VMware Cloud Well-Architected Framework for VMware Cloud on AWS

AWS

VMware Cloud Well-Architected Framework

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

https://docs.vmware.com/

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

Contents

VMware Cloud Well-Architected Framework for VMware Cloud on AWS 8

1 Plan Pillar 9

- Planning Principles 9
  - Organizational Principles and Culture 9
    - Communication 10
    - Executive Sponsorship and Alignment 10
    - Organizational Readiness 11
  - Service Level Agreements and Objectives 12
  - Assessing Existing Workloads and Infrastructure 13
    - Discovery 13
    - Analysis 15
  - Designing for Scale, High Availability, and Recoverability 19
    - Design Process 19
    - Scalability 20
    - High Availability 21
    - Recoverability 22
- Spend Management 23
  - Terminology 23
  - Common Considerations 24
    - Seller of Record 24
    - Payment Method 24
    - Renewal Planning 25
    - Licensing 25
    - Terms of Service 25
    - Spend Management 25
  - VMware Cloud on AWS 26
- Shared Responsibility Model 27
  - VMware Cloud on AWS 27
  - Responsibilities 27
    - Customer Responsibility: Security in the Cloud 28
    - VMware Responsibility: Security of the Cloud 28
    - AWS Responsibility: Security of the Infrastructure 28
    - Shared Responsibility Matrix 28
  - References 30
- Infrastructure and Application Observability and Services 30
  - Infrastructure and Application Assessment and Native Services 30
    - Planning for Infrastructure Services in the Cloud 30

Retain Infrastructure Services Locally 30
Migrate to Cloud-Native Services 31
Centralized/Shared Infrastructure Services 32
Assessing Existing Infrastructure Services 32
Securing Infrastructure Services 33
Authentication and Directory Services 33

2 Build Pillar 35
Deploying VMware Cloud on AWS 35
Identity and Access Management Services for VMware Cloud on AWS 36
Identity & Access Management 36
Cloud Services Portal Roles 36
Granular Permissions for the VMware Cloud on AWS Service 37
SDDC Permissions 37
Infrastructure Services for VMware Cloud on AWS 38
Introduction 38
DNS 38
DHCP 38
Authentication Services 38
Automation of Infrastructure, Workload, and Security Services for VMware Cloud on AWS 38
Automation for VMware Cloud on AWS 39
Backup and Disaster Recovery for VMware Cloud on AWS 39
Backup and Disaster Recovery 39
VMware Site Recovery Manager (SRM) 40
VMware Cloud Disaster Recovery (VCDR) 40
Best Practices for Sustainability & Carbon Reduction for VMware Cloud on AWS 40
Sustainability Practices 40

3 Modernize Pillar 42
Modernize Introduction 42
Introduction 42
The Modernization Continuum 43
Future Application State 44
Use Known Patterns 45
Consider Portability 45
Start Fast and Learn 46
Modernization Through Rehosting 46
Rehost/Migrate 46
Network Strategy 46
Determine an IP Addressing Strategy 47
Transition Plan 47

