover provided by VMware vSphere High Availability. For more details, see the *VMware vSphere Metro Storage Cluster Case Study* at http://www.vmware.com/files/pdf/techpaper/vSPHR-CS-MTRO-STOR-CLSTR-USLET-102-HI-RES.pdf.

3.1.5 Consumption Models

Service providers can tailor their public cloud offering based on the customer's desired consumption. The following models are available:

- On-demand cloud
- Virtual private network
- Dedicated cloud

3.1.5.1 On-Demand Cloud

The customer does not subscribe to cloud resources up front, and pays based only on consumption. This is usually beneficial for bursty or seasonal workloads, which scale up and down in time. The customer consumes the resources from a seemingly infinite pool that the provider must maintain. Predicting such demand might be challenging for the provider. Therefore, the provider compensates by making such resources more expensive or more oversubscribed.

This offering directly maps to the organization virtual data center (Org VDC) pay-as-you-go allocation model.

3.1.5.2 Virtual Private Cloud

The customer buys a chunk of resources in terms of virtual private cloud (VPC) consisting of compute (CPU and memory), storage, and networking. Inside the VPC, they deploy workloads up to the total capacity of resources purchased. This consumption model is beneficial for steady workloads because the customer knows their needs and can commit to buying the resources for minimum period of time (usually at least a month).

The oversubscription of resources is managed by the provider and the customer has no control over it. The size of VPC can be arbitrary from as small as a portion of one host to as large as multiple vSphere clusters, thanks to the sharing of underlying physical and vSphere virtual infrastructure among different tenants.

This offering directly maps to the Org VDC Allocation Type allocation model.

3.1.5.3 Dedicated Cloud

In cases where the customer does not run workloads on the same physical infrastructure as other tenants, or because of licensing reasons, they can purchase a dedicated cloud. Multiple physical hosts (which form a vSphere cluster) are dedicated to the tenant, who has full control over oversubscription of workloads that are deployed in their dedicated cloud.

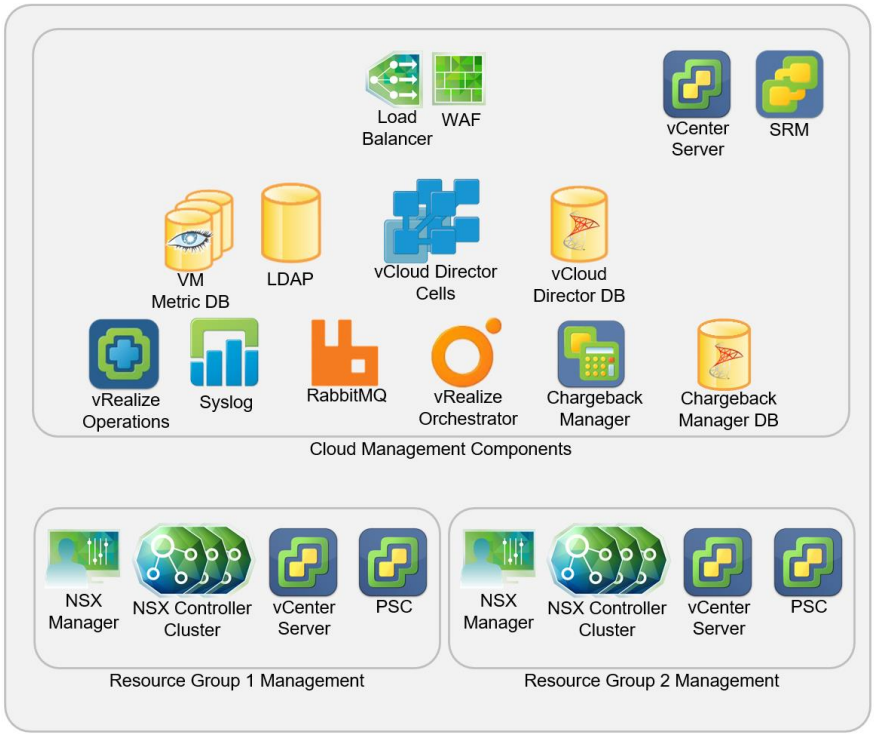
This offering directly maps to the Org VDC Reservation Type allocation model.

Architectural Overview

The typical public cloud architecture consists of management components that are usually deployed in the management cluster, and of resource groups for hosting the tenant workloads. Some of the reasons for the separation include:

- Different SLAs (availability, recoverability, performance) for management components and for tenant workloads
- Separation of duties (resource groups are managed by vCloud Director)
- Consistent management and scaling of resource groups

Figure 7. Management Components



The management cluster runs the cloud management components and resource group management components.

Note Some management components (such as VMware NSX Controller™ cluster nodes) must be deployed to the resource groups.

Resource groups are independent infrastructure domains represented by virtualized compute, networking, and storage, each managed by its own vCenter Server.

Cloud Management Components

5.1 Management vCenter Server

In all but small environments, VMware recommends that the management cluster be managed by its own vCenter Server. This allows disaster recovery protection with Site Recovery Manager and further enhances the separation of resource groups. The use of dedicated storage (for example Virtual SAN) means that resource groups do not affect the storage performance of management components.

5.2 vCloud Director

5.2.1 vCloud Director Cells

vCloud Director functionality is provided by stateless cells—Linux (Red Hat Enterprise Linux or CentOS) machines with the vCloud Director binaries. Each cell contains a set of services such as transfer service, console proxy, vCenter listener, UI services and others. Each cell usually has at least two IP addresses—a primary for the vCloud Director user interface or API, and a secondary for VMware remote console proxy because both run by default on port TCP 443. However, it is possible to move services to non-default ports. The cells communicate with each other through an ActiveMQ message bus on the primary interface. They also share a common vCloud Director database where the cells persist configuration and state data. The transfer service requires that all cells have access to a common shared folder—usually an NFS mount.

The following are vCloud Director cell design considerations:

- Use at least 2 vCPUs and 6 GB RAM for the cell VM.
- Deploy at least *N*+1 cells where *N* is number of resource groups or *n*/3000+1 where *n* is the expected number of powered-on VMs (use the larger of the two).
- To avoid a split-brain scenario, verify that the cells can communicate with each other through the message bus on the primary network interface.
- vCloud Director starts vCenter Server proxy (listener) for each connected vCenter Server. Distribute the vCenter Server proxy among the cells so none is running more than one proxy. This can be done manually by triggering reconnection of the vCenter Server. For the reconnected vCenter Server, a new vCenter Server proxy is started on the least utilized vCloud Director cell.
- Use the same *consoleproxy* certificate on all cells.
- It is possible to steer the load-balanced traffic to a specific cell. However, the cells are not site-aware and the tasks are randomly distributed among them. VMware recommends keeping the cells on one site with their database and transfer share and recovering them together in case of disaster recovery.
- Use a Web application firewall to terminate vCloud HTTPs traffic at the load balancer and to apply Layer 7 firewall rules. You can filter based on URL, source IP, or authentication header to protect access to certain organizations or API calls (provider scope). The traffic between the load balancer and cells must be HTTPs-encrypted as well.
- Enable *X-Forwarded-For (XFF)* HTTP header insertion on the load balancer to track the source IP of requests in vCloud Director logs.
- The VMware remote console proxy traffic cannot be terminated at the load balancer and must be passed through to the cells because it is a proprietary socket SSL connection. The WebMKS (native HTML 5 web console used exclusively as of vCloud Director 8.20) requires the use of TCP port 443 on the load balancer virtual IP address.
- The sticky sessions on the load balancer are recommended (for performance reasons), but not required, because the session state is cached at the cell level and also stored in the vCloud database.

- Use round-robin or least-connection load-balancing algorithm to share the load.
- Use the following load balancer health checks for the cell pool:

  GET `http://<cell_HTTP_IP>/api/server_status` (expected response is "*Service is up*").

  GET `https://<cell_consoleproxy_IP>` (expected response 200) or a simple TCP 443 check.
- After installing the first vCloud Director cell, back up certificates and the `$VCLOUD_HOME/etc/responses.properties` file, which contains all necessary information to deploy new or additional cells (for example, database password).
- Verify that cell transfer share is accessible for all cells and that the Linux vCloud user has write permissions. The size of the transfer share must be large enough to store all concurrent OVF or ISO imports/exports or migrations between resource groups (for example 10 concurrent 50 GB transfers require up to 500 GB of transfer share capacity). If catalog publishing with early catalog export is used, extend transfer share capacity by the size of the exported catalog.
- Redirect vCloud Director logs to an external syslog by editing the `$VCLOUD_HOME/etc/log4j.properties` file or by installing the vRealize Log Insight agent on the cell.
- For large environments deploy more vCloud Director cells (scale out approach). To scale up a single cell it is possible to increase its vCPU, memory, JVM heap size, database connection pool, and jetty connections in `$VCLOUD_HOME/bin/vmware-vcd-cell`. See the following table.

Table 2. vCloud Director 8.20 Cell Performance Tweaks

Attribute	Location	Default Value	Recommended Value for large environments
Cell vCPU	Cell VM	2 vCPU	4 vCPU
Cell Memory	Cell VM	6 GB RAM	12 GB RAM
JVM Heap Size	`$VCLOUD_HOME/bin/vmware-vcd-cell`	JAVA_OPTS:--Xms1024M -Xmx4096M	JAVA_OPTS:--Xms2048M -Xmx8192M
Database.pool.maxActive	`$VCLOUD_HOME/etc/global.properties`	75	200
vcloud.http.maxThreads	`$VCLOUD_HOME/etc/global.properties`	128	200
vcloud.http.minThreads	`$VCLOUD_HOME/etc/global.properties`	25	32
vcloud.http.acceptorThreads	`$VCLOUD_HOME/etc/global.properties`	2	16

5.2.2 vCloud Director Database

The following are vCloud Director database design considerations:

- Microsoft SQL and Oracle database servers are supported with following high availability options. See the VMware Knowledge Base article *Supported high availability options for the vCloud Director database (2037802)* (http://kb.vmware.com/kb/2037802).
  - Oracle RAC
  - Microsoft SQL Failover Cluster

  vCloud Director 8.20 additionally supports Microsoft SQL AlwaysOn Availability Groups (https://kb.vmware.com/kb/2148767) in single subnet mode.
- Follow the recommended practices on configuration of the vCloud Director database as specified in the VMware Knowledge Base article, *Installing and configuring a vCloud Director 5.1 or 5.5 database (2034540)* (http://kb.vmware.com/kb/2034540).
- Collocate the vCloud Director database with vCloud Director cells on the same site to minimize network latency.
- For Microsoft SQL, additional performance improvements can be achieved by changing MS SQL TDS to 1472 (to avoid fragmentation) or to increase it to 8060 while enabling jumbo frames on the vCloud cell – MS SQL network.
- Make sure that the database supports necessary number of connections generated by vCloud Director cells.

5.3 VM Metric Database

Beginning with vCloud Director 5.6, virtual machine performance and resource consumption metrics are collected and historical data is provided for up to two weeks.

Table 3. Virtual Machine Performance and Resource Consumption Metrics

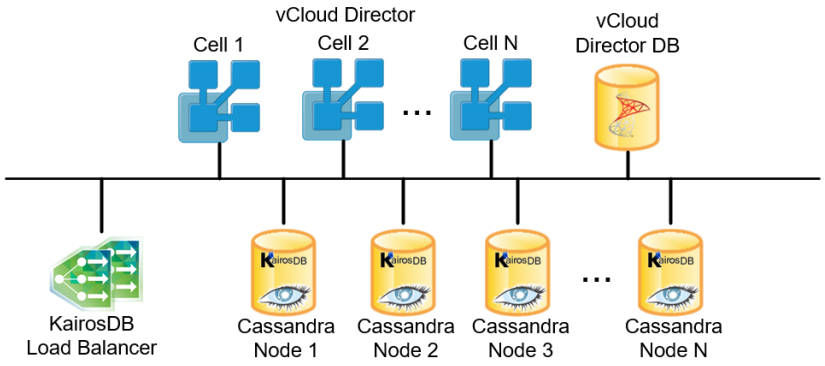
Metric Name	Type	Unit	Description
cpu.usage.average	Rate	Percent	Host view of average actively used CPU as a percentage of total available
cpu.usagemhz.average	Rate	Megahertz	Host view of actively used CPU as a raw measurement
cpu.usage.maximum	Rate	Percent	Host view of maximum actively used CPU as a percentage of total available
mem.usage.average	Absolute	Percent	Usage as percentage of total configured or available memory
disk.provisioned.latest	Absolute	Kilobytes	Storage space potentially used
disk.used.latest	Absolute	Kilobytes	Storage space actually used
disk.read.average	Rate	Kilobytes per second	Read rate aggregated across all datastores
disk.write.average	Rate	Kilobytes per second	Write rate aggregated across all datastores

Retrieval of both current and historical metrics is available through vCloud API. The current metrics are directly retrieved from the vCenter Server database with the Performance Manager API. The historical metrics are collected every 5 minutes (with 20 seconds granularity) by the StatsFeeder process running on the cell with vCenter Server Proxy and pushed to persistent storage—Cassandra NoSQL database cluster with KairosDB database schema and API.

Note The usage of KairosDB will be deprecated in the future vCloud Director releases.

The following figure depicts the recommended VM metric database design. Multiple Cassandra nodes are deployed in the same network. A KairosDB database runs on each node, which also provides an API endpoint for vCloud cells to store and retrieve data. For high availability, load balance all KairosDB instances behind a single virtual IP address that is configured by the cell management tool as the VM metric endpoint.

Figure 8. VM Metric Database Design



The following are VM metric database design considerations:

- Currently only KairosDB 0.9.1 and Cassandra 1.2.x/2.0.x are supported.
- Minimum cluster size is three nodes (number of nodes must be equal or greater than the replication factor). Use the scale-out rather than scale-up approach because Cassandra performance scales linearly with the number of nodes.
- Estimate I/O requirements based on expected number of VMs, and size the Cassandra cluster and its storage properly.

  n – Expected number of VMs

  m – Number of metrics per VM (currently 8)

  t – Retention (days)

  r – Replication factor

  Write I/O per second = $n \times m \times r / 10$

  Storage = $n \times m \times t \times r \times 114$ KB

  For 30,000 VMs, the I/O estimate is 72,000 write IOPS and 3,288 GB of storage (worst-case scenario if data retention is 6 weeks and the replication factor is 3).
- Enable Leveled Compaction Strategy (LCS) on the Cassandra cluster to improve read performance.
- Install JNA (Java Native Access) version 3.2.7 or later on each node because it can improve Cassandra memory usage (no JVM swapping).

- For heavy read utilization (many tenants collecting performance statistics) and availability, VMware recommends increasing the replication factor to 3.
- Recommended size of one Cassandra node: 8 vCPUs (more CPU improves write performance), 16 GB RAM (more memory improves read performance), and 2 TB storage (each backed by separate LUNs/disks with high IOPS performance).
- KairosDB does not enforce data retention policy. Therefore, old metric data must be regularly cleared with a script.

The following example deletes one month's data:

```
#!/bin/sh

if [ "$#" -ne 4 ]; then
    echo "$0 <kairosdbvip> port month year"
    exit
fi

let DAYS=$(( ( $(date -ud 'now' +'%s') - $(date -ud "${4}-${3}-01 00:00:00"
+'%s')  )/60/60/24 ))
if [[ $DAYS -lt "42" ]]; then
 echo "Date to delete is in not before 6 weeks"
 exit
fi
METRICS=( `curl -s -k http://$1:$2/api/v1/metricnames -X GET|sed -e
's/[{}]/''/g' | awk -v k="results" '{n=split($0,a,","); for (i=1; i<=n; i++)
print a[i]}'|tr -d '[":]'|sed 's/results//g'|grep -w
"cpu\|mem\|disk\|net\|sys"` )
echo $METRICS

for var in "${METRICS[@]}"
do
for date in `seq 1 30`;
  do
    STARTDAY=$(($(date -d $3/$date/$4 +%s%N)/1000000))
    end=$((date + 1))
    date -d $3/$end/$4 > /dev/null 2>&1
    if [ $? -eq 0 ]; then
       ENDDAY=$(($(date -d $3/$end/$4 +%s%N)/1000000))
       echo "Deleting $var from " $3/$date/$4 " to " $3/$end/$4
       echo '
       {
          "metrics": [
          {
            "tags": {},
            "name": "'${var}'"
          }
          ],
          "cache_time": 0,
          "start_absolute": "'${STARTDAY}'",
          "end_absolute": "'${ENDDAY}'"
       }' > /tmp/metricsquery
       curl http://$1:$2/api/v1/datapoints/delete -X POST -d
@/tmp/metricsquery
    fi
  done
done
```

```
rm -f /tmp/metricsquery > /dev/null 2>&1
```

Note The space gains are not seen until data compaction occurs and the delete marker column (tombstone) expires (by default 10 days). This can be changed by editing `gc_grace_seconds` in the `cassandra.yaml` configuration file.

- KairosDB v0.9.1 uses the Quorum consistency level both for reads and writes. Quorum is calculated as rounded down (replication factor + 1) / 2 and, for both reads and writes, the quorum number of replica nodes must be available. Data is assigned to nodes through a hash algorithm and every replica has equal importance. The following table provides guidance on replication factor and cluster size configurations.

Table 4. Cassandra Configuration Guidance

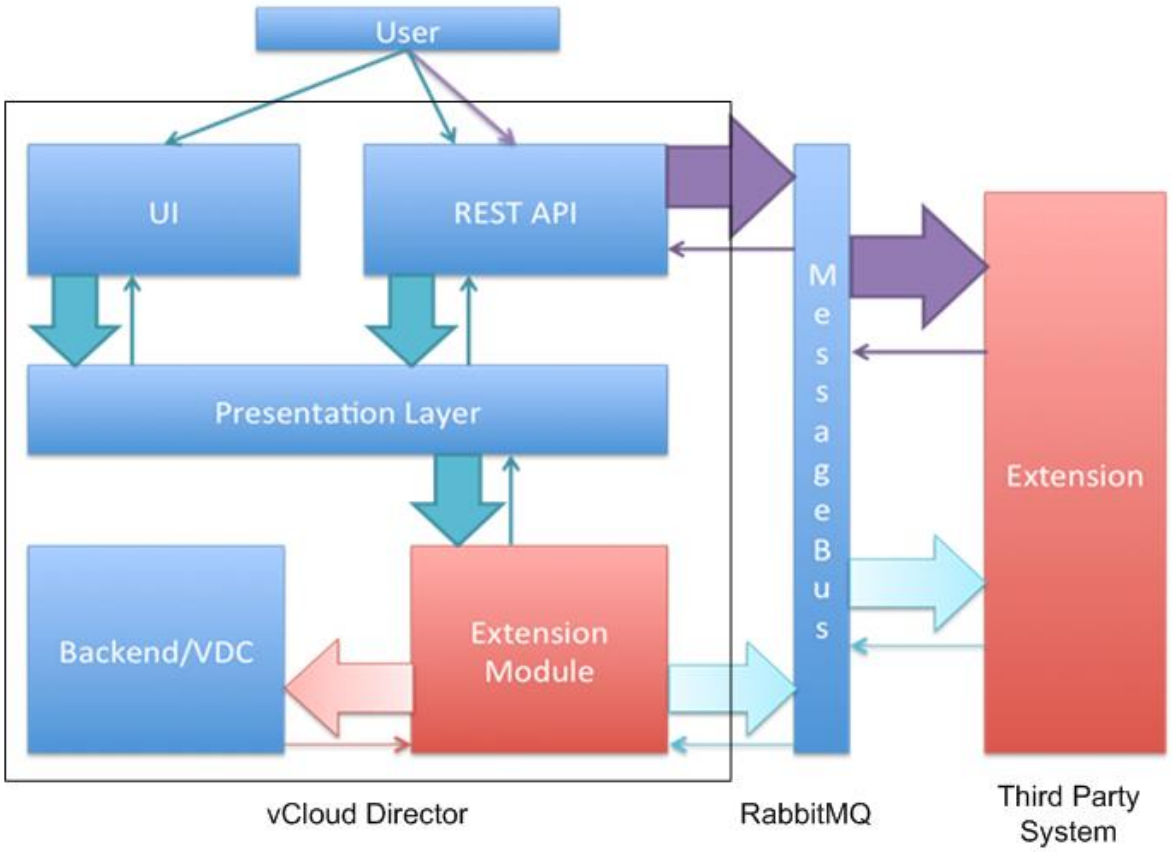
Repl. Factor	Cluster Size	Node Amount of Data	Quorum	Availability
1	1	100%	1	Does not tolerate any node loss
1	2	50%	1	Does not tolerate any node loss
1	3	33%	1	Does not tolerate any node loss
2	2	100%	2	Does not tolerate any node loss
2	3	67%	2	Does not tolerate any node loss
2	4	50%	2	Does not tolerate any node loss
3	3	100%	2	Tolerates loss of one node
3	4	75%	2	Tolerates loss of one node
3	5	60%	2	Tolerates loss of one node
4	4	100%	3	Tolerates loss of one node
4	5	80%	3	Tolerates loss of one node
5	5	100%	3	Tolerates loss of two nodes
5	6	83%	3	Tolerates loss of two nodes

5.4 Pivotal RabbitMQ

vCloud Director functionality can be extended in two different ways:

- vCloud Director and the vCloud API include a framework for integration of extensions services that a vCloud API client can access as though they were native services. In addition to service-specific objects or operations they provide, extension services can implement new operations for existing API objects.
- vCloud messages provide the capability to connect vCloud Director with external systems by posting notifications or blocking task messages to AMQP-based enterprise messaging brokers.

