vCloud Automation Center Reference Architecture

vCloud Automation Center 5.1

This document supports the version of each product listed and supports all subsequent versions until the document is replaced by a new edition. To check for more recent editions of this document, see http://www.vmware.com/support/pubs.

EN-001070-00

You can find the most up-to-date technical documentation on the VMware Web site at:

http://www.vmware.com/support/

The VMware Web site also provides the latest product updates.

If you have comments about this documentation, submit your feedback to:

docfeedback@vmware.com

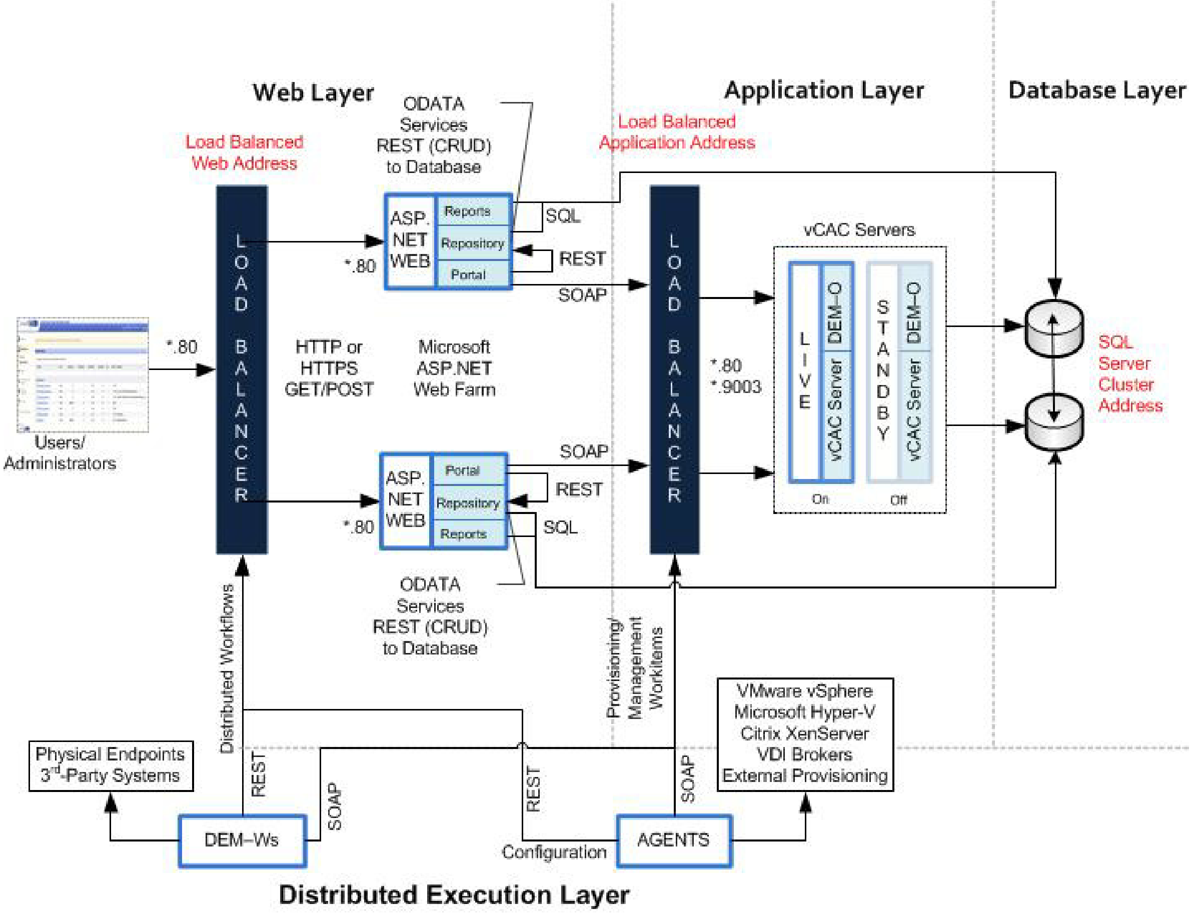
VMware, Inc.
3401 Hillview Ave.
Palo Alto, CA 94304
www.vmware.com