Maintain 48

Hybrid Transition 48

Automate Workload Migration 48

Migration Waves and Events 48

VMware Migration Solutions 49

Modernization Through Re-platforming 53

Deliver Modernization Through Re-Platforming 53

Containerize the Monolith 54

Adjacent Cloud Services 55

Modernization Through Refactoring 57

Deliver Modernization through Refactoring 57

Deconstruct the Monolith 57

Backing Services 57

Processes 57

Concurrency 58

Disposability 58

Event Streams 58

Dependencies 58

Build Microservices 59

Modernization Operations 60

Operate in the Cloud 60

Verifying Application Functionality 60

Verifying Application Performance 60

Post-Migration Cutover Events 60

Post-Refactor or Replatform Events 61

4 Secure Pillar 63

Introduction to Security 63

What is Security? 63

Information Security Concepts 64

Authentication 64

Authorization 64

Availability 64

Compensating Control 64

Confidentiality 64

Data Breach 65

Identification 65

Incident 65

Integrity 65

Lateral Movement 65

Least Privilege 65

Non-repudiation 65

Recovery Point Objective (RPO) 65

Recovery Time Objective (RTO) 65

Security Control 65

Separation of Duties 66

Vulnerability 66

Shared Responsibility Model 66

Compliance 67

Regulatory Compliance 67

Compliance vs. Security 67

How Compliance is Achieved 67

VMware Cloud Trust Center 68

Design Considerations 68

Infrastructure Dependencies & Design Considerations 68

DNS Availability and Records 68

IP Addressing and Management 70

Network Address Translation 70

Storage Availability & Security 71

Templates and Container Registry 72

Backup and Restore 72

Management Interface Availability 73

Incident Response and Business Continuity 74

Administrative Access 74

Managing Administrative Access 75

Cloud Console Account Management 75

Role-Based Access Control (RBAC) 76

Virtual Private Network (VPN) 76

Private Network Links 77

Connected Accounts and Virtual Private Clouds (VPCs) 77

Network Perimeter Controls 78

Management Gateway 79

Management Appliance Access and Authentication 80

Endpoint Security 82

Endpoint and Workload Security 82

Microsegmentation 83

VMware Tools 84

In-Guest Controls 84

In-Guest Data-at-Rest Protections 85

Virtual TPM 85

Storage Policies 85

Multicast 86

Workload Resilience 86

Conclusion 86

5 Operate Pillar 88

Operate Introduction 88

Introduction 88

Infrastructure Operations and Service Control 88

Day-to-day Infrastructure Operations 89

Control Plane Management 89

Compute & Workload Management 90

Storage Management 90

Network Management 90

Performance & Capacity Management 90

Availability Management 91

Infrastructure Observability & System Health 91

Overview 92

Underlying Hardware Infrastructure & VMware Control Plane 92

Common metrics 92

Automation 94

Introduction 94

Benefits in Automation of operational tasks 94

CI/CD Pipeline & Automation 94

Continuous Assessment and Optimization 95

Workload rightsizing 95

Workload Optimization 95

Capacity Planning 96

OS & Application Patch Management 96

Cost Management 96

Introduction 97

Defining Cloud Cost Management 97

Sustainability 98

Introduction 98

Achieving Workload Energy Efficiency 98

Making Energy Visible 98

Maximizing Host Utilization 99

Designing Compute-Efficient Applications 99

Achieving Workload Carbon Efficiency 99

Workload Placement and Scheduling 100

Carbon-aware Workloads 100

VMware Cloud Well-Architected Framework for VMware Cloud on AWS

The *VMware Cloud Well-Architected FrameworkVMware Cloud Well-Architected Framework* provides a set of best practices and design principles for organizations to ready themselves for VMware Cloud Infrastructure.

Intended Audience

This document provides design principles and guidelines and is meant for a variety of audience within an organization. From the CTO, CFO, architects, and implementers /Operators, customers interested in understanding the end-to-end cloud journey of their organization will benefit from this framework.

Plan Pillar

1

The plan pillar discusses the importance of the various infrastructure services that are in use in an on-premises deployment. When a VMware SDDC is deployed in the cloud, default configurations related to infrastructure services are implemented at a rudimentary level to support the VMware components that make up the SDDC. There is no requirement to establish infrastructure services in the cloud before deploying a VMware SDDC because the Service Provider has accounted for these services as part of the initial deployment.

This chapter includes the following topics:

- Planning Principles
- Spend Management
- Shared Responsibility Model
- Infrastructure and Application Observability and Services

Planning Principles

Whether you call it Digital Transformation, Cloud First or Application Modernization, most businesses are going through some form of transformation to become more competitive in this digital era. According to a recent survey by VMware , 91% of executives agree their major application initiative in 2021 is to migrate and modernize legacy applications. The promise of cloud is its potential in transforming an organization to create and deliver new digital experiences for their end users.

Organizational Principles and Culture

Consumption of cloud services and new tooling will demand organizations learn new skills, adapt existing processes, and update automation workflows. The speed in which an organization can achieve cloud transformation will heavily depend on executive sponsorship, organizational readiness, and clear communications.

A traditional operating model relies predominately on managing the procurement of physical assets, lifecycle management, end-to-end visibility, and ownership across the entire technology stack. Transitioning to a cloud operating model will change the organizational dynamics that many organizations are currently operating under.

Communication

Cultural transformation starts with a clear purpose, supported by succinct and continuous communication. It is imperative that as organizations embark on the necessary changes to support their move to a cloud operating model. The reasoning for the move should be clearly understood. Document the business outcomes and ensure it is visible and available to all stakeholders in the organization.

To change the culture and operating model of an organization takes time. Prioritizing key workstreams, attaching timelines, and assigning clear ownership and accountability is critical to the transformation of the organization. Involvement from key stakeholders across the business is important in ensuring all requirements are communicated and captured early. Ambiguity will cause delays, budget overages, and may negatively impact business outcomes.

How an organization communicates can either inhibit or accelerate the desired culture and business outcomes. While not an exhaustive list, it is important to consider the following as an organization plans their communications to support a cloud operating model:

- Communication Promotes Motivation
  - Clearly articulate expectations and outcomes
  - Communicate the importance of each contribution against business outcomes
- Ownership Drives Accountability
  - Provide clear and consistent lines of accountability across areas of change
  - Provide a clear and documented scope of responsibilities
  - Provide clear guidance, measurements, and criteria for success
- Inclusion Fosters Success
  - Ensure those impacted by the transformation have their perspectives included
  - Create an environment for immediate feedback and contrarian views

With any organizational change, it is not uncommon for new requirements to surface, introducing unknown constraints during execution. Organizations should have processes in place to incorporate new requirements into existing workstreams.