Figure 9. vCloud Director Extensions

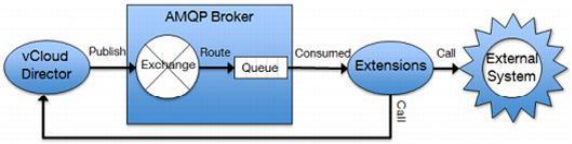


Both options rely on an AMQP message broker, which must be installed. The AMQP service must be configured in vCloud Director. VMware recommends Pivotal RabbitMQ as the AMQP broker.

vCloud Director AMQP service (the publisher) sends messages to AMQP exchange which then routes them to specific queues based on a routing key. The external systems (the consumers) connect to queues and listen to the messages. See VCD-nclient VMware Fling[1] for a quick introduction to the feature.

[1] https://labs.vmware.com/flings/vcd-nclient

Figure 10. AMQP Messages Architecture

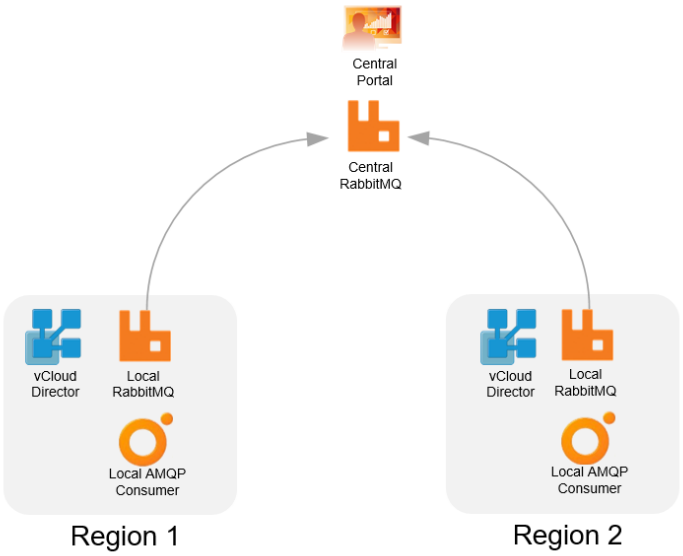


The following are RabbitMQ design considerations:

- vCloud Director cells retry delivering messages to the AMQP broker until the configured timeout period is reached.
- vCloud System Administrator is informed by email when the connection to the AMQP server is lost.
- RabbitMQ broker can be clustered for high availability. All definitions (exchanges, bindings, users, and so on) are mirrored across the entire cluster. You can either use load-balanced RabbitMQ nodes (configure AMQP service in vCloud Director with the virtual IP of the load balancer), or each vCloud cell can have RabbitMQ service installed (point AMQP service to the localhost address).

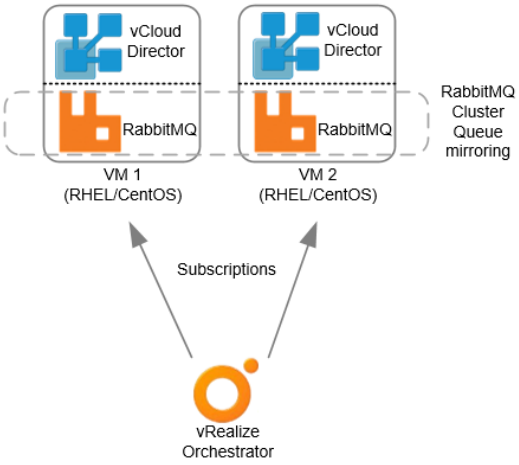
  Clustered RabbitMQ does not handle network partitions well. Therefore, deployment of nodes across sites is not recommended. Split-brain scenario causes some messages not to be delivered properly. Monitoring for cluster state is essential. Take care when RabbitMQ nodes are powered off and on. The last clustered node that shuts down must be powered on first, otherwise the cluster is not started.
- By default, queues reside only on the node where they were declared. For high availability, they must be mirrored across nodes with one master and multiple slaves—this is enabled through a policy. The messages published to the queue are replicated to all slaves. Consumers are connected to the master regardless of which node they attach to, with slaves dropping messages that have been acknowledged at the master.
- The consumer subscribes to the same queue through all RabbitMQ brokers or by using a load-balanced broker and is able to retrieve messages in case of a node failure.
- While VMware still offers download of VMware vFabric® RabbitMQ, the Pivotal Web site offers the latest version. Pivotal provides RabbitMQ support as well.
- RabbitMQ scales up to thousands of messages per second, which is much more than vCloud Director is able to publish. Therefore, there is no need to load balance RabbitMQ nodes for performance reasons.
- RabbitMQ can use SSL both for communication with vCloud Director and the consumers. Enable this when such communication goes over unsecured networks.
- In some rare cases, tenants might need to access organization-specific vCloud messages to connect their own orchestration tools. While this scenario is more likely to be used in dedicated cloud environments, it is possible to accommodate such a request with the RabbitMQ Shovel plug-in. The provider deploys the tenant's own instance of RabbitMQ and the provider RabbitMQ proxies a subset of vCloud messages to the tenant instance.
- The Shovel plug-in can be also used for aggregation of messages from multiple vCloud Director instances into one RabbitMQ receiver. This might be useful in multi-region set-up where the federation portal consumes messages from a central aggregated RabbitMQ instance.

Figure 11. Multi-Region RabbitMQ Example



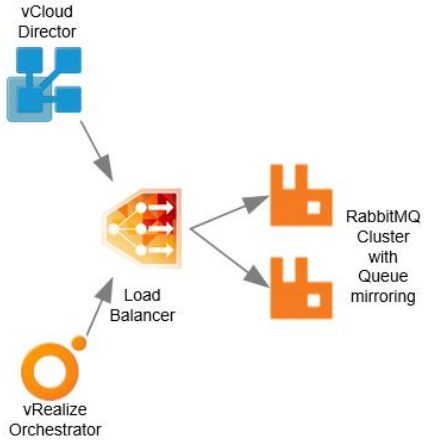
An example of a highly available RabbitMQ design is shown in the following figure. RabbitMQ is co-installed with vCloud Director on each cell virtual machine in a clustered configuration and queue mirroring is enabled. The AMQP Service in vCloud Director points the AMQP host to a localhost URL. The consumer (for example vRealize Orchestrator) establishes a subscription with each node. No load balancer is needed.

Figure 12. RabbitMQ Design Example



For environments with a larger number of cells and where RabbitMQ consumers cannot establish multiple subscriptions (for example vCloud Availability for vCloud Director), VMware recommends using dedicated load-balanced RabbitMQ nodes.

Figure 13. Load Balanced RabbitMQ Cluster



5.5 VMware vCenter Chargeback Manager

VMware vCenter Chargeback Manager™ provides the metering capability to enable cost transparency. In service provider environments, it is typically used to meter tenant usage of the resources and to provide raw data to the existing billing systems.

vCenter Chargeback connects to the resource group vCenter database and retrieves usage data every 30 minutes. The integration with vCloud Director is handled through two additional data collectors, which collect data in 5-minute intervals:

- The vCloud data collector connects to the vCloud Director through the vCloud API to collect resource usage data for each organization. Hierarchies for each organization are generated automatically by vCenter Chargeback.
- The vShield data collector talks to NSX Manager and gathers information on network resources used by vCloud tenants.

The resources described in the following table can be metered.

Table 5. vCenter Chargeback Metrics

Resource	Chargeback Metrics
CPU	• CPU usage/allocated (GHz) vCPU (count)
Memory	• Memory usage/allocated (GB)
Network	• Count of networks • Network services (DHCP, static and dynamic routing, firewalling, NAT, IPSec, VPN, L2 VPN, SSL VPN, load balancing, edge gateway scale and availability, distributed firewall) • External network I/O traffic (transmit/receive, GB/hour)
Disk	• Storage usage (GB) • Disk I/O read/write usage (GB/hour)

Resource	Chargeback Metrics
Custom	• Fixed cost • Licensing

Billing policies define the amount of chargeable computing resources units to be considered. Pricing models consist of resource rates and billing policies. Reports provide cost or usage of particular hierarchy entity during timeframe based on given pricing.

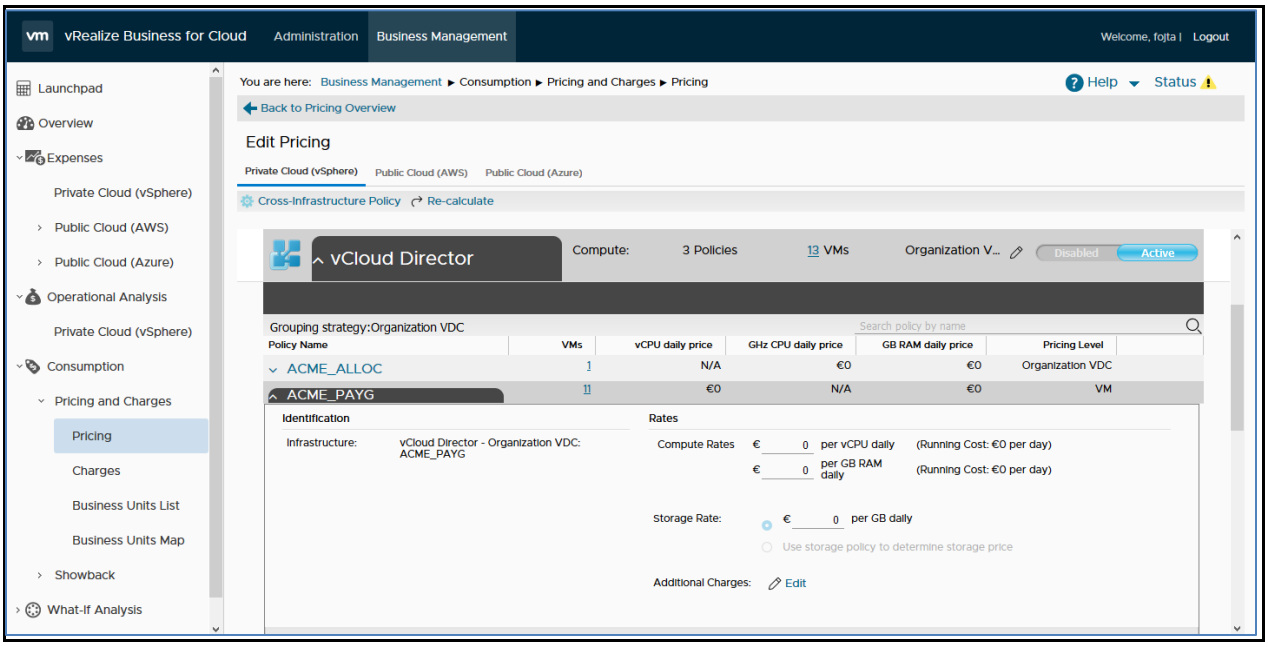
The following are vCenter Chargeback design considerations:

- One vCenter Chargeback data collector can connect up to 5 vCenter Server instances and collect information from up to 15,000 VMs. One vCenter Chargeback instance can connect up to 10 vCenter instances and collect information from 5,000 hierarchies and 35,000 VMs. vCenter statistics collection level might be raised up to level 3 in certain charging scenarios (disk read/write, network transmit) with a retention of at least 30 minutes. The interval vCenter Chargeback collector retrieves data from the vCenter database.
- vCenter Chargeback data collectors can be installed on multiple servers for increased scalability. Deploy at least two of each kind for high availability.
- Metered data is persisted in the vCenter Chargeback database with regular rollup jobs. Data retention with 5-minute granularity is available only for the past 24 hours. It is then rolled up to 1-day averages, which are retained forever.
- MS SQL or Oracle can be used for the vCenter Chargeback database. A sizing tool is available at https://www.vmware.com/support/vcbm/doc/CBM%20DB%20Size%20Calculator.xlsm.
- Chargeback server nodes can be load balanced by a built-in load balancer. An external load balancer is not supported. For high availability, deploy the load balancer separately to one vCPU VM protected by VMware vSphere Fault Tolerance.
- Data collectors can be deployed externally or embedded on vCenter Chargeback server nodes. Load is evenly distributed among them.
- The tenants are typically not provided with the ability to directly view cost and usage reports.
- vCenter Chargeback integrates with Active Directory-based LDAP but the vCenter Chargeback Super User role can be assigned only to a local user. vCenter Chargeback uses resource-based authorization. Therefore, before an LDAP user can access a given vCenter Chargeback hierarchy, they must be assigned to it by Super User.
- Reports can be scheduled through a REST-like API and exported to a billing system in XML format.
- An SDK is available at https://www.vmware.com/pdf/cbm_api_prog_guide_2_5_0.pdf.
- vCenter Chargeback is in deprecated mode and will be replaced by vRealize Business for Cloud in the future.

5.6 vRealize Business for Cloud

vRealize Business for Cloud is a consumption metering, analysis, and reporting tool for both private and public cloud costing. While it can connect and meter multiple solutions, in the vCloud Director context it collects usage data from resource group vCenter Server nodes and NSX Manager nodes and from vCloud Director.

Figure 14. vRealize Business for Cloud



vRealize Business for Cloud is a virtual appliance based with internal PostgreSQL and MongoDB databases and optional external Data Collectors (also virtual appliances). The authentication is provided by a VMware Identity Manager virtual appliance. Alternatively, for non-production use, local authentication can be used.

As of version 7.3, vRealize Business for Cloud still did not achieve full vCenter Chargeback Manager feature parity. Therefore, the service provider must evaluate if the missing features are necessary for the provider use cases. Some of the missing features are as follows:

- External network I/O traffic pricing
- Storage policy pricing
- Allocation pool overage pricing
- API only for report generation
- Does not support clustered deployment

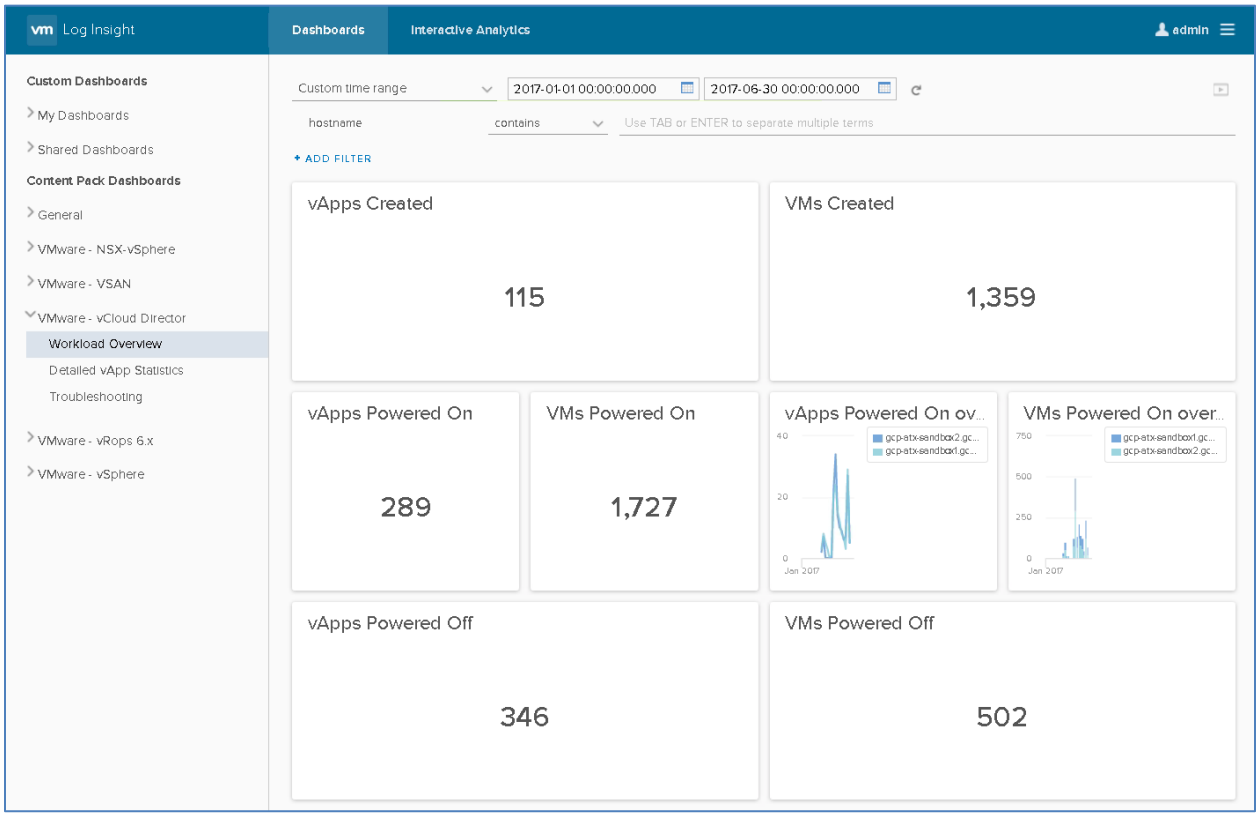
vRealize Business for Cloud supports up to 10 vCenter Server nodes and 20,000 virtual machines.

5.7 vRealize Log Insight

vRealize Log Insight delivers heterogeneous and highly scalable log management with actionable dashboards, analytics and broad third-party extensibility. It provides deep operational visibility and faster troubleshooting across the whole environment.

Content packs of built in or custom dashboards provide additional visibility into unstructured log data through collection of chart, field table, and query list widgets.