Overview

This document provides recommendations about deployment topology, hardware specifications, and scalability. For software requirements and supported platforms see the *vCloud Automation Center Support Matrix*.

vCAC Deployment Profiles

The following diagram shows a typical deployment topology for a vCAC instance:



The port numbers shown in the diagram are for an HTTP deployment. In HTTPS, all communication with the web servers and the Manager Service takes place over port 443. For a full reference to TCP ports used by vCAC, see the *vCloud Automation Center Installation Guide*.

vCAC Machines

The following table shows describes which components to install on each server profile in your deployment, along with their recommended hardware specifications.

Server Role	vCAC Components	Recommended Hardware Specifications
Database Server	vCAC Database	CPU: 2.4 GHz 4-core or equivalent RAM: 8 GB Disk: 40 GB Network: 1 GB/s
Web Server	Model Manager (Web and Data) Portal Website Reports Website	CPU: 2.4 GHz 4-core or equivalent RAM: 4 GB Disk: 40 GB Network: 1 GB/s
vCAC Server	Manager Service DEM Orchestrator	CPU: 2.4 GHz 4-core or equivalent RAM: 4 GB Disk: 40 GB Network: 1 GB/s
DEM Machines	(one or more) DEM Workers	CPU: 2.4 GHz 4-core or equivalent RAM: 4 GB Disk: 40 GB Network: 1 GB/s
Agent Machines	(one or more) vCAC Agents	CPU: 2.4 GHz 4-core or equivalent RAM: 4 GB Disk: 40 GB Network: 1 GB/s

High Availability Configuration

For high availability environments, install at least one redundant instance of each server profile above. For details, see the *vCloud Automation Center Installation Guide*.

Load Balancer Considerations

VMware recommends deploying at least two web servers in a production environment. Single–web server deployments should be used for development and testing environments only. Use the “least response time” method for balancing traffic to the web servers.

VMware also recommends enabling session affinity (or “sticky sessions”) to direct subsequent requests from each unique session to the same web server in the load balancer pool. If you do not want to enable session affinity on the load balancer, you can configure the vCAC web components in Web Farm Configuration using a session state database. For details, see the *vCloud Automation Center Installation Guide*.

A load balancer is also recommended to handle failover for the Manager Service. Because only one Manager Service is active at any one time, session affinity is not required for the Manager Service.

The load balancer can determine the availability of the vCAC components on each host by attempting to connect to the appropriate port, depending on the service, or making a GET request to a URL.

The following table summarizes the recommended balancing method for each server profile and how to determine the availability of the host.

Server Role	Balancing Method	Availability (HTTP)	Availability (HTTPS)
Web Server (Portal Website, Reports Website, Model Manager)	Least response time	Port 80	Port 443
vCAC Server (Manager Service)	Failover	Port 9003	https://<managerhost>/VMPS2

Initial Deployment Recommendations

This section describes a general deployment configuration for vCAC. These recommendations should be considered a starting point for deployment. After initial testing and deployment to production, you should continue to monitor performance and allocate additional resources if necessary, as described in the Scalability Considerations section below.

Database Deployment

VMware recommends that you always deploy with a dedicated database server to host the vCAC database.

Portal Website Configuration

The initial number of web servers depends on the expected size of your deployment and the number of machines that you plan to manage through vCAC.

Deployment Size	Web Server Recommendation
Small (up to 1,000 machines)	1 active web server, with 1 server for failover **Note:** In this configuration, the Manager Service can be cohosted with the web components.
Medium (up to 10,000 machines)	2 dedicated web servers with the Manager Service deployed separately
Large (more than 10,000 machines)	3+ dedicated web servers

Data Collection Configuration

The default data collection settings provide a good starting point for most implementations. After deployment to production, continue to monitor the performance of data collection to determine if any adjustments need to be made.