Executive Sponsorship and Alignment

Digital transformation and Cloud Initiatives are most often born and driven by executive leadership. At the highest level of any organization, there is often strong alignment on these strategic initiatives, or at a minimum, agreement to proceed with organizational transformation.

As organizations prepare for and execute against their transformational initiatives, issues will arise. It is imperative that support, alignment, and clear sponsorship at the executive level are present and accessible to those driving the supporting workstreams and keep the following considerations in mind:

- Hold themselves publicly accountable for its success

- Champion and lead by the example the cultural changes taking place
- Make timely decisions to positively impact progress
- Remove roadblocks to ensure consistent forward progress
- Allocate resources (people, infrastructure, tooling, licensing, etc.) as needed
- Remove stalemates to keep momentum and remove tension
- Stay up to date, and course correct where nessisary

As an organization prepares to pursue a transformative initiative, alignment and sponsorship will have a direct and measurable impact. It is important to continually have open discussions to ensure the vision, strategy and execution have alignment across stakeholders. Consider the following points as you prepare for the transformation:

- How do organizational stakeholders envision their overall, future Cloud position?
  - Are there broad understanding of what success looks?
- Who will lead the organization through its strategic planning including goals, objectives, and actions?
  - How are decisions communicated? Is feedback welcomed and open?
- How are success criteria for the cloud journey defined, and what are the specific and measurable KPIs to action against?
  - What are the measurements of success as defined by leadership, and do they correctly translate into KPIs that can be measured and managed?
- How are strategic management processes documented, regularly analyzed, and improved?
  - Is this information gated, or made openly available to all individuals in the organization?
- How does the current culture, and go-forward strategy align for long-term success of the organization during, and post transformation?
  - Does the organizational culture foster long-term transformational success?

Organizational Readiness

The ability for an organization to transform its culture is ultimately driven by its people. Processes must continually be adapted to meet new requirements, and the use of technology must continue evolve to meet the needs of the business, but the people in an organization is the constant in any transformation.

Individual stakeholders must be empowered to drive and lead transformation across an organization. Below are some example questions to review internally with relevant stakeholders. These questions are used to help gauge the readiness of the organization for executing against its tranformation initative:

- Is there complete alignment across the organization on which infrastructure service provider(s) the business will align?
  - If not, how will this risk be mitigated to reduce the likelihood of delays to progress, rogue cloud deployments and associated costs?
- Do existing agreements exist with the providers of choice, or do they need to be negotiated?
  - Does the individual expertise exist internally to properly negotiate said agreements?
- Are finance and procurement teams aligned on how to move from a Capital Expenditure (CapEx) model, to one driven predominately by Operational Expenditures (OpEx) for infrastructure resource consumption?
  - Do budget allocations and structure easily allow for this shift?
- Do key individuals possess the technical skills across operations teams to support the businesses critical workloads and applications that are moving to, or will be developed in the cloud?
  - If so, how quickly can IT change their runbooks, their tooling, and extend their automation to factor in remote and unique infrastructure components?
  - If not, how do these individuals acquire the necessary skills – Retrain individuals? Hire desired skills into existing teams? Outsource?
- What is the internal perception of public cloud, and is it the same across all facets of the business?
  - Are there cloud adverse stakeholders as acting members of the decision tree?
- Does all existing systems and tooling have sufficient licensing to be deployed and run on remote, cloud provider owned infrastructure?
  - If not, what is involved in ensuring the proper cloud licensing is acquired?

The answers to these questions and many others will provide an organization an idea of the potential tradeoffs and/or risks. Organizations that engage in an open and early dialogue will gain a better understanding of the required changes to successfully execute their transformation.

Service Level Agreements and Objectives

One of the key differences between an on-premises environment and a VMware Cloud SDDC is in the responsibility of the infrastructure management. In an on-premises environment, an organization is responsible for managing the physical infrastructure, virtual infrastructure, and workloads.

With VMware Cloud, an organization is primarily responsible for managing and operating their workloads, while the VMware Cloud Provider manages the physical and virtual infrastructure.

Note For more details, please refer to the VMware Cloud shared responsibility model.

It is important for an organization to understand the Service Level Agreements (SLAs) and Service Level Objectives (SLOs) for a given VMware Cloud Provider to ensure it satisfies the needs of the organization. In VMware Cloud, an SLO defines the quality of service that a VMware Cloud Provider delivers to an organization. An SLA is a legally binding contract between a VMware Cloud Provider and an organization with specific terms and conditions of the SLOs.

When analyzing existing workloads, an organization should agree on a set of SLOs with the respective application owners. An organization should design a VMware Cloud environment that meets the SLOs for their end-users, while aligning to the established SLAs of the VMware Cloud Provider. In addition, an organization should monitor and log key metrics to ensure that a VMware Cloud Provider is meeting their SLAs.

Note VMware Aria Operations (SaaS) and VMware Aria Operations for Logs can be used to monitor the VMware Cloud SDDC and its workloads.