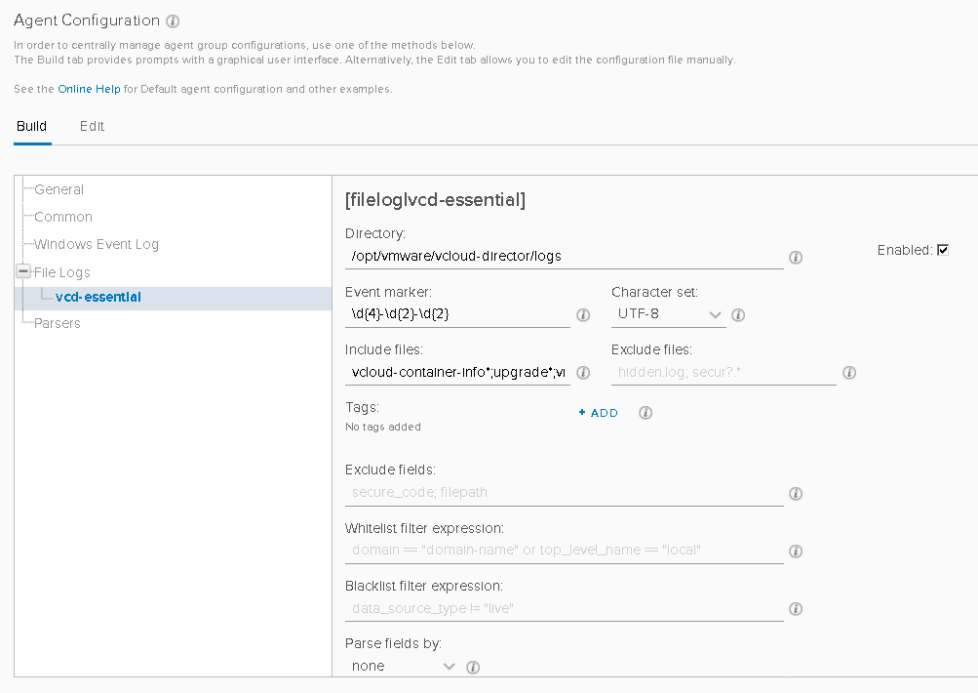
Figure 15. vRealize Log Insight – vCloud Director Dashboard



While it can act as a regular syslog target, it also provides agent based collection of log files, or events for applications that do not provide syslog message logging redirection. Agents are provided for Windows and Linux (RPM, DEB and BIN formats) operating systems.

Collection of logs from vCloud Director cells is much easier with the agent-based method: API request logs can be included, and the selection of log verbosity (info/debug) can be made centrally.

Figure 16. Log Insight Agent Configuration for vCloud Director Cells



5.8 vRealize Orchestrator

vRealize Orchestrator provides service orchestration. It can automate tasks across VMware products, leveraging vCloud API, VIM API, VMware NSX APIs or the vCenter Chargeback API. Using generic plug-ins (SSH, SOAP, HTTP REST, SQL, PowerShell, Active Directory) or third-party specific plug-ins, it can orchestrate other systems as well. vRealize Orchestrator provides a large library of workflows and actions in its base configuration, and its library grows with each newly installed plug-in. Powerful workflows can be built with little or no knowledge of an API.

vRealize Orchestrator can have two distinct roles in vCloud Director environments.

- It can act as an extension that is subscribed to RabbitMQ and consumes vCloud Director messages. In this role, vRealize Orchestrator extends vCloud Director by providing additional services (backup, additional controls, CMDB integration, and so on).
- It acts as an orchestrator for common onboarding or tenant lifecycle tasks. The tasks are triggered by an external portal (VMware vRealize Automation Advanced Service Designer, or through the vRealize Orchestrator REST API). By utilizing plug-ins, the tenant service can be configured end-to-end.

The following are vRealize Orchestrator design considerations:

- As of vRealize Orchestrator v 7.3 can be installed only as a virtual appliance.
- vRealize Orchestrator requires a database. While external MS SQL and Oracle databases are still supported, they are in deprecated mode, scheduled for removal in future releases. Internal PostgreSQL database is preconfigured on the appliance and is production ready.
- High-load production environments and clustered highly available vRealize Orchestrator deployments still require a shared external database.
- In highly available cluster mode, multiple vRealize Orchestrator server nodes with identical server and plug-in configurations work together as a cluster and share one database. Only the active nodes respond to client requests and run workflows. All server nodes communicate with each other by

exchanging heartbeats (by default every 5 seconds, with a 12 heartbeat threshold) through the database. If an active instance fails to send heartbeats, it is considered non-responsive and one of the inactive instances takes control to resume all of the workflows from the point at which they were interrupted. A network load balancer must be used to send the client requests to an active node.

- While it is possible to have more than one active node in a cluster, vRealize Orchestrator Client cannot be used due to concurrency issues. Default number of active nodes is one (recommended maximum of three).
- It is possible to scale out with independent vRealize Orchestrator instances that are each deployed for a specific task (for example, consuming dedicated AMQP queue).
- A multinode plug-in can be used to coordinate workflows between independent vRealize Orchestrator instances.
- LDAP authentication is no longer supported. Instead VMware Identity Manager must be used to integrate with external LDAP Identity Providers.
- The following are recommended external vRealize Orchestrator plug-ins:
  - vCloud Director
  - VMware NSX
- Scalability – A single vRealize Orchestrator instance supports up to 300 concurrent workflow instances in the running state, with 35,000 managed virtual machines in the inventory.

5.9 vRealize Operations Manager

The role of vRealize Operations Manager is to monitor performance and capacity of management cluster and resource groups as well as integrate with vCloud Director to provide useful provider centric metrics. Examples include provider VDC utilization, oversubscription, and organization VDC utilization.

Additionally, service providers can monitor (subset) of tenant workloads for managed services use cases.

The following management packs can be used to extend vRealize Operations Manager monitoring beyond vSphere objects:

- MP for Endpoint Operations (guest level monitoring)
- MP for vCloud Director (correlation of vSphere workloads with vCloud Director objects)
- MP for VMware NSX for vSphere
- MP for VMware vSAN
- MP for vRealize Log Insight (integration for faster troubleshooting)

vRealize Operations Manager provides a way to exclude monitoring of certain resources and thus optimize VMware Cloud Provider Program licensing. The actual VMware Cloud Provider Programming licensing reporting can be then provided through vRealize Operations Manager super metrics and reports.

5.9.1 Service Provider Internal Use Case

This use case provides monitoring of relevant resources for service provider internal usage (performance, availability, capacity, and so on). It excludes tenant virtual machines.

Management cluster monitoring includes the following:

- All VMs in management clusters
- Guest level monitoring for VMs and physical servers with Endpoint Operation MP
- Capacity calculations based on VM demand

- VMware Cloud Provider Program licensing based on monitored VMs
- Physical servers licensing is based on Operating System Instance (OSI)

Resource groups monitoring provides the following:

- All Tenant VMs in resource clusters are excluded from monitoring
- Capacity calculations are based on ESXi host demand
- Licensing is based on physical OSI

5.9.2 Managed Services Use Case

A subset of resource group (tenant) virtual machines can be configured for monitoring as a managed services offering. This can be achieved by leveraging vCenter Server permissions that limit vRealize Operations visibility of the workloads:

- Leverage Provider VDC structure in vCenter Server (clusters / resource pools) and assign access only to those that contain workloads to be monitored
- To avoid incorrect capacity calculations, set **Exclude Virtual Machines from Capacity Calculations** to **true** in the Advanced Settings of vSphere Adapter
- Guest level monitoring for virtual machines
- Capacity calculations is based on ESXi host demand
- Licensing is based on monitored virtual machines

5.10 vCloud Usage Meter

vCloud Usage Meter is a virtual appliance that collects usage data from vCenter Server and vCloud Director to provide data for VMware Cloud Provider Program licensing. It must have access to vCenter Server instances, vCloud Director, NSX Manager instances, and vRealize Operations Manager.

Resource Groups

A resource group is a set of compute, networking, and storage resources dedicated to tenant workloads and managed by a single vCenter Server / NSX Manager pair. vCloud Director manages the resources of all attached resource groups through API communication with vCenter Server and NSX Manager.

Provisioning resources in standardized groupings provides a consistent approach for scaling vCloud environments. A separate vCenter Server instance is recommended to manage resource groups for all except very small environments. If a single vCenter Server is used to manage both management components and resource groups, place all vCloud management components in a separate cluster.

The decision to create a new resource group instead of scaling out an existing resource group is based on:

- Fault zone domain separation
- Multisite requirements
- Scalability limits (both for virtual and physical components)
- vCloud Director object boundaries (provider VDCs or organization VDC networks cannot span vCenter Server domains)

6.1 Resource Group Management Components

The management components of resource groups are installed in the management cluster together with cloud management components or in their own management cluster in a multisite configuration (see Section 3.1.2.1, Distributed Resource Groups). The management components typically consist of vCenter Server, NSX Manager, and supporting systems (vCenter Server database, VMware vSphere Update Manager™, VMware vCenter Single Sign-On™, distributed syslog, AD, and so on).

6.2 Compute Resource

The compute resource is represented by a set of vSphere clusters with VMware vSphere High Availability and VMware vSphere Distributed Resource Scheduler™ (DRS) enabled. For details, refer to design considerations in the *Architecting a VMware vSphere Compute Platform for Service Providers* document.

The following are compute design considerations:

- When calculating the required compute capacity, account for virtualization overhead (VMkernel processes including vSAN and NSX, and virtual machine memory overhead) and resources needed for edge gateway virtual appliances, which are not charged against tenant consumption.
- VMware recommends not mixing CPU generations within the same service offering (provider VDC) because the customer might have a different performance experience due to vCloud Director allocating CPU based on clock rate in GHz.
- Selection of HA admission control policy depends on the required SLA. The only policy that guarantees restart of all workloads from a failed host is “Specify Failover Hosts.”
- The number of hyperthreaded physical cores dictates the maximum number of vCPUs that tenants can allocate to their virtual machines.
- Due to security implications in multi-tenant environments, disable memory transparent page sharing. For more information, see the VMware Knowledge Base article *Additional Transparent Page Sharing Management capabilities in ESXi 5.5, 5.1, and 5.0 patches in Q4, 2014 (2091682)* at http://kb.vmware.com/kb/2097593.
- Oversubscribed physical memory can cause memory ballooning or hard drive swapping, which can affect other tenants. In the Reservation type allocation model, the tenant can oversubscribe their

actual allocated VDC. Therefore, VMware recommends always using Reservation type allocation mode on tenant dedicated hardware.

- As of vCloud Director 8.10 and extended in 8.20, you can influence vSphere placement decisions with VM-VM affinity groups (tenant exposed), VM-host affinity groups (only provider exposed) and VM-host tagging through metadata (tenant exposed). While this opens new use cases (clustered workload availability, OS licensing, low latency workloads), it must be accounted for in provider operational practices.

6.3 Networking

vCloud Director creates and manages virtual networks and network services through NSX Manager – the management component of VMware NSX for vSphere which offers full backward compatibility and a seamless upgrade path for VMware vCloud Networking and Security™ used in older vCloud Director environments.

VXLAN based logical networks, Edge Gateways and distributed firewalls are deployed and managed by NSX Manager, while VLAN and vCloud Isolation (VCDNI) based logical networks are deployed and managed directly by vCloud Director.

vCloud Director 8.20 offers access to Edge Gateway and distributed firewall NSX services to tenants, while the provider can leverage non-exposed NSX services for provider-managed services.

The following table summarizes the transition between vCloud Director releases and vCloud Networking and Security / NSX interaction.

Table 6. vCloud Director Networking Platform Transition

vCloud Director	5.5/5.6	8.0	8.10	8.20
vCloud Networking and Security	●	●		
NSX for vSphere	●	●	●	●
Edge Gateway version	5.5	5.5	6*	6*
VXLAN	●	●	●	●
VCDNI	●	●	●	deprecated

* Edge Gateway version 5.5 is still supported, however newly deployed Edge Gateways are version 6.

Note NSX version 6.3 and later no longer support legacy networking APIs, and therefore, Edge Gateway version 5.5 is no longer supported. The provider must upgrade all Edges to version 6 before upgrading to NSX version 6.3!

6.3.1 Transport Zones

A transport zone defines the scope of a VXLAN logical switch. It consists of one or more vSphere clusters. Transport zones can be created manually. However, vCloud Director automatically creates for each provider VDC one transport zone, which matches the clusters that are added to the provider VDC and associates it with a VXLAN network pool. When the organization VDC is created by the vCloud system administrator or from a VDC template, a network pool must be assigned—all organization VDC and vApp networks will span the transport zone scope.

Note The transport zone created by vCloud Director always uses VXLAN control plane mode multicast. It can, however be changed manually to unicast/hybrid.

6.3.2 NSX Edge Cluster

VMware NSX overlay networks allow the creation of logical networks over an existing IP network fabric. This enables a highly scalable network design using a leaf and spine architecture, where the boundary between Layer 2 and Layer 3 networks is at the rack level (leafs) and all communication between racks is Layer 3 only through a set of spine routers.

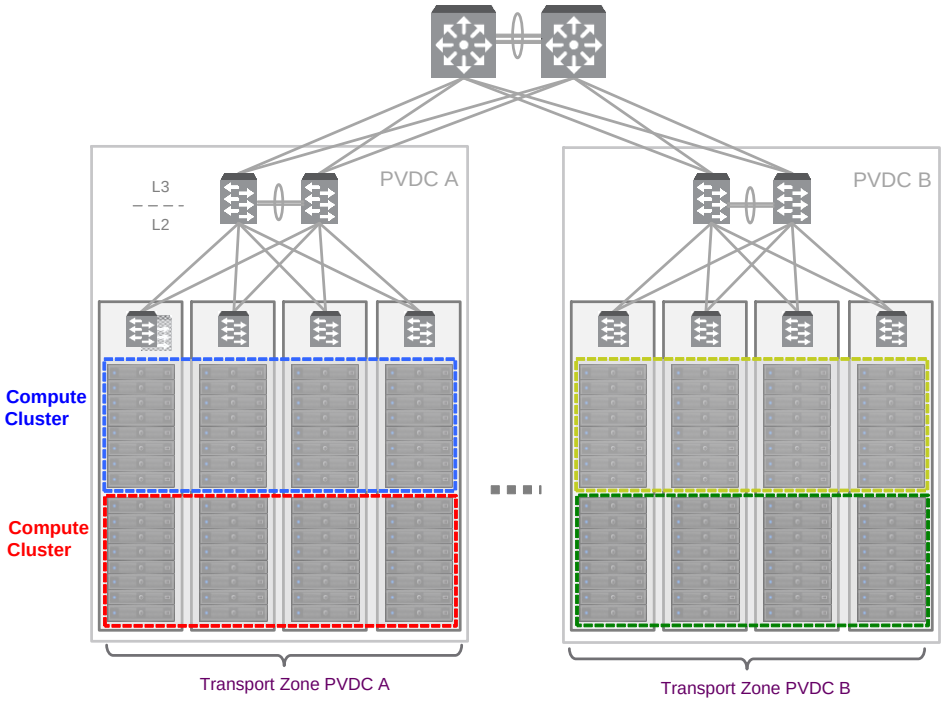
VMware NSX logical networks span across all racks. However, there is a need to connect virtual workloads from the logical networks to the outside physical world (WAN, Internet, co-located physical servers, and so on). These networks are represented by a set of VLANs, and because there is no stretching of Layer 2 across the racks, they cannot be trunked everywhere. They are connected only to one (or two for redundancy) racks, and become the NSX Edge cluster.

The purpose of the NSX Edge cluster is to host virtual routers—edge service gateways that provide the connectivity between the physical world (VLANs) and virtual world (VXLAN logical switches). This does not mean that every NSX Edge gateway needs to be deployed there. If an NSX Edge gateway provides connectivity between two VXLAN logical switches, it can be deployed anywhere because logical switches span all clusters.

6.3.2.1 Design Option 1 – Traditional

In the traditional access/aggregation/core network architecture, the Layer 2 / Layer 3 boundary is at the aggregation switches. This means that all racks connected to the same set of aggregation switches have access to the same VLANs. There is no need for an edge cluster because the edge VM connecting the VLAN with VXLAN based networks can run on any rack. In vCloud Director, if the external networks (VLANs) are trunked to aggregation switches, edge placement is not a concern. The set of racks (clusters) connected to the same aggregation domain usually maps to a vCloud Director provider VDC. The transport zone is then identical to the aggregation domain. The drawback of this design is that provider VDCs cannot span multiple aggregation domains.

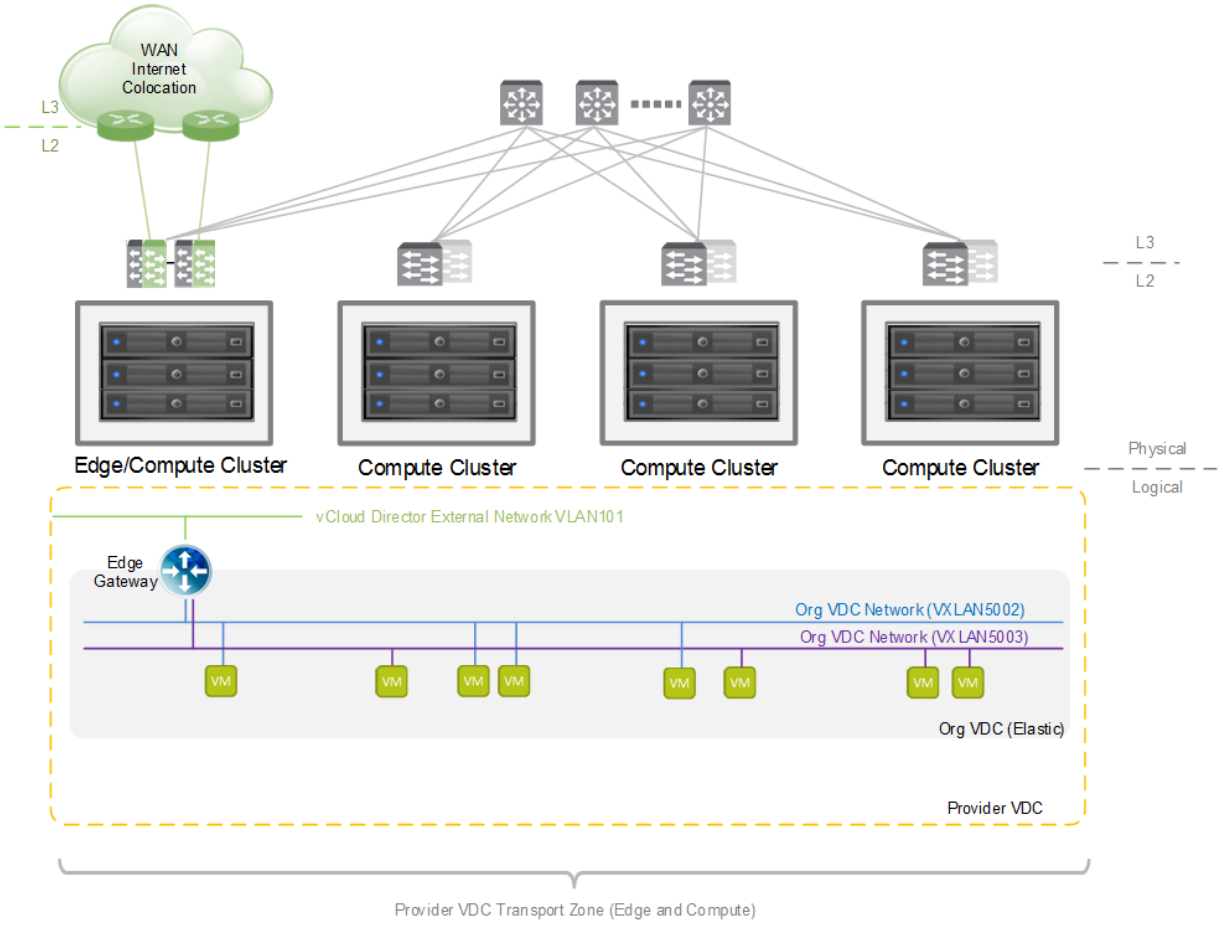
Figure 17. Traditional Access/Aggregation/Core Architecture



6.3.2.2 Design Option 2a – Combined Edge/Compute Cluster

When the leaf and spine network architecture is used, VLANs that back vCloud Director external networks are trunked only to the Edge/Compute cluster. The vCloud Director placement engine deploys edge VMs to a cluster based on VLAN connectivity. vCloud Director automatically places all edge gateways into the Edge/Compute cluster because it is the only cluster where the external connectivity (VLANs) exists. However, vCloud Director will also opportunistically place regular tenant VMs into this cluster (hence its name: Edge/Compute).