Distributed Execution Manager Configuration

In general, DEMs should be located as close to the Model Manager host as possible. The DEM Orchestrator must have strong network connectivity to the Model Manager at all times. VMware recommends that you have two DEM Orchestrator instances (one for failover) and begin with two DEM Workers (or at least one per data center location) in your implementation.

If a DEM Worker must execute a location-specific workflow, the Worker should be installed in that location. Assign skills to the relevant workflows and DEMs to ensure that those workflows are always executed by DEMs in the correct location. For information about assigning skills to workflows and DEMs using the vCloud Automation Center Designer console, see the *vCloud Automation Center What's New Guide*.

For best performance, DEMs and Agents should be installed on separate machines. For additional guidance about installing vCAC Agents, see the *vCloud Automation Center Installation Guide*.

Scalability Considerations

The following sections describe various performance characteristics of vCAC. It provides recommendations for your initial deployment based on anticipated usage and guidance for tuning performance based on actual usage over time.

Portal Website Scalability

Website latency, network traffic, and CPU usage on the web server and database server hosts increase with the number of concurrent users and the number of machines each user owns.

Performance Analysis and Tuning

In most cases, performance problems with the portal website can be alleviated by adding more web servers to the load balancer pool. In the case where the web server or database server is CPU-bound, performance can be improved by increasing the processing power of the server.

Data Collection Scalability

The time required for data collection to complete depends on several factors, including the capacity of the compute resource and the number of machines on the compute resource or endpoint, current system and network load, among other variables. The performance scales at a different rate for different kinds of data collection.

Performance Analysis and Tuning

As the amount of resources to be data collected increases, the time required to complete data collection may become longer than the interval between data collections, particularly for state data collection. You can determine whether data collection is not completing in time and being queued by checking the Data Collection page for a Compute Resource or Endpoint. If the Last Completed field always displays “In queue” or “In progress” instead of a timestamp when data collection last completed, you may need to decrease the data collection frequency (that is, increase the interval between data collections).

Alternatively, you can increase the concurrent data collection limit per agent. By default, vCAC limits concurrent data collection activities to two per agent and queues requests that are over this limit. This allows data collection activities to complete quickly while not affecting overall performance. It is possible to raise the limit to take advantage of concurrent data collection but this should be weighed against any degradation in overall performance.

If you do increase the configured vCAC per-agent limit, you may want to increase one or more of these execution timeout intervals. For more information about configuring data collection concurrency and timeout intervals, see the *vCloud Automation Center Operating Guide*.

Data collection is CPU-intensive for the Manager Service. Increasing the processing power of the Manager Service host can decrease the time required for data collection overall.

Data collection for Amazon EC2 in particular can be very CPU-intensive especially when running data collection on multiple regions concurrently and when data collection has not previously been run on those regions. This can cause an overall degradation in website performance. Decrease the frequency of Amazon inventory data collection if it is having a noticeable effect on performance.

Workflow Processing Scalability

Average workflow processing time (from the time that the workflow is ready to be preprocessed by the DEM Orchestrator to when it finishes executing) increases with the number of concurrent workflows. Workflow volume is a function of the amount of vCAC activity, including machine requests and some data collection activities.

Performance Analysis and Tuning

You can use the Distributed Execution Status page to view the total number of workflows that are in progress or pending at any time, and you can use the Workflow History page to determine how long it takes to execute a given workflow.

If you find that there is a large number of pending workflows at any one time, or that workflows are taking longer to complete, the general solution is to add more DEM Worker instances that are available to pick up workflows. Each DEM Worker can process 15 concurrent workflows; beyond this limit, workflows are queued for execution.

Additionally, you can adjust workflow schedules so as to minimize the number of workflows that are scheduled to be kicked off at the same time. For example, rather than scheduling all hourly workflows to execute at the top of the hour, you can stagger their execution so that they do not compete for DEM resources at once.

Some workflows, particularly certain custom workflows, can be very CPU-intensive. If the CPU load on the DEM Worker machines is high, consider increasing the processing power of the DEM machine or adding more DEM machines to your environment.