Assessing Existing Workloads and Infrastructure

The initial assessment of the existing infrastructure and workloads is critical to enable an organization to successfully onboard into a VMware Cloud SDDC.

The assessment consists of two main phases: discovery and analysis. Information about the existing infrastructure and workloads will be collected during discovery, and then the appropriate cloud migration strategy will be determined during the analysis phase.

Discovery

The first step of discovery is to build an accurate inventory within an organization's infrastructure.

Infrastructure Inventory

This includes, but not limited to, the following:

- Physical workloads (i.e., bare-metal nodes, firewalls, networks/VLANs)
- Virtual workloads (i.e., virtual machines, containers)
- Applications
- Third-party tools and integrations
- The inventory should include the application owner information, which can be used to conduct a more detailed interview.

An accurate inventory will help build a successful migration strategy. A combination of tools and/or interviews can be leveraged to create and validate the inventory.

Note Tools such as the vSphere Client, PowerCLI, VMware Aria Operations, VMware Aria Operations (SaaS), Application Transformer for VMware Tanzu, internal Change Management Databases (CMDB), and other third-party solutions can all be used to gather an inventory of both physical and virtual workloads.

Data Collection

Upon completing the higher level inventory collection, the next step is to collect detailed information about each application. Before an organization can determine the migration strategy for a particular application, a comprehensive understanding of its functionality is required. This involves the following information to be collected and validated through a combination of tools and interviews:

- Business function and criticality
  - How important is this application to the business?
  - What is the business impact if this application is not functional for a certain period?
  - The business impact can be tangible (i.e., lost inventory, legal penalties, lost transaction revenue) or intangible (i.e., brand damage, decrease in stock value, loss of employees)
- Is there an established SLA/SLO from the respective line of business within the organization?
- Compute and storage capacity requirements and resource utilization
  - What are the minimum capacity requirements?
  - What is the expected capacity growth over the next x years?
  - What is the current resource utilization?
  - Are there certain days and times or months in a year where there are peak demands for resources?
- Performance requirements (compute, storage, network)
  - What is the current performance baseline for compute, storage, and network?
  - Are there specific performance requirements that need to be met? (i.e., minimum IOPS, number of concurrent connections, provisioning time, etc.)
- Ingress and egress traffic flows and network utilization
  - Which network ports are required for traffic?
  - What are the average and peak network utilizations?
  - Are there periodic spikes in network utilization due to scheduled events, such as backup?
- Service dependencies (i.e., application dependency, third-party integration)
  - What is the current application architecture?

  - Does the application depend on other services and/or workloads for functionality?
  - How often does the application communicate over the network?
- Business continuity and disaster recovery requirements
  - Is there any single point of failures (SPOFs) for the application that should be mitigated?
  - What are the recovery time objective (RTO) and recovery point objective (RPO) requirements?

Note Tools such as VMware VMware Aria Operations, VMware Aria Operations (SaaS) and VMware Aria Operations for Networks can be used to collect this information. VMware Aria Operations can analyze the current resource consumption of virtual machines and recommend sizing for virtual machines, which can be used in cloud migration planning. VMware Aria Operations for Networks can provide data on traffic usage and help identify or validate firewall requirements. Other third-party solutions can also be used to collect similar information.

Analysis

Information gathered during the inventory collection is used to build a list of requirements for each application.

Requirements Gathering

For each application, the requirements can be organized by design attributes, that includes, but is not limited to, the following:

- Scalability - The ability for a system to continue providing the same level of performance or functionality when there is a change in utilization.
  - For example, an application architect has a requirement that the underlying infrastructure must be able to scale out dynamically if there is an increase in traffic load. This may translate to a scalability requirement where the underlying infrastructure must be able to add or remove a compute capacity within an acceptable time frame.
- Availability - The ability for a system to continuously operate and function for an extended time without interruption.
  - For example, the Virtual Machine workloads has a requirement that the underlying virtual infrastructure can provide an SLA of 99% uptime per month for management. This may translate to an availability requirement of a VMware Cloud SDDC having a minimum of 99% uptime or greater.
- Recoverability - The ability for a system to recover from a disaster or a failure.
  - For example, an application architect requires the data for the application cannot tolerate data loss for more than 2hrs. This may translate to a recoverability requirement where the date for the Recovery Point Objective (RPO) must not exceed 2hrs.

- Manageability - The measure of how easily a system can be deployed, configured, and controlled.
  - For example, an organization has a requirement to provide end users with a solution that enables self-service workload provisioning with governance. This may translate to an evaluation of a cloud management platform (CMP) that integrates with a VMware Cloud SDDC.

Note Tools such as VMware Aria Automation Cloud and other third-party solutions that are supported with VMware Cloud can be used as a CMP.