Figure 18. Leaf and Spine with Edge/Compute Cluster



This design option has all the scale advantages of leaf and spine architecture. However, the drawback is the possibility of tenant workloads taking up limited space of the Edge/Compute cluster. There are two options to remediate this:

- vCloud Director edge gateways are always deployed by the vCloud system administrator. Prior to edge gateway deployment, the administrator can verify that there is enough capacity in the Edge/Compute cluster. If not, some tenant workloads can be migrated to another cluster. This must be done from within vCloud Director (*Resource Pool* / *Migrate to* option). Live migration is possible only if the Edge/Compute cluster shares the same VXLAN prepared VMware vSphere Distributed Switch™ (VDS) with the other clusters. This setup requires at least four network uplinks on the Edge/Compute cluster hosts (two uplinks for the edge VDS with external VLANs and two uplinks for the VXLAN VDS that spans all clusters).
- Artificially limit the size of the Edge/Compute cluster so the placement engine does not choose it for regular tenant workloads. This can be achieved by leveraging the resource pool, which is created manually by the system administrator in the Edge/Compute cluster and attached to the provider VDC instead of the whole cluster. An artificial limit is set by the system administrator and is increased only when a new edge gateway needs to be deployed.

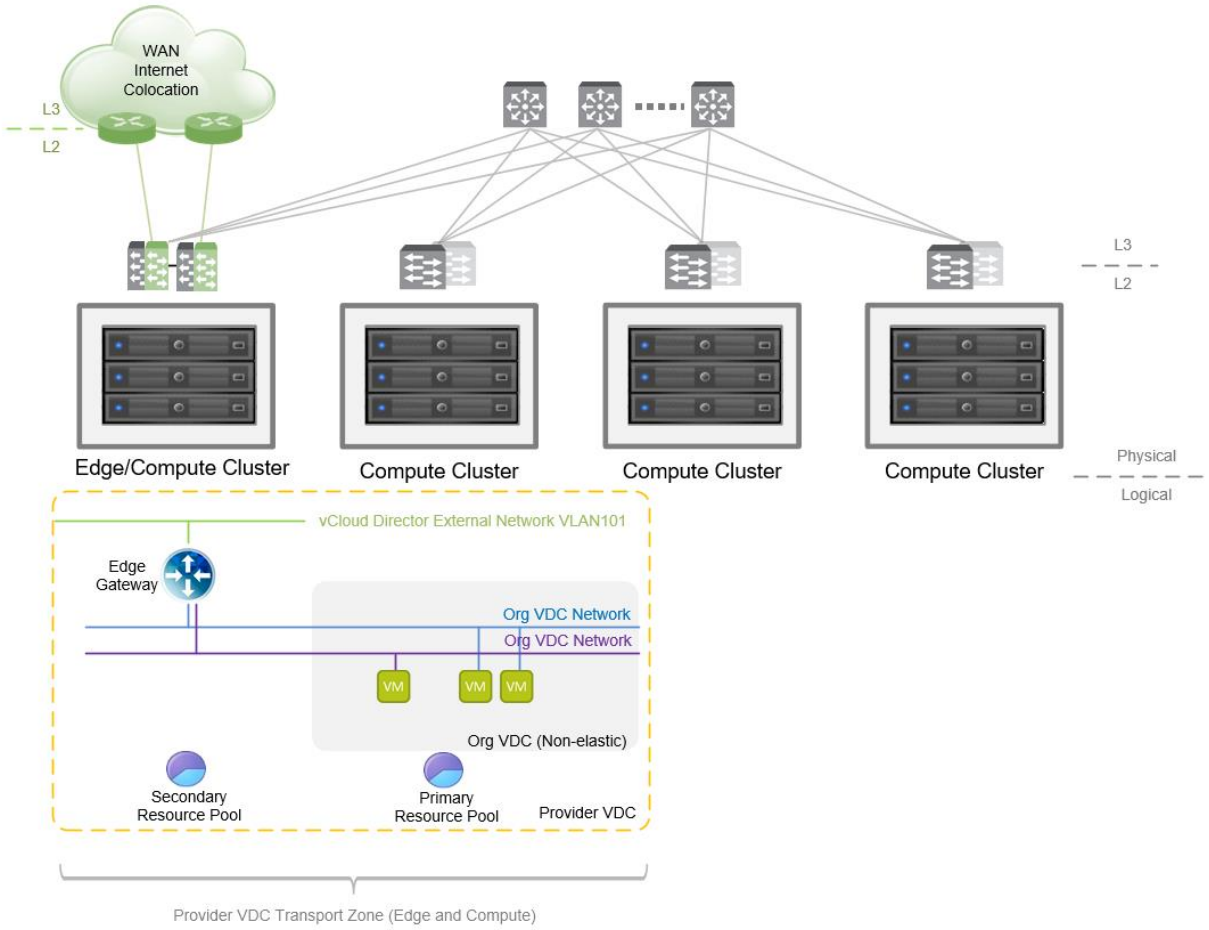
Both options unfortunately involve significant operational overhead therefore are not recommended..

6.3.2.3 Design Option 2b – Combined Edge/Compute Cluster with Non-Elastic VDC

Elastic Org VDC types (such as pay-as-you-go or allocation) can span multiple clusters. Consider the impact of a non-elastic VDC, such as a reservation pool.

In a non-elastic Org VDC, all tenant workloads are deployed into the primary provider VDC resource pool. Edge VMs can be deployed into secondary resource pools. If the Edge/Compute cluster is added as a secondary resource pool into a provider VDC, this design option can be used.

Figure 19. Leaf and Spine with Edge/Compute Cluster and Non-Elastic VDC



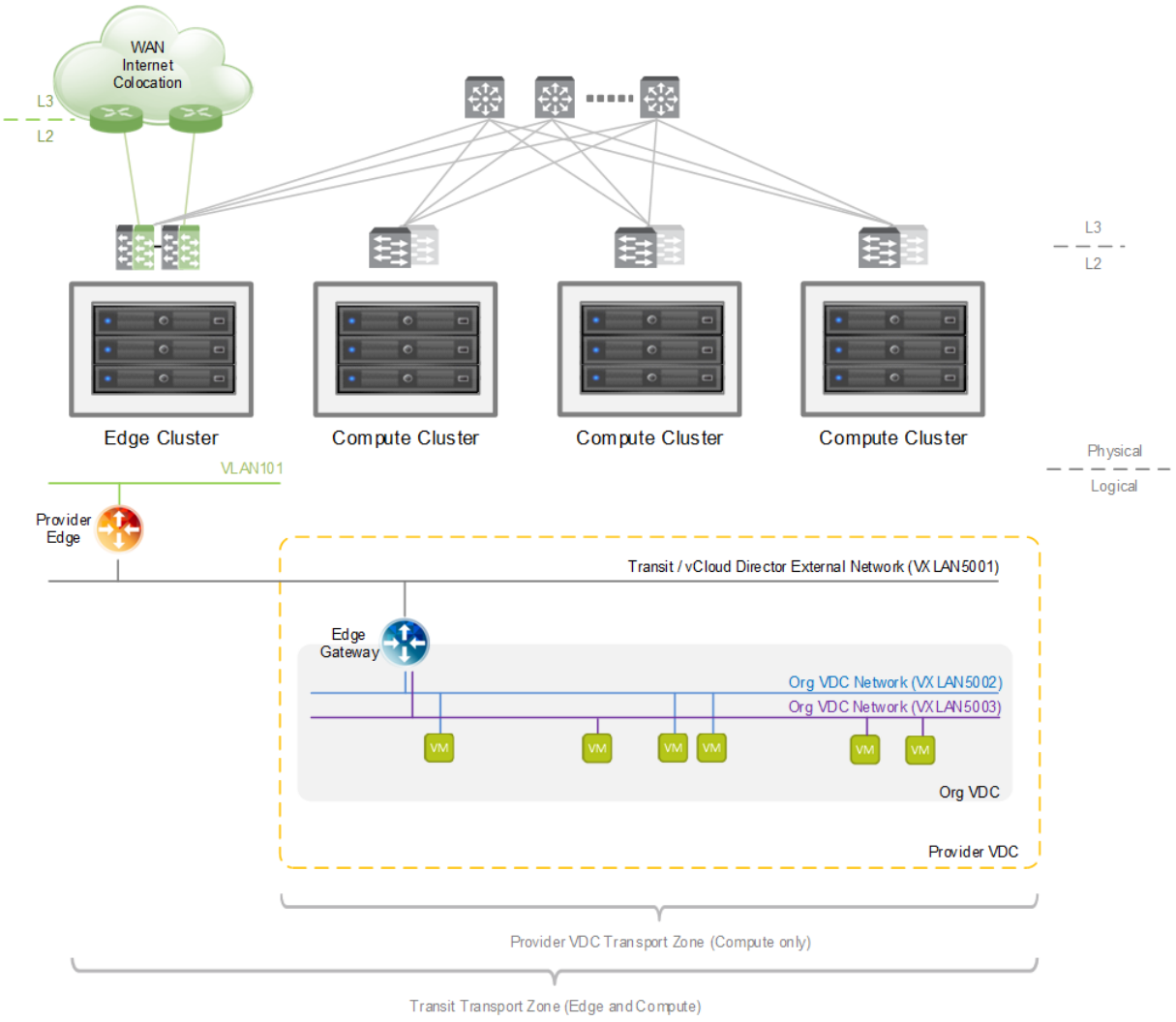
6.3.2.4 Design Option 3a – Dedicated Edge

This design option has a dedicated Edge cluster that is not managed by vCloud Director and introduces a new edge gateway type—provider edge. Provider edges are manually deployed by the service provider outside of vCloud Director into the Edge cluster. Provider edge external uplinks are connected to external VLAN-based networks, while internal interfaces are connected to a transit VXLAN logical switch spanning all Compute and Edge clusters (using manually created transport zone with all clusters). The transit networks are then consumed by vCloud Director as external networks.

The provider edges can provide all VMware NSX functionality (dynamic routing protocols on external uplinks, Layer 2 bridging, Layer 2 VPN, and so on). They can scale as additional vCloud Director external networks are added (the current maximum in vCloud Director 8.20 is 1999 external networks). The edges

deployed by vCloud Director then go into compute clusters because all their interfaces connect to VXLAN logical switches spanned everywhere in the provider VDC.

Figure 20. Leaf and Spine with Dedicated Edge Cluster



6.3.2.5 Design Option 3b – Dedicated Edge Cluster with ECMP Edges

In the previous design option, there was one provider edge in the edge cluster for each transit vCloud Director external network to which Org VDC edge gateways are connected.

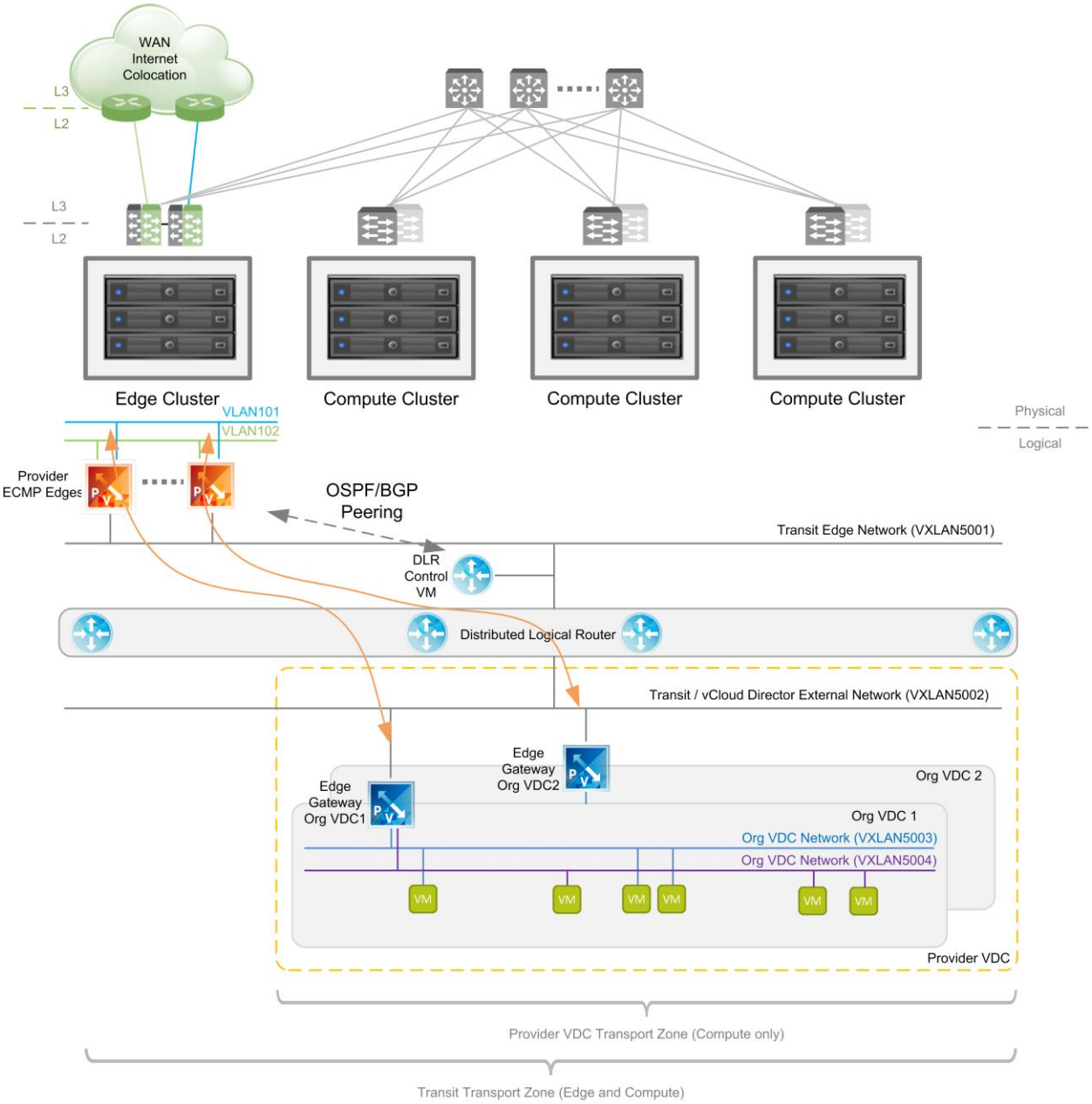
To provide access to a shared service (for example, the Internet) where multiple Org VDC edge gateways of different tenants are connected to the same external network, all external network traffic must go through a single provider edge, which can become a bottleneck.

VMware NSX Edge gateways can be deployed in an Equal Cost Multi-Path (ECMP) configuration where the bandwidth of up to 8 edges (8x10 GB = 80 GB throughput) can be aggregated. High availability of ECMP edges is achieved with a dynamic routing protocol (BGP or OSPF) configured with aggressive timing for short failover times (3 seconds) which will quickly remove failed paths from the routing tables.

The problem is that to take advantage of multiple paths, the tenant Edge Gateways must set up peering with the provider Edge to exchange routing information and availability of the paths. This is, however, not

manageable in a shared environment where each newly deployed tenant Edge Gateway must have peering with the Provider Edge set up. The following design works around this limitation by deploying a distributed logical router (DLR) between provider and organization VDC edges. The DLR then provides a single distributed, highly available default gateway for all tenant Edge Gateways.

Figure 21. Leaf and Spine with Dedicated Edge Cluster and ECMP Edges



The previous figure shows two provider ECMP edges (can scale up to 8) with two physical VLAN connections each. These connections are to the upstream physical router and one internal interface to the transit edge logical switch. The DLR then connects the transit edge logical switch with the transit vCloud Director external network to which all tenant Org VDC edge gateways are connected. The DLR has ECMP routing enabled as well as OSPF or BGP dynamic routing peering with the provider edges. The DLR provides two (or more) equal paths to the upstream provider edges and chooses one based on a hashing algorithm of source and destination IP addresses of the routed packet.

The two shown Org VDC edge gateways (which can belong to two different tenants) then take advantage of all the bandwidth provided by the Edge cluster (indicated with the orange arrows).

The figure also depicts the DLR control VM. This is the protocol endpoint that peers with Provider Edges and learns and announces routes. Routes are then distributed to the VMware ESXi™ host VMkernel routing process by the VMware NSX Controller™ cluster (not shown in the figure). The failure of the DLR control VM has impact on routing information learned through OSPF/BGP protocols, even if the DLR is highly available in active standby configuration. This is due to the protocol aggressive timers (DLR control VM failover takes more than 3 seconds). Therefore, a static route is created on all ECMP provider edges for the transit vCloud Director external network subnet. That is enough for North-South routing, because Org VDC subnets are always behind the tenant Org VDC edge gateway that provides Network Address Translation (NAT). North-South routing is static because the Org VDC edge gateways are configured with a default gateway defined in the external network properties.

The other consideration is placement of a DLR control VM. If the VM fails together with one of ECMP Provider Edges, ESXi host VMkernel routes are not updated until DLR control VM functionality fails over to the passive instance. In the meantime, the route to the non-functioning provider edge is blackhole traffic. If there are enough hosts in the Edge cluster, deploy DLR control VMs with anti-affinity DRS rule to all ECMP edges. Most likely, there are not enough hosts, so deploy DLR control VMs to one of the compute clusters. The VMs are very small (512 MB, 1 vCPU). The cluster capacity impact is negligible.

6.3.3 Summary of Edge Cluster Deployment Options

Table 7. Summary of Edge Cluster Deployment Options

Design Option	Pros	Cons
No Edge Cluster	Simplicity Ideal for traditional networking architectures (three tier)	VLANs trunked to all provider VDC hosts limit scale and extend network failure domain
Combined Edge/Compute Cluster	Spine-leaf network architecture support Scale Deployement simplicity	Administration overhead around capacity management of Edge/Compute cluster Four network uplinks (for two vDS switches) needed in order to be able live migrate workloads away from Edge/Compute cluster
Dedicated Edge Cluster	Spine-leaf network architecture support Support for provider managed Edge Gateway	Provider edge could become a bottleneck in large deployments
Dedicated Edge Cluster with ECMP Edges	Spine-leaf network architecture support Highly scalable Support for additional provider managed NSX features	Deployment complexity

6.3.4 NSX Controller Cluster

The NSX Controller cluster is the control plane component that is responsible for managing the switching and routing modules in the ESXi VMkernel. The following table shows which VMware NSX features require the NSX Controller cluster.

Table 8. NSX Controller Cluster Feature Requirement

NSX Feature	NSX Controller Cluster Requirement
VXLAN Transport Control Plane	
Multicast	✗
Hybrid	✓
Unicast	✓
Distributed firewall	✗
NSX Edge gateways	✗
Distributed logical router	✓
VXLAN – VLAN bridging	✓
ARP suppression	✓

For migration from vCloud Network and Security to VMware NSX, the NSX Controller cluster must be deployed before any of the advanced NSX features that require it are used.

The following are NSX Controller cluster design considerations:

- NSX Controller cluster consists of NSX Controller nodes, which are deployed by NSX Manager to the vSphere environment which the NSX Manager is paired with. Therefore, the NSX Controller is running in the resource group vSphere clusters.
- NSX Controller cluster consists of three nodes, which are virtual machines deployed by NSX Manager. An NSX Controller cluster with one VM can be used only for training and demo purposes. An even number of controllers is not supported because there must always be a quorum.
- For high availability purposes, place each NSX Controller node on a different host. This can be achieved with a manually created anti-affinity DRS rule.
- The NSX Controller node VM must be connected to a standard or distributed port group. It cannot be connected to a VXLAN-based port group (logical switch).
- NSX Controller instances must have network connectivity to NSX Manager and ESXi management vmknics. They do not need to be deployed in the same Layer 2 subnet.
- NSX 6.3.2 and later offers Controller Disconnected Operation (CDO) mode which provides resiliency against control plane (Controller Cluster) failure[2]. CDO mode is currently supported if there is only

[2] https://blogs.vmware.com/networkvirtualization/2017/03/nsx-v-6-3-control-plane-resiliency-cdo-mode.html/

one transport zone of vSphere Distributed Switch, which means only one Provider VDC VXLAN network pool.

6.3.4.1 Design Option I – Edge Cluster

The recommended placement option is to deploy NSX Controller nodes to the Edge cluster (see Section 6.3.2.4, Design Option 3a – Dedicated Edge.) This separates the management component from tenant workloads running in compute clusters. The Edge cluster also has more VLAN connectivity options than Compute clusters. Due to the three-node NSX cluster requirement, VMware recommends having at least four ESXi nodes in the Edge cluster.

6.3.4.2 Design Option 2 – Compute Clusters

When a dedicated Edge cluster is not used, the NSX Controller nodes can be deployed into the compute clusters. The disadvantage is that this setup combines tenant workloads with the management components. For logical separation, use resource pools with specific limits and reservations to guarantee NSX Controller performance and allow vCloud Director compute capacity metering. Provider VDC is then mapped to the resource pool and not to the vSphere Cluster (see Section 7.1, Provider Virtual Data Centers).

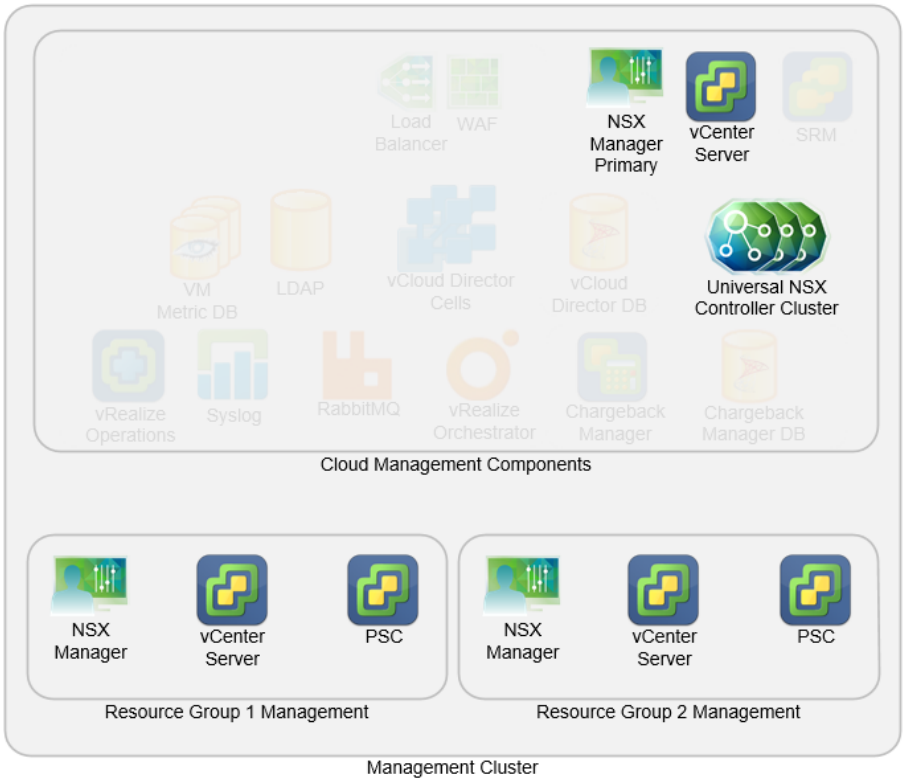
Network connectivity must be VLAN-based with routing to NSX Manager and ESXi management vmknics. The ESXi management network is recommended.

6.3.4.3 Design Option 3 – Universal NSX Controller Cluster

VMware NSX version 6.2 and later supports combining multiple vCenter Server / VMware NSX domains under one management domain. While vCenter Server – NSX Manager coupling still exists, one NSX Manager is acting as primary. Only one universal NSX Controller cluster is deployed by the primary NSX Manager. Secondary NSX Manager instances do not deploy the NSX Controller cluster.

The management cluster NSX Manager acts as primary, and all resource group NSX Manager instances are secondary.

Figure 22. Universal NSX Controller Cluster



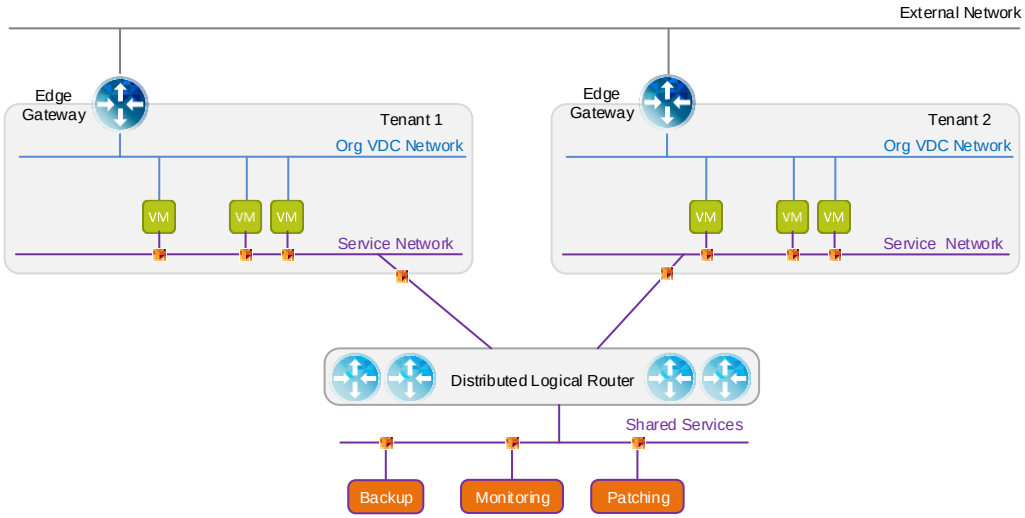
6.3.5 NSX Distributed Firewall and Logical Routing

The usage of NSX features directly by the provider outside of vCloud Director can be shown in an example combining Distributed Firewall (DFW) for segmentation of L2 networks with Distributed Logical Routing (DLR).

A use case example is where shared services, such as monitoring, patching, or backup, are available on the service network for tenant workloads. The tenant connects workloads with a secondary network interface to a dedicated service network with routable access to shared services network. Because there is no need for NAT, this approach works with any monitoring or backup solutions.

The distributed logical router provides scalable routing, while distributed firewall combined with VMware NSX SpoofGuard provides the necessary multitenant security enforcement at the tenant VM vNIC level.

Figure 23. Shared Services with DFW and DLR



Note VMware NSX DFW automatically excludes all VMware NSX created virtual machines. It is not possible to create DFW rules that apply to edge gateways.

6.3.6 Other Network Services

While most of Edge Gateway services are available to tenants, the service provider can still leverage additional VMware NSX capabilities outside of vCloud Director:

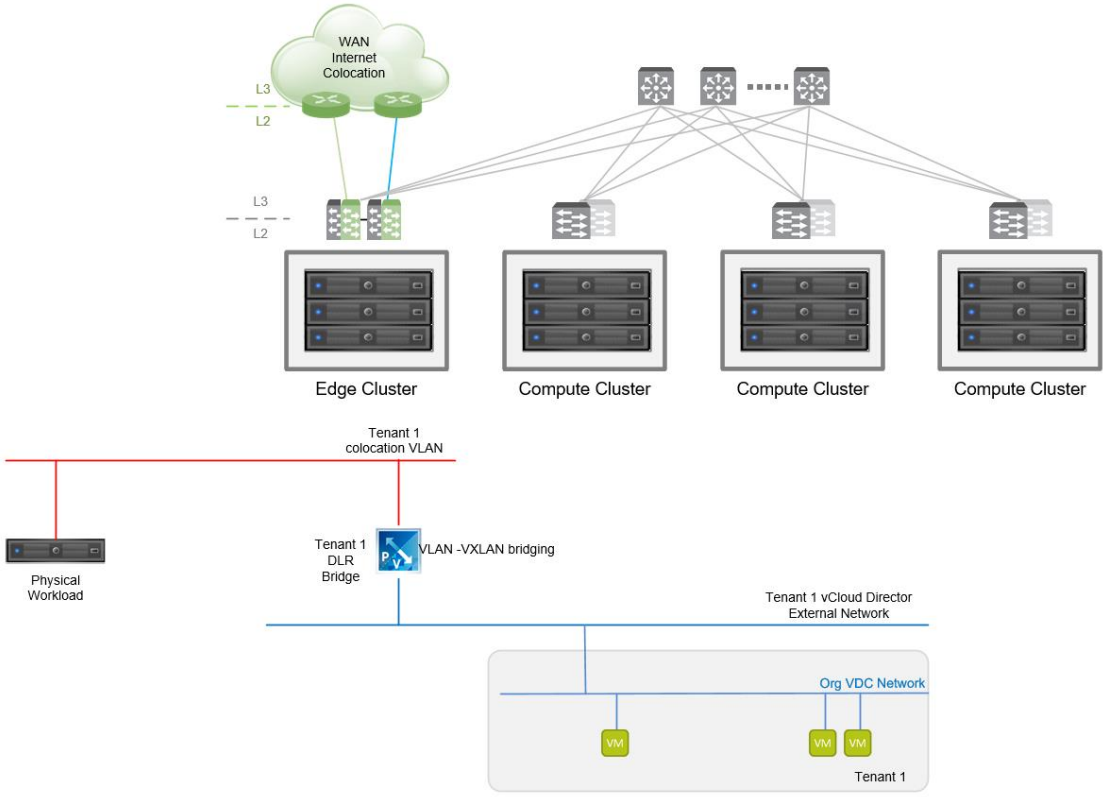
- Layer 2 bridging for connection of tenant Org VDC workloads with co-located physical server in the same broadcast domain.

  The service provider deploys the NSX DLR bridging instance in the edge cluster and configures a VLAN – VXLAN bridging for the physical server VLAN and Org VDC network VXLAN networks. The bridge is directly connected through a dedicated vCloud Director external network to the customer Org VDC.

- IPv6 routing

  The service provider can deploy an NSX Edge gateway for each tenant in the Edge cluster and configure advanced services manually. The tenant is directly connected to its NSX Edge through a dedicated vCloud Director external network.

Figure 24. Provider Managed NSX Services



6.4 Storage

In a vCloud Director environment, users can self-provision vApps to virtual data centers that contain a pool of storage capacity. The physical storage details are abstracted from the user. Precise control of where vApps reside in the physical infrastructure is removed due to provisioning responsibility shifting to the end user. Datastore sizing is a balancing act between availability, recoverability, performance, and manageability requirements. When thin provisioning, fast provisioning, or snapshots are enabled, existing VM storage requirements can grow and create out-of-space datastore conditions. Leverage vCloud Director datastore yellow and red thresholds, vSphere alarms, and other storage monitoring tools.

Storage support as of vCloud Director 8.20:

- VMFS5 supported. Version VMFS6 not supported
- NFS v3 and v4.1 supported
- vSAN supported
- VMware vSphere Virtual Volumes™ not supported

6.4.1 Storage Tiering

To handle the elasticity and scalability required for a vCloud implementation, a modular tiered storage approach is recommended. This involves designing for future growth while optimizing for performance. Storage differentiation for each service tier is determined, and then pools of tiered storage are presented to vCloud Director.

The following are storage tiering design considerations:

- VM storage policies must be enabled on every resource group vSphere cluster.
- User-defined storage tags (for example, bronze, silver, gold) are created and assigned to vSphere datastores. Each vSphere VM storage policy is mapped to corresponding tags with a set of rules. Alternatively, capability-based storage policies can be created for storage resources that support them. Upon synchronization, the storage policies can be assigned to provider VDCs in vCloud Director. In general, do not the universal storage profile * (any). (If it is used, disable the local datastore.)
- vCloud Director references datastores by the vSphere storage policy names. Do not rename them afterwards in vSphere.
- A single vCloud Director VM can span multiple datastores. For example, VM disk 1 can be on datastore A and VM disk 2 on datastore B, where datastores A and B belong to the same or a different storage policy. When fast provisioning is enabled at the Org VDC level, the VM cannot span multiple datastores.
- A single disk can be attached to only one VM at a time. For shared storage between VMs, in-guest IP storage can be used (NFS, iSCSI).
- Independent disks (available only through the vCloud Director API) can be easily moved between VMs.
- Storage policies can be used for placement of VMs to a subset of ESXi hosts due to licensing restrictions (Windows as opposed to Linux hosts). Storage policy of a particular type (for example Linux) is assigned to a datastore that is attached only to hosts licensed for Linux workloads.

6.4.2 Datastore Clusters

A datastore cluster is a collection of datastores with shared resources and a shared management interface. Datastore clusters are to datastores what vSphere clusters are to hosts. Datastore cluster is required to use VMware vSphere Storage DRS™ to load balance storage resources.

The following are datastore cluster design considerations:

- Ongoing space balancing on LUNs that are thinly provisioned by the storage array creates unusable white space. VMware vSphere Storage vMotion operation relocates VMs disk (VMDK) files at the vSphere layer. However, the storage array is not aware that the source LUN freed space can be reused. The deleted blocks must be reclaimed manually by running the `vmkfstools` command from the ESXi CLI prompt.
- A datastore cluster with Storage DRS can simplify datastore management and migrations with Storage DRS maintenance mode. It is, however, no longer possible to manage datastores at individual level (enable/disable, disk thresholds, or VMware vSphere Storage APIs – Array Integration for fast provisioning).
- All datastores in a datastore cluster must belong to the same storage policy.
- Limited NFS 4.1 support.

6.4.3 Cloning and Copy Operations

When designing a datastore (LUN, NFS export, vSAN), take into account layout and storage network considerations (especially bandwidth) with regard to expected cloning and copy operations:

- Importing or exporting templates to and from vCloud Director, or movement of VMs or ISO images between vCenter Server instances, always goes through vCloud Director. Cells transfer shared volume over the management network to the ESXi host's management vmknic.

- Clone and copy operations between vSphere clusters that do not have common shared storage use the ESXi host management vmknic, or in the case of vSphere 6.x, vmknic with provisioning service enabled.
- ISO media is stored as a file on the catalog datastore. The media cannot be mounted to a VM that is running on a host that does not have access to the datastore.
- vSAN datastore is accessible only by the vSphere cluster it belongs to.
- When using port binding, iSCSI traffic cannot be routed.
- vSphere 5.5 does not support IP storage over IPv6 networks
- When an elastic Org VDC is used, the provider can migrate running VMs from one cluster / resource pool to another. This operation, however, does not support enhanced vMotion compatibility and shared storage is required between clusters.

6.4.4 vSAN

vSAN is supported in vCloud Director and can be used as a storage tier for tenants through manually-created storage policies that relate to vSAN capabilities.

The following are Virtual SAN design considerations:

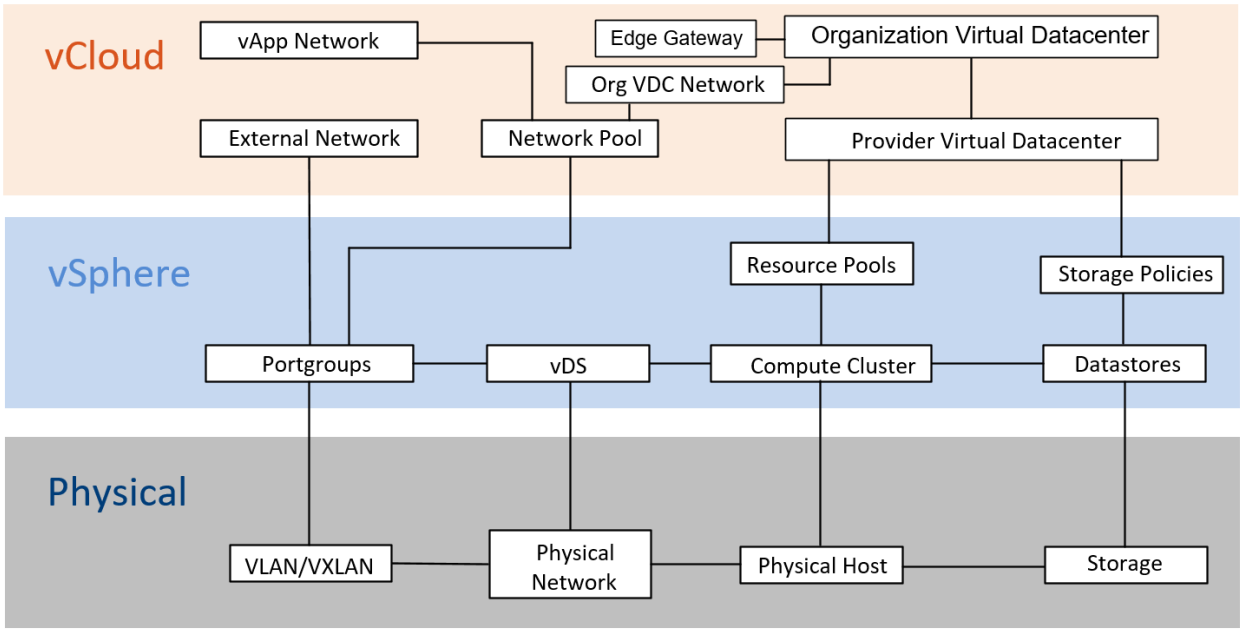
- vSAN datastore is an object that is tightly related to a single vSphere cluster. It cannot provide storage to different vSphere clusters.
- VMware does not recommend using vSAN storage for catalogs. All catalog media images (ISO files) are uploaded by vCloud Director as file objects into a directory structure under a single folder. If the same vSAN datastore is used as a storage policy for different catalogs, they all share one VM Home Namespace object with a maximum size of 255 GB.
- Third-party virtual storage appliances that consume vSAN storage and provide NFS file services can be used to provide catalog across clusters with scale above 256 GB.

vCloud Director Design

vCloud Director adds a layer of resource abstraction to facilitate multi-tenancy and provide interoperability between clouds built to the vCloud API standard.

- Physical compute, storage, and network resources are passed to the vSphere layer where resource pools, virtual switches, and storage policies are created.
- Resource pools and datastores are then passed up to vCloud Director and attached to provider virtual data centers.
- Pure virtual compute and storage resources are exposed to users through virtual data center constructs. Users consume pure virtual resources from virtual data centers through various allocation models.

Figure 25. Physical, Virtual, and Cloud Abstraction Relationships



For multi-tenancy, the following key constructs are introduced by vCloud Director.

Table 9. Virtual Data Center Definitions

Term	Definition
Organization	The unit of multi-tenancy representing a single logical security boundary. An organization contains users and virtual data centers.
Provider virtual data center	A grouping of compute and storage resources from a single vCenter Server. A provider virtual data center can be composed of one or more resource pools. It combines resource pools with one or more storage policies and can share resources to multiple organizations.

Term	Definition
Organization virtual data center	A sub-grouping of compute, memory, storage resources, networks and network routers, allocated from a provider virtual data center. A virtual data center is a deployment environment where vApps can be instantiated, deployed, and powered on. Virtual data centers cannot span multiple organizations.

7.1 Provider Virtual Data Centers

The virtual data center (VDC) is the standard container for a pool of compute, storage and network resources. There are two types of virtual data centers—provider and organization. Provider virtual data centers are assembled from resource pools and datastores represented by storage policies managed by a single resource group vCenter Server.

A provider VDC can span multiple resource pools (or dedicated clusters) and form elastic organization VDCs. However, this is supported only for pay-as-you-go Org VDCs and for Allocation Pool VDCs. Reservation pool VDCs are never elastic and are constrained by the size of the primary resource pool (cluster) of the provider VDC.

Note The vCloud administrator must enable elastic allocation pool VDCs. This is a system-wide setting.