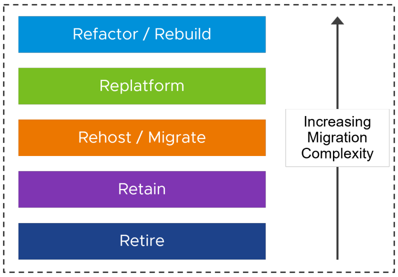
- Performance: the measure of how well a system accomplishes a given task.
  - For example, the application has a requirement to deliver at least 100 concurrent requests per minute to satisfy its service SLA. This will translate to the supporting virtual infrastructure provisioned with the necessary compute, network, and storage resources to meet the application requirement.

The finalized list of requirements must be validated with the appropriate stakeholders within the organization through workshops or interviews before continuing with the assessment.

VMware Cloud Migration Strategy

With clear requirements and a holistic understanding of the application inventory, an organization can now determine an appropriate migration strategy for each application based on the needs of the business.

Figure 1-1. Common Migration Strategies



- Refactor / Build involves changing the application at the source code level. Typically, applications are re-written to take advantage of cloud microservices architecture and to incorporate new services such as IoT, machine learning, and others
- Replatform involves changing the operating system, such as going from Windows to Linux, modifying the application middleware, such as going from a self-managed database to a cloud provider managed database or from a virtual machine to a container image
- Rehost / Migrate involves either changing the hypervisor. (e.g., migrate applications from one virtualized environment to another) which is known as Rehost or moving an application without changing the underlying hypervisor or application at a source code level (e.g., migrate VMs from one virtualized environment to another without requiring changes) which is known as Relocate

- Retain means leaving workloads and/or applications in a private cloud environment
- Retire means decommissioning workloads and/or applications, which can involve eliminating them altogether or converting to SaaS

It is important to understand that application modernization is not one specific approach but can be a combination of approaches. A common strategy that organizations have adopted to help accelerate their application modernization journey is first to migrate their existing workloads to a VMware Cloud SDDC and then modernize the underlying application.

Certain migration strategies, such as Refactor and Replatform, provide an opportunity for an organization to modernize their applications after migrating to VMware Cloud. The speed at which a modernization project is executed largely depends on an organization's business outcomes and timelines.

Organizations will be most successful in achieving their application modernization (app modernization) goals by leveraging a Migrate and Modernize strategy. By migrating existing Virtual Machine workloads to VMware Cloud, organizations will now have a modern infrastructure platform. Organizations will now be able to focus on their app modernization efforts.

Migration Network Connectivity

For each batch of workloads to be migrated, the migration path and method must be determined.

There are various network connectivity options to create a migration path from an on-premises environment to a VMware Cloud SDDC, such as using a VPN or setting up a direct, private connection. Analyzing findings from the infrastructure assessment would help identify feasible network connectivity options for an organization's specific business needs and requirements.

Note VMware HCX can be used to provide a private, secure connection and migrate workloads between an on-premises environment and a VMware Cloud based SDDC. VMware HCX also provides different migration methods, such as cold migration, bulk migration, and live migration.

Migration Wave Planning

Wave planning is the process of grouping workloads that will be migrated concurrently based on business critically and application dependencies to help create a high-level migration schedule. Workloads can be migrated based on the application SLAs, for example non-mission critical workloads can be migrated initially.

It is also critical to understand the different types of migration methods and an organization should select the one based on the needs of the business. For example, a Dev/Test workload which can afford downtime during the evenings, a production workload that cannot afford any downtime, and a staging workload which can have minimal downtime when scheduled. From a migration execution standpoint, you would then select three different migration types as mentioned below for each of the respective workloads, maximizing the speed at which you can migrate the workloads and maintaining the application service level agreements (SLA).

There are different methods for migrating workloads such as hot, warm, and cold migration:

- A hot migration is referred to as a live migration and is the most familiar to VMware administrators. It is a staged migration where the virtual machine stays powered on during the initial full synchronization and the subsequent delta sync, using the VMware vSphere® vMotion® feature.
- A warm migration is a virtual machine that is actively running while it is being replicated to ensure minimal downtime. After the migration completes, you either start a manual or automated cutover to make the replicated virtual machine available on the cloud provider. Cutover is a process of powering on the virtual machines at the cloud provider site after the warm migration gets completed. This cutover operation includes a final sync and import of the migrated VM into a destination VMware Cloud SDDC.
- A cold migration is a virtual machine that is in a powered-off state before starting the migration. Exporting and importing virtual machine images is another form of cold migration.

Migration waves should also incorporate the related application dependencies and network communication traffic to keep intra-application traffic within the same environment and limit traffic across an on-premises data center and/or a VMware Cloud based SDDC.

In addition to grouping by applications, isolated waves should be created for large or complex workloads, such as database virtual machines or virtual machines with a high data change rate. These workloads tend to require more network bandwidth for migration and could impact other migrations if performed concurrently.

Note VMware Aria Operations for Networks can be used to validate application dependencies and traffic flows. VMware HCX integrates with VMware Aria Operations for Networks can automatically create a VMware HCX Mobility Group which migrates pre-defined set of workloads based on a migration wave planning.

VMware Cloud Design

After a thorough assessment of the existing infrastructure and workloads, the results will guide an organization in creating a VMware Cloud SDDC using design decisions based on compute and storage sizing, service location selection, and network connectivity.

The requirements gathered during the assessment will guide the design decisions for all aspects of a VMware Cloud SDDC, such as compute and storage sizing, service location selection, and network connectivity.

Appropriate compute and storage sizing must be determined for the VMware Cloud SDDC include resources for management components and overhead as well as the expected growth of workloads when sizing the VMware Cloud environment.

Note VMware Cloud Sizer should be used to help an organization size the VMware Cloud based SDDC. The VMware Cloud Configuration Maximums document should also be referenced to ensure scalability of a VMware Cloud SDDC.

Service location for a VMware Cloud SDDC should be chosen depending on the requirements, such as the following:

- Availability of services: not all VMware Cloud services are available in every region
- User locations: depending on the business needs (i.e., market expansion, local security compliance) and application requirements (i.e., service feature availability, minimum latency for optimal performance), an organization’s service(s) may need to be in close proximity to their end users.

Network connectivity for a VMware Cloud SDDC depends on the business needs and requirements. If an organization decides to keep the on-premises environment and use VMware Cloud for bursting to meet unexpected demands or for disaster recovery of the on-premises environment, permanent network connectivity may be needed between the two environments. If an organization decides to decommission the on-premises environment, then only network connectivity for the migration will be needed. In addition to deciding on the longevity of a network connection, data on network utilization and application traffic flows collected during the assessment should be utilized to determine the type of network connectivity and the required bandwidth.

It is important to remember that a VMware Cloud SDDC design must meet all the identified requirements. A detailed example of how a VMware Cloud based SDDC can be designed to meet availability, recoverability, and scalability requirements is discussed in the next section: Designing for Scale, High Availability, and Recoverability.

Designing for Scale, High Availability, and Recoverability

Similar to an on-premises infrastructure, a VMware Cloud environment must also be designed for high availability, recoverability, and scalability.

The specific technical configurations will vary based on the specific VMware Cloud Providers. The design process and considerations discussed in the following sections apply to any VMware Cloud based environment.

Design Process

There are many ways to design each part of a VMware Cloud environment, such as host clusters and network connectivity.

Regardless, the final design should meet identified requirements within constraints and assumptions. Constraints are any limiting factors that may affect the design. They can be project-related, such as budget or timeline, or technical, such as an existing application architecture that cannot be altered. Assumptions can be made during the discovery, as not all the necessary information may be available immediately. Identified assumptions must be validated to make design decisions based on correct information.

In addition, any risks, whether they are technical or non-technical, that may arise from deciding on a particular design should be documented and addressed with potential mitigation strategies. Managing risk can vary based on the impact and level of effort. High impact risk items should likely be mitigated, whereas the organization may accept low impact risk items.

Once a particular design is finalized, it is important to document the decision, including justification and any related risks. It is also valuable to note which requirements have been fulfilled by a particular design to ensure that the final design meets all the identified requirements and business needs.

Scalability

Scalability does not have a standard metric used across the industry. Generally, performance or load testing can be done on a system to determine how flexible it can handle changes in demand.

For example, performance testing can be done to measure how long a system takes to add more resources and meet peak demand. A system that only takes one minute to add more resources is more scalable than a system that takes an hour to perform the same action.

Designing for Scale

There are several options for scalability of a VMware Cloud SDDC. An organization may begin with scaling up a vSphere cluster by adding hosts to meet an increase in demand. Organizations can enable automatic resource allocations such as Elastic Distributed Resource Scheduler (eDRS) which is available in a VMware Cloud on AWS SDDC. If an automated resource allocation service is not available in a VMware Cloud SDDC, the process of adding a host can be automated by leveraging the VMware Cloud SDDC APIs.

If scaling up a vSphere cluster reaches the VMware Cloud configuration maximums or does not fulfill the business needs, the next option is to scale out by creating additional vSphere clusters within the same VMware Cloud SDDC. It is important to determine when a new vSphere cluster should be created to plan for day two operations and manage the growth of an environment.

Depending on the business needs and the estimated growth, an organization may choose to create multiple VMware Cloud SDDCs instead of scaling up a single SDDC. It is essential to identify any workloads that must communicate with each other spanning several VMware Cloud SDDCs to design the network connectivity between these environments appropriately.

In addition to the virtual infrastructure, the VMware management Virtual Machines should also be considered when designing for scalability. Typically, the management Virtual Machines, such as the vCenter Server and NSX managers, will be deployed with a pre-defined set of compute and storage resources. Based on the workload requirements and expected utilization collected during the initial assessment, the VMware management Virtual Machines may need to be resized if this capability is available within a VMware Cloud SDDC.