The following are provider virtual data center design considerations:

- Mixing organization VDC allocation types within the provider VDC is possible, but complicates capacity management, particularly when different allocation types have different performance SLAs.
- From a manageability perspective, VMware recommends backing the provider VDC with vSphere clusters. It is also possible to divide a cluster into multiple resource pools and assign these to different provider VDCs. This might be useful when non-vCloud Director managed VMs are deployed into the same cluster. In this case, manually set resource pool limits so vCloud Director understands the amount of resources each provider VDC can consume from the subdivided vSphere cluster.
- Each availability zone must have its own provider VDCs. This allows creation of site-specific Org VDCs.
- Create dedicated Org VDCs on top of a customer-dedicated cluster in a customer-dedicated provider VDC.
- Elastic provider VDCs allow seamless expansion (and possible migrations) for allocation and pay-as-you-go Org VDCs.
- A VXLAN network pool from which organization VDC and vApp networks are provisioned is created automatically with the creation of each provider VDC. Each VXLAN network pool is represented by VXLAN network scope spanning clusters that belong to a given provider VDC. An organization VDC can use any network pool (but only one).
- Host preparation does not push the vCloud agent to the ESXi host provided that no vCloud Director Network Isolation (VCDNI)-backed network pools are defined. However, it sets specific custom attribute to control host – VM compatibility.

7.1.1 Placement Engine

When a virtual machine is created, the placement engine puts it in a provider virtual data center resource pool that best fits the requirements of the virtual machine. A sub-resource pool is created for this organization virtual data center under the provider virtual data center resource pool, and the virtual machine is placed under that sub-resource pool.

When the virtual machine powers on, the placement engine checks the provider virtual data center resource pool to confirm that it still can power on the virtual machine. If not, the placement engine moves the virtual machine to a provider VDC resource pool with sufficient resources to run the virtual machine. A sub-resource pool for the organization virtual datacenter is created if one does not already exist. If the movement requires transfer of virtual machine files, NFC protocol transfer over management network or provisioning network is used.

7.1.2 Tiers

A single provider virtual data center can provide various tiers of storage represented by storage policies, but only one tier of compute. The provider VDC can span multiple vSphere clusters (resource pools), however it is bound to a single vCenter Server.

Multiple provider VDCs are created for the following reasons:

- The cloud requires more compute capacity than a single vCenter Server can provide.
- Tiered compute is required. Each provider VDC maps to vSphere clusters with different characteristics (performance or SLA).
- Requirement for workloads to run on a physically separate infrastructure (availability zones).

7.2 Organizations

Organizations are the unit of multi-tenancy within vCloud Director and represent a single logical security boundary. A vCloud Director organization maps to an end customer. Organizations can use local vCloud Director accounts or direct LDAP integration (with optional SSPI), or can integrate with SAML2 or OAuth with a compatible identity provider (VMware Identity Manager, Microsoft Active Directory Federation Services, OpenAM, and so on).

An administrative organization is typically created to provide global catalogs and possibly other shared services (licensing servers, patching repositories, and so on).

vCloud Director system administrator access can be managed by using integrated LDAP authentication or by federation with vCenter Single Sign-On service, which allows seamless management integration between vCloud Director and vSphere objects. If vCenter Single Sign-On is federated, when trying to log in, the vCloud Director system administrator is redirected to the VMware vSphere Web Client for the authentication. Therefore, proper network connectivity is required for the administrators not only to vCloud Director but also to a particular vSphere Web Client. vCenter Single Sign-On can also provide two-factor authentication[3] for provider access.

7.2.1 User Management

The following are user management design considerations:

- Local vCloud Director accounts have limited password policy enforcement options.

[3] https://blogs.vmware.com/vsphere/2016/04/two-factor-authentication-for-vsphere-rsa-securid.html

- The user or organization administrator can change the user's password only if it is a local vCloud Director account.
- Organization LDAP settings must be configured by the system administrator. vCloud Director cells must have network access to LDAP servers.
- Active Directory SSPI integration allows single sign-on for tenants who are already authenticated in the Active Directory domain.
- SAML 2.0 identity provider (IdP) can be configured by the organization administrator. The tenant can use its own IdP (Active Directory) without requiring network connectivity between vCloud Director cells and the IdP servers.
- OAuth 2.0 authentication allows user to authenticate with single token provided by an external identity provider. The service provider can revoke the right to configure OAuth from the organization administrator and manage it on their behalf for third-party portal integration or federation of multiple vCloud Director instances.
- User management rights can be revoked from the organization administrator, which might be useful if the provider manages accounts in a centralized identity provider and does not want to allow tenants create local accounts.

7.2.2 OAuth Authentication

OAuth simplifies user access management especially in federated multi vCloud Director environments.

Typical workflow:

1. The system administrator enables vCloud Director organization for OAuth authentication
2. User access to the organization and roles are managed in central identity provider (for example LDAP).
3. A user who wants to access a given organization must first be authenticated by the central identity provider. The identity provider will issue a bearer OAuth2 token which gives access to the specific resource to anyone who has the token.
4. The OAuth token consists of three base64 encoded text string sections delimited by a dot ('.'). The first part is JWS (JSON Web Signature) header, the second part is *claims set*, and the third part is the signature.
5. The *claims set* section must contain *authz* field which provides information to which organizations the user has access and under which role.
6. The user makes vCloud API call to vCloud Director passing the OAuth token in the Authorization header of the HTTP API request together with the vCloud Director organization name.

   ```
   Authorization = Bearer <Base64 encoded OAuth Token>;org=<organization name>
   ```

7. vCloud Director extracts the token and performs expiration and signature validation and retrieves the role information to set the users security context. A vCloud authorization token (`x-vcloud-authorization`) is issued, which can be used for subsequent API requests or for browser portal access if stored as a `vcloud_session_id` cookie.
8. If the user does not exist in the vCloud Director organization, it is automatically imported.
9. The requested API call is performed in the proper user security context.

Note The API call does not need to be login session request (`POST /api/sessions`). It can be any API request. For example, `GET /api/session` would return a session object containing the user name and URL link to the user's organization object.

Table 10. OAuth Token Claims

Claim	Description	Notes
jti	OAuth token id	A new session is created if no session exists already associated with jti
sub	User ID of the user being logged in	Universal identifier for the subject of the token
email	User's email	
uname	User name/UPN that the user logs in as	Unique, 1:1 with User ID
cid	tenant/company/customer id that the user belongs to	Not used
tvr	OAuth token version	vCloud Director supports only 2.0
iat	Token issuance time, in seconds	Token must be presented at or after this time
exp	Token expiration time in seconds.	Token must be presented before this time
iss	Token issuer ID	Used to verify that the token is issued by the configured issuer
authz	Represents the set of roles for each specific service instance	
instances	Service instances	Organization IDs
roles	User role	vCloud Director user role

The *authz* section must have the following format:

```
"authz" : {
            "com_vmware_vchs_compute" : {
                  "instances": {
                        "34691574-7ccd-4fc1-b940-0bd2388bf3a5": {
                              "roles" : [
                                    "Organization Administrator"
                              ]
                        },
                        "48df38a4-aec8-4a34-b25a-b8f372bd8c33": {
                              "roles": [
                                    "Organization Administrator"
                              ]
                        }
                  }
```

```
            }
        }
```

Where `34691574-7ccd-4fc1-b940-0bd2388bf3a5` and `48df38a4-aec8-4a34-b25a-b8f372bd8c33` represent Organization IDs where the user has Organization Administrator role access.

Note The `com_vmware_vchs_compute` string is mandatory.

The following are OAuth authentication design considerations:

- While a vCloud Director organization can use multiple identity providers at the same time, an organization user can be associated with only single identity provider. For example, it is not possible for the same user to log in through OAuth and integrated LDAP authentication.
- The service provider can use OAuth authentication for federation of multiple vCloud Director instances with the central identity provider, while the tenant can still use SAML authentication to federate tenant users with their company Active Directory (with Active Directory Federation Services). The SAML users will not exist in the provider's central identity directory.
- External tools that use vCloud API (such as vRealize Automation) and that rely on basic authentication do not work with OAuth authentication. To enable OAuth, the service provider must implement the following process:
    a. Intercept API authentication calls (`POST /api/sessions` and `/api/login`).
    b. Get the Authorization header. If it is not basic authentication, pass it to the vCloud API endpoint.
    c. If it is basic authentication parse and Base64, decode the header to get <username>@<org>:<password> values.
    d. Use the credential values to authenticate against provider's central identity provider.
    e. Retrieve the OAuth token and replace the Authorization header of the original request with the Base64 encoded OAuth header (`Bearer` <OAuth-token>;`org=`<org>).
    f. Forward the request to the vCloud API endpoint.

7.2.3 Granular Role-Based Access Control

vCloud Director 8.20 introduces the possibility to create granular roles at tenant and system level. This is important for service providers who want to differentiate which tenants have access to specific features (for example advanced networking services). This also allows tenants to create their own roles that correspond to their team structure (for example, network administrator). And last, system the administrator can create additional roles in system context with access to a subset of features.

A role is a set of rights which can be assigned to a user or a group. A tenant rights example is to configure IPSEC VPN. A system admin rights example is to enable/disable the host.

Prior to vCloud Director 8.20, the following limitations existed:

- Only global roles could be created by a system administrator in addition to a handful of predefined roles (vApp Author, Organization Administrator, and so on).
- Every organization would have access to the global and predefined roles.
- The organization administrator could assign the roles to organization users.
- The service provider could not differentiate access to features among different tenants.
- There was only one system administrator role with access to everything.

With vCloud Director 8.20, the following capabilities exist:

- Roles are no longer global, but instead are organization specific.
- Former global and predefined roles become *role templates*.
- The service provider can create new role templates.
- Role templates are used to instantiate organization specific roles.
- The service provider can selectively grant rights to specific organizations.
- Organization administrators can create their own organization specific roles from a subset of granted rights.
- New roles can be created in the system context from subset of system administrator rights.

The transition from pre-vCloud Director 8.20 role management happens during the upgrade to 8.20. Existing roles are transferred to role templates and each organization has its own roles instantiation based on the role templates. The UI has changed and now includes an Organization column and filter. A new System organization is added with default System Administrator role.

When a new organization is created, it has access to all rights that are used in role templates. The system administrator can grant additional rights to the organization with the vCloud API only:

GET `/api/admin` … get references to all rights in VCD instance

GET `/api/admin/org/<org-id>/rights` … get references to all rights in the organization

PUT `/api/admin/org/<org-id>/rights` … edit rights in the organization

System administrator or Organization Administrator can create new roles in its organization with vCloud API only:

POST `/admin/org/<org-id>/roles`

Note While the system administrator can edit tenant roles in the UI, editing of a role based on a role template changed the role template and therefore changes it for all organizations.

The vCloud Director 8.20 graphical user interface no longer allows creation of global roles. Only organization-specific roles can be created and only by the system administrator. However, the legacy API (version 20.0 or earlier) can be still used to create (and edit) a global role which will in fact become a role template.

The system administrator can edit a role in a particular organization that is based on a role template directly in the GUI. This affects other organizations. If a right is removed, all organizations have that right removed from the role. If a right is added, the role in existing organizations does not have the new right added unless the organization already had access to the right. New organizations created after the role edit will inherit it completely.

7.3 Organization Virtual Data Centers

An organization virtual data center is a subgrouping of compute, storage, and network resources allocated from a provider virtual data center and mapped to an organization. Organization virtual data centers are deployment environments where vApps can be instantiated, deployed, and powered on.

7.3.1 Non-Elastic Allocation Pool VDC

The following are non-elastic allocation pool VDC design considerations:

- Allocation pool Org VDC is constrained by the size of primary resource pool (a cluster of provider VDC).
- Upon creation of non-elastic allocation pool Org VDC, a child resource pool is created with CPU and memory limits set to Org VDC allocation and reservations set to guaranteed *percentage x allocated value*.
- Tenant cannot over allocate memory.
- Tenant can over allocate vCPUs (CPU usage is limited by Org VDC CPU allocation).
- VM memory overhead is charged to the tenant.

7.3.2 Elastic Allocation Pool VDC

The following are elastic allocation pool VDC design considerations:

- Allocation pool Org VDC can spread over multiple resource pools/clusters of the provider VDC.
- vCPU speed in GHz must be specified (default 1 GHz). This unit is used to calculate how many vCPUs can be deployed into Org VDC (*allocation (GHz) / vCPU speed (GHz)).*
- Tenant cannot over allocate memory or vCPUs.
- One vCPU is not limited by the defined speed of vCPU. Instead, vCPUs share all allocated GHz.
- VM memory overhead is not charged to tenant.

7.3.3 Pay-As-You-Go VDC

Pay-as-you-go VDC provides instant committed capacity on demand. Resources (vCPU, memory and storage) are committed only when virtual machines or vApps are instantiated within the VDC. The tenant is charged only based on the committed resources (CPU and memory for powered-on VMs, storage for deployed VMs).

Set limits on the maximum size of the VDC to prevent a sudden spike of resource usage of one customer which would prevent other tenants from deploying new VMs.

7.3.4 Reservation Type VDC

A reservation type VDC is used for a dedicated service offering where the tenant obtains its own compute resources (vSphere cluster). The tenant can set reservations, limits, and shares at the VM level and manage resource oversubscription.

The maximum size of the VDC is the useable cluster capacity (after taking out HA N+1 and hypervisor overheads) lowered by necessary resource for edge gateway deployment.

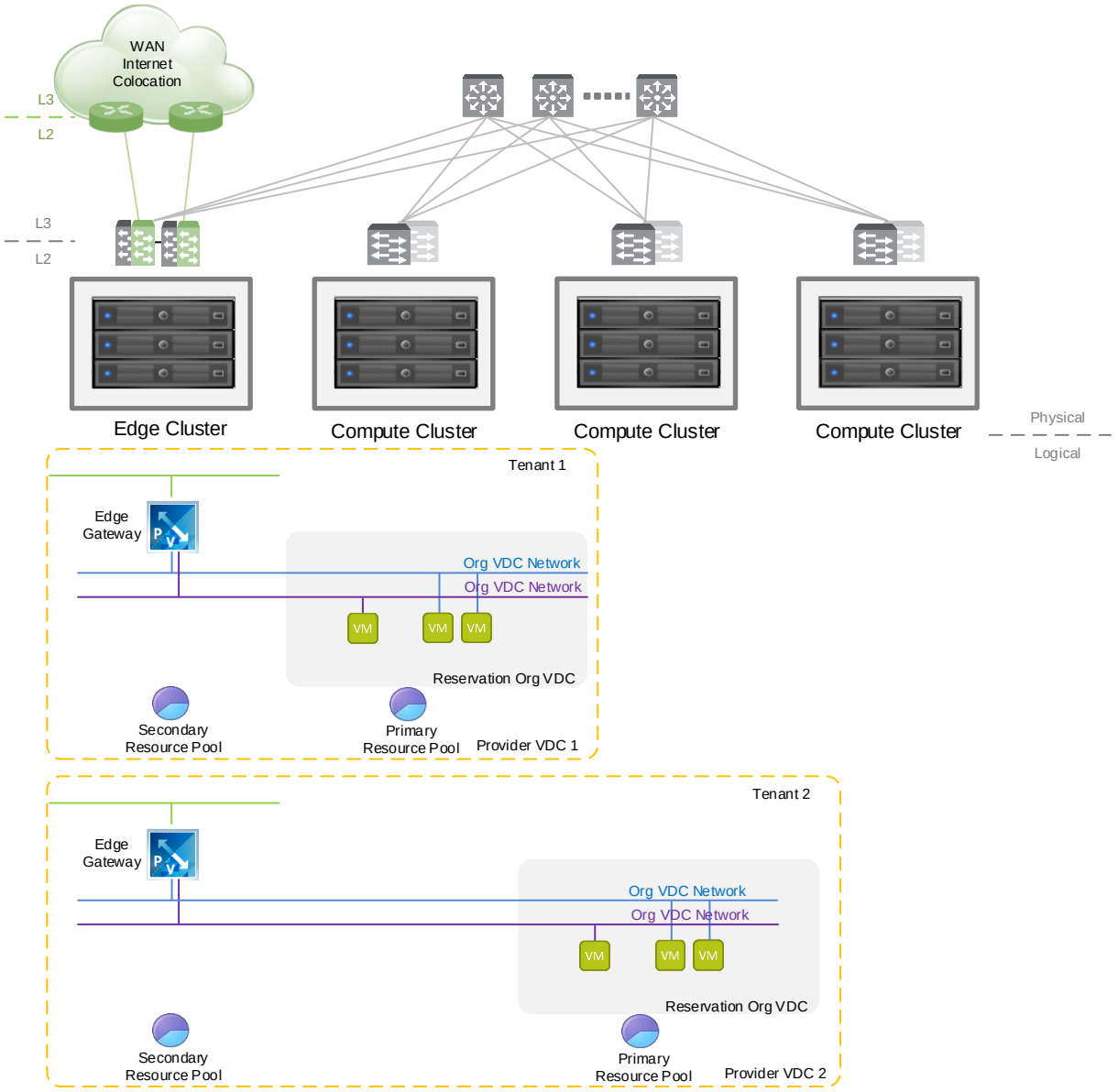
The edge gateways might be deployed into provider VDC secondary resource pools, which is useful in NSX Edge cluster scenarios (see Section 6.3.2.3, Design Option 2b – Combined Edge/Compute Cluster with Non-Elastic VDC). In this scenario, the cluster for edge gateways can be shared among many different provider VDCs by dividing it into manually created resource pools.

In the following figure, the primary resource pool of each provider VDC is a dedicated cluster, which can be as small as two hosts. The NSX Edge cluster contains multiple small resource pools, which are added to each Reservation Org VDC as secondary resource pools.

This design allows use of highly available NSX Edge instances (active/passive) when a tenant wants to buy only two host clusters and pay software licensing (for example, database software) for one host with

the second used only for failover. The vSphere HA admission control is then set to the dedicated failover host.

Figure 26. Shared Edge Cluster for Reservation Org VDCs



VMware does not recommend using a reservation type VDC in a shared cluster. The tenant can create an arbitrarily large VM because its physical compute resources are limited by the Org VDC resource pool and not by allocated VDC resources. (For example, inside an Org VDC allocated with 100 GB RAM, it is possible to power on a VM with 200 GB RAM.) If the VM memory is larger than the resource pool limit, the ESXi host running the VM will most likely result in swapping the memory it cannot physically provide. The swapping can create extensive load on the host and its storage, and possibly have a negative impact on other tenant workloads.

7.3.5 Org VDC vSphere Resource Settings

The following table summarizes how each Org VDC allocation type applies compute resource settings at vSphere level.

Table 11. Org VDC vSphere Resource Settings

	Pay as you go	Allocation Elastic	Allocation Non-Elastic	Reservation
Elastic	Yes	Yes	No	No
vCPU speed	Impacts VM CPU limit	Impacts number of running vCPUs in Org VDC	N/A	N/A
RP CPU limit	Unlimited	Org VDC CPU allocation	Org VDC CPU allocation	Org VDC CPU allocation
RP CPU reservation	None, expandable	**Sum** of powered-on VMs (CPU guarantee **x** vCPU speed **x** # of vCPUs)	Org VDC CPU allocation **x** CPU guarantee	Org VDC CPU allocation
RP RAM limit	unlimited	unlimited	Org VDC RAM allocation	Org VDC RAM allocation
RP RAM reservation	None, expandable	**Sum** of powered-on VMs (RAM guarantee **x** vRAM), expandable	Org VDC RAM allocation **x** RAM guarantee	Org VDC RAM allocation
VM CPU limit	vCPU speed **x** # of vCPUs	unlimited	unlimited	custom
VM CPU reservation	CPU guarantee x vCPU speed x # of vCPUs	0	0	custom
VM RAM limit	vRAM	unlimited	unlimited	custom
VM RAM reservation	vRAM x RAM guarantee + overhead	0	vRAM x RAM guarantee + overhead	custom

RP … resource pool

7.4 Networks

To support multi-tenancy, vCloud Director integrates with NSX Manager to manage the creation of isolated Layer 2 networks. All networks created from vCloud Director are backed by vSphere port groups. Access between cloud networks is governed by the NSX Edge virtual routers.

vCloud Director does not support universal NSX objects. However, port groups belonging to a universal logical switch can be used as backing for multiple external networks.

vCloud Director generates unique VM MAC addresses based on its installation ID (1-63) with the following scheme: 00:50:56:ID:xx:xx.

7.4.1 Network Pools

A network pool contains network definitions used to instantiate organization VDC and vApp networks. Networks created must be isolated at Layer 2 from all other networks.

A VXLAN network pool is created automatically when the provider VDC is created, if the VXLAN fabric is prepared in the NSX Manager governing the provider VDC vCenter Server.

A VXLAN network pool can span multiple distributed switches if they are managed by one vCenter Server. The VXLAN transport zone (scope) is configured in NSX Manager by vCloud Director automatically when the provider VDC is created or reconfigured with additional vSphere resource pools (clusters). The VXLAN transport zone always defaults to multicast control plane mode.

Do not us legacy VLAN-based, port group-based, and vCloud Director Network Isolation (VCDNI) based network pools in large service provider use cases because they have limited scale.

vSphere 6.5 does not support VCDNI network pools. vCloud Director 8.20 does not allow creation of new VCDNI network pools. It provides migration option to transition VCDNI backed networks to VXLAN backed networks from the UI or with vCloud API with a minimal network disruption[4].

7.4.2 External Networks

An external network in vCloud Director is a network providing external connectivity to organizations. Each external network is backed by one VLAN or VXLAN based port group on a virtual switch with uplinks to outside networks. This port group must be pre-provisioned before the external network can be created.

Note the following:

- Allow Overlapping External Networks must be enabled in vCloud Director settings to provision multiple VXLAN based external networks because they share the same transport VLAN.
- If the logical switch spans multiple vSphere Distributed Switch instances, it is backed by multiple port groups. The correct port group must be selected based on the needs (for example, compute cluster vSphere Distributed Switch port group).

External networks are used to provide Internet access, dedicated access to a direct connect type service, or access to shared services (syslog, OS licensing, patching).

7.4.3 Organization VDC Networks

Organization VDC networks provide an organization with a private network where vApps can be connected. Directly connected Org VDC networks are created by vCloud system administrators and are directly connected to an external network. Routed or isolated Org VDC networks are instantiated through

[4] https://kb.vmware.com/kb/2148381

network pools by organization administrators. Org VDC networks can optionally be shared among other Org VDCs in the same organization.

7.4.4 vApp Networks

vApp networks are networks that connect virtual machines within a vApp. These networks cannot span beyond a single vApp. The connectivity options include:

- Isolated
- Connected to the Org VDC network (direct or routed)

End users choose whether to place their vApp behind a vApp network. The majority of vApps connect to organization VDC networks. If the end user chooses to fence a vApp, a vApp network is created automatically when the vApp is deployed.

7.4.5 vCloud Director Edge Gateways

vCloud Director deploys edge VMs to provide Organization VDC or vApp network connectivity. The actual deployment is done through NSX Manager, but it is vCloud Director that makes the decision about placement and configuration of the edges. The vCloud Director edge gateway provides connectivity between one or more vCloud Director external networks and one or more Organization VDC networks. It is deployed inside the provider VDC in a special System VDC resource pool on a datastore belonging to the Org VDC default storage policy. The vCloud Director placement engine selects the most appropriate cluster where the edge gateway VM is deployed based on which clusters belong to the provider VDC, their available capacity, and most importantly, their access to the appropriate storage and external networks.

In vCloud Director 8.0 and earlier, Organization VDC and vApp edge gateways are deployed in vShield (legacy) compatibility mode (NSX Edge version 5.5.4). In vCloud Director 8.10 and 8.20, edge gateways and vApp edges are deployed as full NSX Edge nodes (version 6.x) with the same feature set, accessible through the user interface or API, as legacy NSX Edge nodes.

Legacy edge gateways deployed before the upgrade to vCloud Director 8.10/8.20 are still supported. VMware recommends redeploying the old edges in vCloud Director or upgrading them in VMware NSX to leverage the more efficient message bus communication mode with NSX Manager as opposed to the legacy VIX API mode. If the NSX Edge nodes are upgraded directly in VMware NSX, verify that vCloud Director is still running because it needs to be notified about the NSX Edge version change.

vCloud Director 8.20 enables additional NSX services on Org VDC Edge Gateways by converting them to Advanced Gateway. If the backing NSX Edge Gateway is still version 5.5.4, it is redeployed to version 6.x. vApp Edges cannot be converted to Advanced Gateway. The following table provides comparison for the Org VDC Edge Gateway before and after conversion to Advanced Gateway.

Table 12. Org VDC Edge Gateway Feature Set

Feature	Regular	Advanced Gateway
Routing	Static	Static, OSPF, BGP
Firewalling	Basic	Yes, with objects and IP sets
DHCP	Basic	DHCP bindings, relay
NAT	Basic	TCP/UDP/ICMP/any protocol NAT IP/CIDR/ranges can be used

Feature	Regular	Advanced Gateway
Load balancing	Layer 4	Up to Layer 7 with SSL termination, X-header forwarding, custom health check, application rules and TCP L4 acceleration
IPsec VPN	Yes	More flexibility in configuration (PSK characters, PFS can be disabled, DH 2, 15 and 14 groups)
SSL VPN-Plus (client-server VPN)	No	Yes
Layer 2 VPN	No	Yes (with another Org VDC Edge Gateway or with Standalone Edge[5])
CLI	No	Read only
Syslog	API only	Yes
API	vCloud API	vCloud Director API for NSX
UI	Legacy Flash UI	New HTML5 UI

Each Org VDC Edge Gateway can have up to 10 network interfaces that can be external or internal. In vCloud Director 8.20 internal interfaces can be converted to subinterfaces which creates one trunk interface that can have up to 200 subinterfaces.

In vCloud Director 8.20 Org VDC Edge Gateways can be deployed in four form factors (Compact, Large, Quad Large, and X-Large) and optionally in high availability (application active/passive) mode. vApp and DHCP Edge Gateways are always deployed in compact, single-node configuration.

Table 13. Org VDC Edge Gateway Form Factors

Edge Gateway Size	vCPU	RAM (MB)	Purpose
Compact	1	512	Moderate usage of networking services
Large	2	1024	Large number of concurrent SSL VPN sessiosn
Quad Large	4	1024-2048*	High throughput, high connection rate
X-Large	6	8192	Load balancing with millions of concurrent sessions

* Depends on NSX Version

[5] https://docs.vmware.com/en/VMware-NSX-for-vSphere/6.3/com.vmware.nsx.admin.doc/GUID-C9E2B0E4-F1C1-44A7-B142-F814F801FA42.html

External IP addresses are sub-allocated by the vCloud system administrator on the Internet networks. These addresses are used by organization administrators to configure NAT for internal VMs to allow access to and from the Internet.

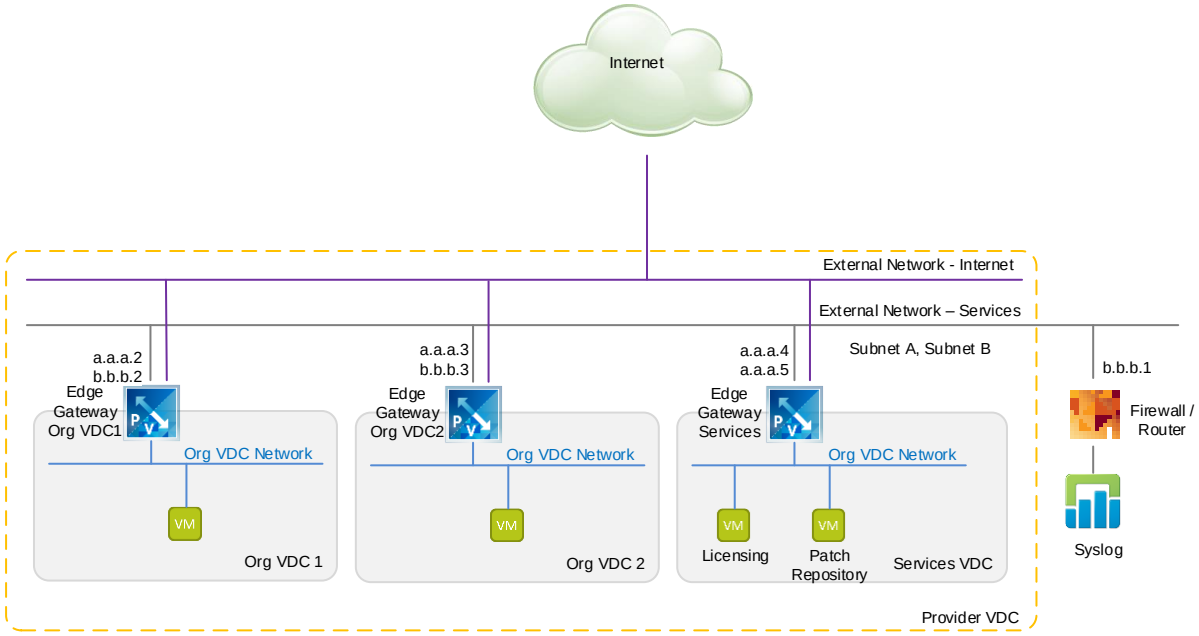
vApp Edge Gateways provide connectivity between an organization VDC network and a vApp network. They always have only one external and one internal interface. They are also deployed by vCloud Director to the provider VDC System VDC resource pool and exist only when the vApp is in deployed mode (powered on).

One arm DHCP Edge Gateways are deployed on isolated Org VDC networks and optionally on vApp networks to provide DHCP service.

7.4.6 Service Network Use Case Example

The following figure shows how shared services (patch repository, license management servers, and logging) can be provided securely to tenants. An external vCloud Director network is provisioned for the service network with two subnets assigned to it. One subnet is used for edge gateways to send their logs to an external syslog, and the other is used for IP address sub-allocation to each customer edge gateway with preconfigured source NAT rules.

Figure 27. Service Network



To provide access to the shared services and send the logs to syslog:

1. Configure the external syslog for edge gateways in vCloud Director under **Administration** > **System Settings** > **General** (IP b.b.b.1).
2. Sub allocate an IP address to the tenant from the external network—services subnet A.
3. Pre-create an edge gateway SNAT rule for this IP address applied on the logging network to reach services in the admin org (SNAT 0.0.0.0/0 > a.a.a.n).
4. Create DNAT rules on the admin edge gateway to reach the internal service VMs (DNAT a.a.a.4 > x.x.x.x, DNAT a.a.a.5 > y.y.y.y).

Tenants can now consume shared services (licensing, patching) on IP addresses a.a.a.4 and a.a.a.5 while Syslog receives only logs from Edge Gateways.

7.4.7 Distributed Firewall

The distributed firewall is a vCloud Director 8.20 feature that applies and enforces firewall configurations in the ESXi VMkernel at the vNIC level of virtual machines. This means the firewall can inspect every packet and frame coming and leaving the configured VMs, and is therefore, completely independent from the network topology and can be used for micro-segmentation of Layer 2 network. Both Layer 3 and Layer 2 rules can be created. It is managed at the Org VDC level from the Manage Firewall link.

At the NSX platform level, each tenant is given a section in the NSX firewall table and can only apply rules to VMs and Edge Gateways in their domain. There is one section for each Org VDC that has the DFW enabled, and it is always created on top (or optionally at the bottom if vCloud API is used to enable DFW at Org VDC level with `?append=true` suffix) of the firewall rule list. Because tenants might have overlapping IP addresses, all rules in the section are scoped to a security group with a dynamic membership of tenant Org VDC resource pools and therefore are applied only to VMs in the Org VDC.

Tenants can create Layer 3 (IP based) or Layer 2 (MAC based) rules while using the following objects when defining them:

- IP address, IP/MAC sets
- Virtual machine
- Org VDC network
- Org VDC

Note that using L3 non-IP based rules requires NSX to learn IP addresses of the guest VM. One of the following mechanisms must be enabled:

- VMware Tools™ installed in the guest VM
- DHCP Snooping IP Detection Type
- ARP Snooping IP Detection Type

IP Detection Type is configured in NSX at the Cluster Level in the Host Preparation tab.

The scope for each rule can be defined in the Applied To column. By default, it is set to the Org VDC. However, the tenant can further limit the scope of the rule to a particular VM, or Org VDC network (note that the vApp network cannot be used). It is also possible to apply the rule to the Org VDC Edge Gateway. In this case, the rule is actually created and enforced on the Edge Gateway as pre-rule which has precedence over all other firewall rules defined at that Edge Gateway.

The tenant can enable logging of a specific firewall rule with API by editing the `<rule … logged="true|false">` element. NSX then logs the first session packet matching the rule to the ESXi host log with a tenant-specific tag (Org VDC UUID subset string). The provider can then filter such logs and forward them to tenants with its own syslog solution.

7.5 Storage

7.5.1 Snapshots

Tenants can create one snapshot of a vApp or individual VM including its running state (memory). The following are snapshot design considerations:

- Org VDC requires enough free storage capacity for the maximum size of the snapshot, which is equal to the total size of provisioned disks and memory (if a snapshot of memory is chosen as well).
- Taking a snapshot has disk I/O performance impact. In time the size of snapshot grows, which must be accounted for capacity management. It is also disables network reconfiguration options on VM.
- It is possible to remove 'Create, Remove and Revert a Snapshot' from tenant roles.

7.5.2 Fast Provisioning

Fast provisioning saves time by using linked clones for virtual machine provisioning operations.

A linked clone is a duplicate of a virtual machine that uses the same base disk as the original, with a chain of delta disks to keep track of the differences between the original and the clone. If fast provisioning is disabled, all provisioning operations result in full clones.

It is possible to offload creation of linked clones during fast provisioning copy or clone operations to the storage array:

- Only NFS arrays are supported.
- NAS Agent must be installed on each ESXi host in the resource group.
- vSphere Storage APIs – Array Integration (VAAI) must be enabled on the storage array. The array might need additional licensing for creation of fast clones.
- VAAI must be enabled in vCloud Director on each datastore or datastore cluster.

The following are fast provisioning design considerations:

- A linked clone cannot exist on a different datastore than the original virtual machine. vCloud Director creates shadow virtual machines to support linked clone creation across vCenter Server instances and datastores for virtual machines associated with a vApp template. A shadow virtual machine is an exact copy of the original virtual machine created on the datastore where the linked clone is created. You can view a list of shadow virtual machines associated with a template virtual machine.
- Automatic pre-creation of shadow VMs to different storage policies or to different vCenter Server environments can be configured by adding the following row into the vCloud Director database `config` table: `valc.catalog.fastProvisioning=true`. This setting applies to all catalog VMs in all organizations, which means it could have large storage capacity overhead. The pre-creation is performed when the VM is inserted into a catalog and after a day is re-checked if the datastore status has changed.
- When a maximum chain length (full clone > linked clone > linked clone > …) is reached, the next clone results in a full clone. The default vSphere linked clone chain length is 30. The default VAAI chain length is 256. If storage arrays support VAAI offloaded fast clone operations with a shorter chain length, the VAAI chain length must be adjusted in the vCloud Director database in the `config` table, row `VirtualMachine.AllowedMaxVAAIChainLength`.

  Note The number of clones from the same full clone is not limited (the chain length is only 2).

- When the yellow threshold is reached on a particular datastore, fast provisioned clones are no longer created. Instead, a full clone is provisioned to another datastore.
- Manual (vSphere triggered) use of Storage vMotion of a linked clone creates a full clone.
- VAAI fast provisioning leverages VAAI native snapshots. Storage vMotion cannot be used on a virtual machine with native snapshot. Any relocations to different storage must be done from within vCloud Director with the VM in powered-off state.
- While creation of a vSphere linked clone is very fast, its I/O performance depends on the chain length. Writes go into a delta file, which is misaligned at the storage level and can create backend I/O overhead. Reads might traverse the entire chain to find the correct block and multiply needed I/O on the backend.
- Storage DRS supports vSphere linked clones both for placement and balancing. Only placement works with VAAI fast provisioning.
- Disks of fast provisioned VMs cannot be resized. If needed, the tenant can add fast provisioned VMs to its catalog and deploy them back. The Add vApp from Catalog action allows disk resizing.

- If an Org VDC is enabled for Fast Provisioning, it is not possible to create VMs with disks on different storage policies.
- The vCloud system administrator can create a full clone from a fast provisioned VM with the Consolidate action.

7.5.3 Datastore Thresholds

There are two thresholds at the datastore or datastore cluster level:

- Red – No more provisioning operations on datastore / datastore cluster
- Yellow – No more fast provisioning operations on datastore / datastore cluster

Because fast and thin provisioned VMs and their snapshots can grow, make sure there is enough reserve to prevent an out-of-space condition on the datastore;

7.6 Catalogs

Catalogs are the primary deployment mechanism in vCloud Director, serving as centralized repository for vApp templates and ISO media. Users self-provision vApps from vApp templates located in internal catalogs or global shared catalogs.

The following are catalog design considerations:

- Catalogs can be shared, published and subscribed. These rights are controlled with organization granularity.
- Catalogs can be shared to all or a subset of organizations. Sharing with a subset of organizations can only be performed by the vCloud system administrator because this is the only role with access to the list of organizations.
- Catalog publishing can be used for staging identical global shared catalogs among provider VDCs (vSphere instances) or vCloud Director instances (availability zones and regions).
- The catalog publish-subscribe mechanism uses vCloud Director cells' transfer share storage. Early catalog export keeps a pre-exported catalog copy on the transfer share of the source vCloud Director instance. This improves synchronization speed.
- The catalog synchronization schedule can be set (start/stop time and interval).
- Catalogs support versioning.
- A catalog can be deployed to a particular storage policy.

7.7 vApps

7.7.1 Overview

A vApp is a container for a software solution and is the standard unit of deployment in vCloud Director. It has power-on operations, consists of one or more virtual machines, virtual networks, and metadata, and can be imported or exported as an OVF package.

When a vApp is started, it deploys virtual networks and virtual machines and consumes organization virtual data center reservations. A partially powered-on vApp is a vApp state where one or more VMs within the vApp have been powered down, but networks and reservations are still consumed. A vApp in Stopped state releases all but storage resources.

7.7.2 vApp Deployment

Standardized vApp templates for common guest operating systems are usually provided by the service provider in a global share catalog. The tenants have an option to create their own vApps or import them through vCloud Connector or OVF import.

The following are vApp deployment design considerations:

- When creating a VM within a vApp, select the correct operating system in the VM properties. An incorrectly specified operating system impacts the presented virtual hardware.
- Install VMware Tools™ to take advantage of guest operating system shutdown action and guest customization scripts (IP assignment, and so on).
- VM configuration limits can be enforced upon deployment via blocking task (see the VMware Knowledge Base article *CPU and Memory Limit enforcement for vCloud Director* at https://communities.vmware.com/docs/DOC-21694. (Third-party Web sites are not under the control of VMware, and the content available at those sites might change.)
- vApp state (memory) can be saved in the catalog. vApp with a state (in Suspended state) can be moved between Org VDCs. However, the state does not persist in the case where OVF import x export is leveraged. (This is also valid for Org VDC migrations between vSphere instances.) vCloud Director does not guarantee compute compatibility (for example, migration between an Org VDC running on an Intel platform and an AMD platform).
- Only SCSI disks can be resized. No guest partition extension is performed. Disks cannot be shrunk. Resizing a disk during deployment from the catalog in a fast provisioning-enabled Org VDC results in a full clone with much slower deployment time.
- Not all OVF extra configuration (see the *Configuring a Whitelist for VM advanced settings in vCloud Director* article at http://www.virtuallyghetto.com/2014/05/configuring-a-whitelist-for-vm-advanced-settings-in-vcloud-director.html) is supported by vCloud Director.

7.7.3 Guest Customization

vCloud Director supports customizing virtual machine guest OS during deployment or a network configuration change. The following actions are supported:

- Changing computer name
- Changing root/administrator password
- Joining domain (Windows only)
- Changing SID (Windows only)
- Changing network settings
- Running a customization script

The following are guest customization design guidelines:

- Guest customization can be enable and disabled at the VM level.
- VM Tools must be installed in order to run guest customization.
- Full guest customization is run upon deployment of a VM from a catalog or when explicitly run by the Power On and Force Recustomization task.
- When network settings change (a new NIC is added to a VM, or an IP address assignment is changed), partial guest customization is performed upon next power on, changing only network settings.
- vCloud Director does not support IPv6 for network IP address assignment.