Overall, a VMware Cloud environment should be designed to be repeatable. A modular design makes an environment easier to replicate or expand across different regions to meet fast-growing demand. An organization should plan how their VMware Cloud SDDC will expand when designing the first VMware Cloud SDDC to expedite the deployment of a new VMware Cloud SDDC as well as to simplify cloud management and day two operations in the future.

Tip VMware Cloud Sizer can be used to size the VMware Cloud based SDDC with data collected during the initial assessment. Configuration maximums can also be referenced to ensure scalability of the VMware Cloud SDDC.

High Availability

High availability can be achieved for physical infrastructure, virtual infrastructure, and application services. High availability is measured by uptime, the amount of time that a service has been operational.

Designing for High Availability

The VMware Cloud Providers typically have multiple physical data centers in various regions throughout the world. The data centers in each region are designed to be independent of one another so that a failure in one data center would not affect another. An organization can choose to deploy their VMware Cloud SDDC in one or more supported regions. The data centers in each region are designed to be independent of one another so that a failure in one data center would not affect another. The VMware Cloud Provider is responsible for providing high availability for the physical infrastructure according to the Service Level Agreements (SLAs).

A VMware Cloud SDDC inherently provides high availability for the Virtual Machines running in the SDDC using vSphere High Availability (HA) and vSAN storage policies. When an ESXi host fails, vSphere HA automatically restarts the Virtual Machines from the failed ESXi host to other ESXi hosts within the same vSphere cluster. vSAN storage policies provide data redundancy through appropriate RAID configurations depending on the number of ESXi host failures an organization can tolerate.

Although a VMware Cloud SDDC provides native high availability capabilities, designing a virtual infrastructure that meets the applications SLAs is ultimately the responsibility of an organization.

An organization can deploy multiple independent VMware Cloud SDDCs, each in a different region where the VMware Cloud Infrastructure Services provider is available. Workloads can be deployed across these environments to satisfy availability requirements. With this method, it is vital to understand the application dependencies and network utilization to design an appropriate network connectivity between the different VMware Cloud SDDCs.

When deploying VMware Stretched Clusters, VMware Cloud SDDCs can span across multiple geographical regions. A VMware Stretched Cluster can achieve higher levels of availability. An additional benefit of deploying a VMware Stretched Cluster is the ability to provide an extra level of local site protection for Virtual Machines by distributing the placement of Virtual Machines across regions. Deploying a VMware Stretched Cluster will incur additional cost due to cross-regional replication traffic.

Recoverability

Recoverability is measured by two primary metrics - 1. Recovery Point Objective (RPO): the amount of data loss an organization can tolerate and 2. Recovery Time Objective (RTO): the amount of downtime an organization can tolerate.

RPO is a point in time where a VMware Cloud SDDC and an organization's Workloads and applications can be restored. RTO is the amount of time it takes to restore a VMware Cloud SDDC and an organization's Workloads and applications after a failure.

Designing for Recoverability

The VMware Cloud Provider is responsible for the recoverability of the VMware Cloud SDDC management Virtual Machines. However, simply relying on the virtual infrastructure SLA may not be sufficient on meeting the application requirements. An organization must be prepared for individual Virtual Machine failures by designing a proper disaster recovery and backup solution.

For disaster recovery planning, applications should be grouped based on business criticality, RPO/RTO requirements, and application dependencies. The recovery process should prioritize mission-critical applications with lower RTOs. Inter-dependent applications should be recovered together to ensure proper functionality. It is important to choose an appropriate disaster recovery solution based on the application RPO requirements. To meet a lower RTO, an organization can automate the recovery process such as Virtual Machine failover and failback.

Note Tools such as VMware Cloud Disaster Recovery and VMware Site Recovery can provide disaster recovery capabilities for a VMware Cloud SDDC.

In addition, proper monitoring must be in place to ensure that a Virtual Machine or an application failure is detected as soon as possible. The monitoring tool can be configured to alert on specific infrastructure and/or application issues and provides a means to notify the responsible recipients.

Workloads and/or application-level backups are critical to having a comprehensive disaster recovery solution. Backup retention policies and backup job scheduling should be configured to meet an organization's RPO requirements. Appropriate network connectivity with sufficient bandwidth should be provisioned to ensure backup jobs do not affect production traffic. A backup window should be scheduled outside of normal business hours to avoid impact to production workloads. Backups should be stored offsite, in a different location from where the workloads are residing in. An organization should regularly test and validate backups to ensure proper workload recoverability.

A VMware Cloud SDDC can also be used as a disaster recovery destination for an on-premises environment. An organization should have appropriate network connectivity to ensure their users can continue to access their workloads after a disaster has been declared. A plan should be in place for an organization to fail back workloads to their original location once the infrastructure has been restored. An organization should regularly test and failover workloads to and from a VMware Cloud SDDC to validate their disaster recovery procedures.

In the next section, learn about managing and accessing costs for cloud infrastructure providers.