- Guest customization of older Windows OS (for example, 2003, XP, and so on) requires sysprep files to be properly installed on each vCloud Director cell.

7.7.4 VM Auto Import

vCloud Director 8.20 provides a VM auto import feature. The system administrator can simply drag any vSphere VM into an Org VDC resource pool and vCloud Director automatically discovers the VM and imports it. The VM can even be in the running state. This enables migration use cases, or allows service providers to easily offer self-service access to their fully managed vSphere only environments by simply connecting them to vCloud Director.

For each imported VM a separate special vApp is created with a *Discovered* prefix. While these vApps resemble regular vApps, they are not real vCloud Director vApps until they are adopted. The adoption takes place when the VM inside the vApp is somehow reconfigured after being imported to vCloud Director for management.

By default, VM discovery is enabled for every Organization in vCloud Director. It can be disabled in General Settings (UI or API) or with CMT command on VCD cell:

```
cell-management-tool manage-config -n managed-vapp.discovery.activated -v false
```

This behavior can be overridden at the Org VDC level with the `<VmDiscoveryEnabled>` API element.

Differences between a discovered and an adopted vApp are described here.

Discovered vApp:

- Looks like a regular vApp
- Can have only one VM per discovered vApp
- vApp contains API element `<autoNature>true</autoNature>`
- When the imported VM is deleted in VC or VCD, its vApp object is automatically purged
- Is owned by the system
- Is not subject to Org lease settings

Adopted vApp:

- Treated as a regular vCloud Director vApp
- Can contain multiple VMs, vApp networks, and so forth
- A discovered vApp is adopted when it is reconfigured

Other considerations include the following:

- The discovery process runs in the background every 3 minutes
- A failed VM import is retried after 60 minutes. This can be changed with the CMT command. The following is an example using 25 seconds:

  ```
  cell-management-tool manage-config -n managed-vapp.discovery.retry-delay-sec -v 25
  ```

- The following VMs cannot be imported: Fault Tolerant VMs, VMs with creation/upload process in vCenter Server, templates, vCloud Director shell VMs
- VMs must be connected to an Org VDC network
- VMs can be running or powered off
- VMs do not need to use an Org VDC storage policy. If it resides on an unknown storage policy, it is automatically relocated to the default Org VDC storage policy during the adoption. However, the VM must be in the powered off state.

- VMs with IDE controllers must be in powered off state
- VM CPU/RAM resources are changed based on Org VDC allocation type
- VM resources are not subject to Org VDC allocation restrictions, but are charged against it
- The VM name in vCenter Server remains intact until it is adopted and renamed
- New vSphere 6.5 guest operating systems are not recognized and are imported as Other (32-bit) OS
- By default, the minimal VM age is configured to 1 hour. This means VMs that were freshly reconfigured are skipped for the import. This is to "settle" the VMs first. The interval can however be changed with the following CMT command. In this example, the age set to 60 seconds.

  ```
  cell-management-tool manage-config -n VM_DISCOVERY_MIN_AGE_SEC -v 60
  ```

- Related to the minimal age, it is important that all cells and the vCloud database are using the same time configuration. Not only must the time be correct, but the time zone must also be configured identically.

Scalability

Achieving economies of scale means scaling vCloud resources in a consistent and predictable manner. The separation of management and customer workload components enables the use of the building block scaling approach.

8.1 Resource Group

The resource group clusters are running tenant workloads. A scale-up approach (increase the compute capacity of the ESXi hosts) is typically feasible only at the end of the life of the hardware components when the ESXi hosts are replaced with newer, more powerful hardware. The hardware in vSphere clusters belonging to the same provider VDC must be identical and not mixed to provide a consistent experience for the customer and for vSphere DRS efficiency. VMware recommends a horizontal scale-out approach. vCloud Director can scale its resources inside a cluster, inside a vCenter Server boundary, or outside vCenter Server boundaries.

8.1.1 Provider VDC Scalability

Customer workloads are running in organization VDCs which are provisioned inside a provider VDCs. Provider VDCs are mapped to vSphere resource pools or to clusters. One provider VDC can be mapped to multiple resource pools / clusters and create an elastic VDC, with the exception of the reservation type of Org VDCs. When a customer outgrows its organization VDCs and requires its extension, the request can be granted only if there is available capacity inside the provider VDCs. If there is no capacity available, a new organization VDC is carved up from another provider VDC and must be provisioned. This can make the customer VM deployment more difficult because the customer has to manage the capacity of multiple organization VDCs.

8.1.2 Cluster Scalability

The current vSphere (HA/DRS) cluster limit is 32 ESXi hosts for vSphere 5.5 and 64 for vSphere 6. Scalability of additional components (storage LUNs, network ports, and so on) might impact cluster sizes. The minimum cluster size is 2 ESXi hosts (with N+1 redundancy). Depending on the procurement cycle for new hardware, a number of standby ESXi hosts might be kept ready to add to any given cluster.

8.1.3 Scalability Within vCenter Server

Although a vCloud Director VXLAN-backed isolated or routed network can span multiple distributed virtual switches, directly-connected organization VDC networks, which are associated with an external network (that is coupled with a distributed switch port group), cannot. Live migration of VMs in the same provider VDC between different clusters is possible only if they share a common distributed virtual switch.

8.1.4 Scalability Across vCenter Server Systems

vCloud Director can manage up to 20 vCenter Server systems. It is possible to horizontally scale the resource group by adding additional vSphere environments. However, other systems or resource limits might impact the maximum scale.

Note Deployments from catalogs or moving/cloning vApps across vCenter Server systems results in a different copy process than within vCenter Server. The copy process within vCenter Server is handled by ESXi hosts directly through a storage network or network file copy protocol. The copy process across vCenter Server systems results in an OVF export that is staged and copied through the vCloud Director cell transfer storage. This process is much slower with a higher impact on the infrastructure (network bandwidth, vCloud Director cell compute, and transfer storage).

8.1.5 Storage

Datastores are assigned to vCloud Director provider VDCs through vSphere storage policies. It is easy to scale storage by adding new datastores to a storage policy in the clusters backing up provider VDCs. The maximum number of vCloud Director managed datastores is 500.

8.1.6 Networking

The supported maximum for number of external networks is 750. This number might impact the number of customers that can be onboarded for direct connect and Layer 2 bridging use cases (see Section 6.3.6, Other Network Services).

8.2 Management Cluster

All vCloud Director management components must be placed within their own ESXi clusters managed by their own vCenter Server. They can scale independently of resource group vCenter Server systems.

8.2.1 vCloud Director Database

There is always only one instance of the vCloud Director database per vCloud Director. Monitor its utilization, size, and storage I/O requirements and provision more resources with the vCloud Director growth. Microsoft SQL clustering (or Oracle RAC) is supported.

8.2.2 vCloud Director Cells

There can be multiple vCloud Director cells for load balancing and high availability. The supported maximum is 10 cells. The recommended minimum is 2 or *N*+1, where *N* is the number of resource group vCenter Server systems. If VMware Remote Console™ or OVF transfers are used extensively by the vCloud Director users, this can impact the vCloud Director cell utilization. Therefore, consider deploying new instances. In the case of a large number of vCenter Server systems with a small number of VMs, the following formula can be used:

Number of cell instances = *n*/3000 +1, where *n* is the number of expected powered on VMs

8.2.3 NSX Manager

There is always one-to-one relationship between the resource group vCenter Server and NSX Manager. Adding a resource group vCenter Server results in deployment of a new NSX Manager virtual appliance. The virtual resources of the NSX appliance (CPU and memory) can be increased, especially in environments with large churn of edge gateways.

8.3 vCloud Director Federation

While single instance of vCloud Director scalability limits can be sufficient for small service providers, VMware recommends a federation of multiple vCloud Director instances (described in Section 3.1.3, IaaS Multiple Regions) for larger service providers, particularly because of the separation of logical failure domains.

Recoverability

9.1 Overview

Business continuity and disaster recovery is a critical component of any cloud implementation. VMware vCloud Director introduces new challenges because of the additional layer of abstraction and the associated vCloud Director metadata. While management components can be protected with a standard mechanism (VMware Site Recovery Manager or third-party VMware compatible backup solution), coordination between backup vendor software, vCenter Server, and vCloud Director is required to complete the restoration of vCloud Director managed objects. This involves integration using the vSphere API, vCloud API, and VMware vSphere Storage APIs - Data Protection (VADP).

9.2 Management Cluster

For vCloud management, traditional methods can be used to back up physical and virtual machines. The following table lists the backup options for various components.

Table 14. Backup of Management Components

Item	Notes
vCloud Director cells	Image level backup of (stateless) vCloud Director cell. Protection of sensitive files (`global.properties` and `responses.properties` located in `$VCLOUD_HOME/etc/` directory) and certificates. Database backups of the vCloud Director database.
NSX Manager	Backup of NSX configuration and database to an FTP or SFTP site using the built-in user interface option in NSX Manager.
vCenter Server instances	Image level backups of vCenter Server systems. Built in backup of vCenter Server Appliance.
Databases	Database backup best practices from the database vendor.
ESXi hosts	vCenter Server Host Profiles.

Note Restoration of all components must be made to the same point in time to protect from inconsistencies that must be resolved manually.

9.3 Tenant Workloads

The provider can offer backup as a service for the tenant workloads. Leverage third-party vCloud Director compatible backup solutions to properly back up virtual machines and related vCloud Director metadata. This service is exposed through a third-party portal or vCloud API extensions.

Security

10.1 Guidelines

Best practices for security include the following tasks:

- Identify and disable unnecessary functionality and software.
- Identify interfaces that are not needed or wanted (VMCI).
- Remove all unnecessary accounts.
- Follow the principle of least privilege for service and administrator accounts.
- Disable unnecessary network services.
- Audit list of open ports and uses.
- Harden virtual machines.

10.1.1 Key Management and Encryption

vCloud Director requires HTTPS encryption for all server-to-client communication. Port 80 is open, but only redirects the connection to the secured connection port. The vCloud Director user interface is secured by SSL over HTTPS (port 443/TCP). The certificate is installed during configuration of vCloud Director and encrypted to file that is located at `$VCLOUD_HOME/etc/certificates`. This file is protected by the system.info value in the `global.properties` file that is located at `$VCLOUD-HOME/etc`.

The console-proxy connection is also secured by SSL on port 443/TCP. During the initiation of the connection, a key is passed to the browser to authenticate the console session to the console-proxy. This key has an expiration on it to help mitigate a replay attack.

Create the SSL certificates with a key length of at least 2048 bits. Protect the key-store file used to configure the certificates by using a complex password and then removing it from the vCloud Director cells after the configuration is completed.

vCloud Director database communication is sent in plain text over the wire. Therefore, access to the network must be restricted.

When using integrated LDAP authentication, the communication with LDAP server must use encrypted secure LDAP (LDAPS). Otherwise, user credentials can be snooped because they are transmitted in plain text.

Communication from vCloud Director cells to vCenter Single Sign-On (PSC), vCenter Server and NSX Manager is encrypted. Furthermore, it is possible to enhance the security by uploading certificates of each component through the JCEKS KeyStore file.

Disable unsecure SSL 3 and TLS 1.0 ciphers on vCloud Director cells.

10.1.2 vCloud Configuration Sensitive Files

vCloud Director installations contain security-sensitive data in the following location: `$VCLOUD_HOME/*.properties`.

The `global.properties` file contains a hashed copy of the database password for vCloud Director, as well as a randomly created string that is used to decrypt certain configuration information held within the database and the certificates file.

The certificates file contains the private and public keys of the SSL certificates for the HTTP and console proxy services in an encrypted format.

10.1.3 Web Application Firewall

vCloud Director Administration Web URL and REST API URL can be filtered from the Internet at the load balancer level to allow administration only from internal networks and to disable any external attacks.

10.1.3.1 Web Portal

Tenants access the Web user interface with the following URL: https://<vcloud_domain>/cloud/org/<company_name> where <company_name> is the short name of the organization.

vCloud Director administrators use the following URL: https://<vcloud_domain>/cloud/.

Table 15. Web Application Firewall Allowed Web Portal URLs

Allowed URLs
/cloud/transfer/.*
/cloud/org/.*
/cloud/vmrc/.*
/transfer/.*
/tenant/.*

10.1.4 vCloud API

The VMware vCloud API provides support for developers who are building interactive clients of vCloud Director using a RESTful application development style. vCloud API clients and vCloud Director servers communicate over HTTPS, exchanging representations of vCloud objects. Among some VMware vCloud ecosystem applications that leverage vCloud API are vRealize Orchestrator (with vCloud Director plug-in), VMware vSphere PowerCLI™, vCenter Chargeback (vCloud Director Data Collector), vCloud Connector, vRealize Automation, vRealize Operations Manager (with vCloud Management Pack) and vcd-cli[6].

Currently there are multiple versions of vCloud API starting with 0.9 up to 27.0 (as of vCloud Director version 8.20). Supported vCloud API versions can be retrieved at https://<vcloud domain>/api/versions. Some versions are deprecated (see vCloud API documentation).

Expose only a limited set of APIs to the Internet. Tenant API calls must be accessible because they are related to the user and organization administrator operations (user scope) and required by ecosystem applications that leverage vCloud APIs. Expose additional API calls related to the provider operations only to vCloud administrators (provider scope). The separation between the user and provider scope can be made based on source IP addresses. Access from the Internet allows only the user scope APIs, while access from a defined group of service provider addresses allows the provider scope APIs.

The Web application firewall (WAF) must be used to filter the URL access based on the scope. It terminates the client SSL session, examines the content, and based on filter rules, allows or rejects the session. If allowed, another load balanced SSL session is created between the WAF and a vCloud Director cell.

[6] https://github.com/vmware/vcd-cli

Note All API calls except /api/versions and /api/sessions must be authenticated and access control is applied based on the account privileges.

10.1.5 vCloud Director API for NSX

vCloud Director 8.20 provides a new API for advanced networking services that is a multitenant-safe proxy API to the underlying platform NSX APIs.

- Tenancy concept, authentication and authorization is provided with existing vCloud API through which a valid session token must be acquired
- Authenticated calls to the NSX Proxy API provide direct access to the networking objects of the particular tenant (Org VDC Edge Gateways and Org VDC Distributed firewall) using vCloud Director UUID identifiers.
- The following URI paths are currently available:

  /network/edges/.*
  /network/firewall/.*

10.2 Audit Logging

vCloud Director stores an activity log in the vCloud Director database. The last 30 days of relevant activity log data is available to tenants.

Use external syslog servers to redirect relevant logs for audit and troubleshooting purposes. vRealize Log Insight can be used because it is a scalable enterprise grade solution that can collect log data through traditional syslog protocols or through agents both from Linux and Windows systems. It provides built in dashboards for quick analytics of collected data for vSphere, vCloud Director, VMware NSX, SQL and other solutions.

Edge gateway logs can be collected either by a provider to a central syslog server (accessible through a vCloud Director external network (see Section 7.4.5, vCloud Director Edge Gateways) or to a tenant syslog server.

The distributed firewall log is not accessible to tenants because the log is generated by the ESXi host enforcing the specific vNIC rule and sent to central ESXi configure syslog target. However, the tagging mechanism is available to filter logs of a particular tenant.

Operational Considerations

11.1 vCloud Director Monitoring

vCloud Director monitoring is performed using custom queries to vCloud Director through the admin API to retrieve consumption data on various vCloud components.

vCloud Director Audit (https://blogs.vmware.com/vsphere/2012/03/audit-reporting-with-vcloud-director.html) or similar PowerCLI based tools can be used. vRealize Operations together with the vCloud Director Management Pack can be used to provide performance and capacity usage dashboards both for provider or tenant contexts.

11.1.1 vCloud Services Monitoring

Monitor the state of the following services with an enterprise monitoring solution.

Table 16. vCloud Director Cells Monitoring

Parameter	Reason for Monitoring
Process Down	Verify that the following processes are running: • vmware-vcd • vmware-vcd-watchdog
Disk Space Used	Free space available in `$VCLOUD_HOME/data/transfer` Free space available in `$VCLOUD_HOME/logs/`
Port Health	Following TCP ports must be answering: • 80 (http redirect to https Portal / API) • 443 (https Portal / API) • 443 (VMRC proxy)

11.1.2 vCloud Log Monitoring

By default, vCloud Director cell logs are located in `$VCLOUD_HOME/logs`. The configuration of loggers that produce some of the log files is stored in `$VCLOUD_HOME/etc/log4j.properties` file. It is possible to set up additional loggers in order to redirect logs to external logging solutions – the *VMware Knowledge Base article Enabling Centralized Logging in VMware vCloud Director* at http://kb.vmware.com/kb/2004564.

Table 17. vCloud Director Logs

Log Name	Description
`cell.log`	Console output from the vCloud Director cells.
`cell-runtime.log`	Runtime log messages from the cell.
`vcloud-container-debug.log`	Debug-level log messages from the cell.

Log Name	Description
vcloud-container.info.log	Informational log messages from the cell. This log also shows warning on errors encountered by the cell.
vcloud-console-debug.log	Debug-level log messages from the cell remote console proxy process.
statsfeeder.log	VM metric retrieval (from vCenter Server) and storage (KairosDB) informational and error messages.
cell-management-tool.log	Cell Management Tool log messages from the cell.
vmware-vcd-watchdog.log	Informational log messages from the cell watchdog. It records when the cell crashes, is restarted, and so on.
diagnostics.log	Cell diagnostics log. This file is empty unless diagnostics logging is enabled in the local logging configuration.
YYYY_MM_DD.request.log	vCloud API request log.
cloud-proxy.log	vCloud Availability Cloud proxy log messages from the cell.
console-proxy.log	Remote console proxy log messages from the cell.
server-group-communications	Server group communications from the cell.

Each vCloud Director cell exposes a number of Mbeans through Java Management Extension (JMX) to allow for operational management of the server and to provide access to internal statistics. Any JMX client can be used to access the vCloud Director JMX service (for example, JConsole). For more information about the exposed MBeans, see the *VMware Knowledge Base article vCloud Director MBeans* at http://kb.vmware.com/kb/1026065.

11.1.3 Time Synchronization

Time synchronization is important for the following scenarios:

- An event occurs that requires the reconstruction of other past events with respect to time.
- An event does not take place because it was dependent on other events that should have executed in a specific time sequence but did not due to lack of time synchronization or incorrect time.

The ability to reconstruct past events from related infrastructure components provides administrators the opportunity to analyze problems and implement effective corrective actions.

Configure NTP synchronization to a common stratum layer source in the admin zone for all ESXi hosts and vCloud components. For virtual machines, use w32time for Windows and NTP for Linux.

Use the same time zone for vCloud Director cells and database.

11.2 VMware vCloud Director Patching

Update management is a key factor in maintaining the health, performance, and security of the cloud infrastructure. Keeping an infrastructure updated can be a daunting task for IT administrators, who must frequently track patch levels and apply security fixes. However, the entire infrastructure is at risk if updates are not performed dependably and routinely.

11.2.1 vCloud Director Cells

vCloud Director cells are patched manually by running executable file. Prior to the upgrade, the cell must be quiesced and put in maintenance mode with the cell management tool. If the update requires a database schema change, the script that updates the database must be performed only on one cell in vCloud Director installation and requires downtime of the entire vCloud Director environment. vCloud Director 8.20 provides an orchestrated upgrade of a multi-cell group which can be triggered from a single place and operation.

Patching a cell operating system can be done non-disruptively by leveraging load balancer and quiescing and shutting down one cell at a time in a group.

11.2.2 NSX Manager and NSX Edge Instances

NSX Manager is patched manually by uploading a patch file to the appliance.

Existing NSX Edge instances can be updated manually by resetting the NSX Edge vApp network or redeploying the edge gateway in the vCloud Director UI. Upgrading NSX Edge version 5 to version 6 through NSX Manager is currently not supported (as of vCloud Director 8.0).

Upgrade of NSX Controller instances and ESXi host vib components is done from within NSX management interface.