Spend Management

A VMware Cloud environment requires a continuous alignment and collaboration between an organization's internal IT and financial teams. This collaborative approach is especially important for organizations switching from a Capital Expenditure (CapEx) to an Operational Expenditure (OpEx) model, where infrastructure resource has an amortization schedule, and not a strict termination date.

This is a great opportunity for both IT and finance teams to cross-collaborate and evolve their roles and responsibilities to better support the business.

Terminology

This section is written for both a financial and technical audience, certain terms are used which may only be familiar to one group or the other.

Please read these definitions carefully before continuing.

1 Bring You Own License (BYOL): Utilizing a license, not purchased specifically for cloud consumption, when permitted by the terms of service.

2 Enterprise Agreement (EA): Tailored agreement between a company and a service provider for software and services. It may contain or provide a vehicle to create a discount and/or private pricing program.

3 Manufacturer Suggested Retail Price (MSRP): The retail price of a product or service.

4 Microsoft Services Provider License Agreement (SPLA): A mechanism by which service providers and independent software vendors (ISV) can purchase Microsoft software, which includes entitlement on behalf of their customers.

5 Term Commitment: A contractual commitment to a service for a pre-determined period (usually one or three years), usually at a discounted rate.

- Subscription is the term used by VMware Cloud on AWS

6 Terms

7 Subscriptions

8 Subscription Purchase Program Credits (SPP Credits): A unit of value denominated in VMware transacted local currencies that can be purchased by a customer and redeemed for any VMware Subscription Services. (source).

9 Seller of Record (SoR): A vendor which sells a particular product or services to a customer from their inventory. This is distinct from a reseller, for example, a VMware who is authorized to resell SPP credits.

Common Considerations

While each VMware Cloud Infrastructure Provider has its own unique discounting programs, there are potential opportunities to maximize an organization's investment such as: seller of record, payment method, term commitments, and renewal planning.

Note This document is VMware Cloud Infrastructure Provider agnostic, it also does not advise customers on enterprise agreements, which are tailored to an organization's specific needs.

Seller of Record

Whether direct, through a partner or a seller of record, the selected vendor can provide additional saving opportunities.

As part of identifying a potential seller of record, assess whether additional services, software, and support are available for purchase in addition to a VMware Cloud offering. Bundling these additional services along with a VMware Cloud offering may result in potential savings. For more information, please consult a VMware sales or account team.

Payment Method

The next step is to decide the type of payment method, on-demand or a term commitment which is generally one or three years. For example, committing funds up front for services that will be utilized over the term commitment versus on-demand, where organizations are billed on a monthly basis.

Deciding on which payment method will largely depend on how well an organization can forecast its multi-year spend as well as the capability for large up-front investments.

Term Commitments can reduce service costs by up to 50%. Host term commitments are applicable to VMware Cloud based SDDC offerings. Other VMware Cloud services may also be eligible, for example VMware Cloud Disaster Recovery. It is common for organizations to use a combination of hosts term commitments as well as on-demand. Determining the resource requirements for term committed hosts requires both technical and business data to be analyzed.

Note: Paying up-front is generally required for EAs to maximize discounting. To determine the best option for your organization, please contact VMware's Cloud Economics team. Additional resources can be found in the VMware's Cloud Economics team's eBook .

It is important to recognize the benefits and risks of both payment methods. Host term commitments are optimal for long-lived (9+ months) workloads. On-demand host capacity can benefit workloads that are time-bounded with resource utilization spikes, such as end-of-year or end-of-quarter activities or event-based usage such as seasonal demands.

Note VMware Cloud Sizer with LiveOptics data provides an excellent snapshot of an organization's current data, while VMware Aria Operations (SaaS) 's Capacity Analytics can be used to help forecast growth.

Renewal Planning

While it may seem counterintuitive to think about renewal planning during an initial VMware Cloud deployment, it may have an impact on future discounts. When a term commitment expires, the underlying hosts term commitment will revert to an on-demand billing. To prevent this reduction in discounting, monitor for expiring term commitments at least 90 days in advance. The VMware Cloud Notification Gateway can also be used to provide proactive alerting 30 days and 60 days prior to term commitment expiration.

Licensing

Licensing is an important consideration when transitioning to a VMware Cloud-based environment and can be divided into the following categories: VMware infrastructure and Management software, Guest Operating System (OS), and application licensing.

VMware vSphere (vCenter Server and ESXi), vSAN, NSX, and VMware HCX is included in all based VMware Cloud Infrastructure Service offerings. However, licensing levels, versions, and capabilities may vary depending on the VMware Cloud Providers. Organizations with existing VMware infrastructure software licenses can be exchanged for entitlements to VMware Cloud Infrastructure Services. This exchange program is available for VMware Cloud Infrastructure Services that support the VMware Cloud Universal program .

VMware Aria products such as VMware Aria Automation, VMware Aria Operations, VMware Aria Operations for Logs and VMware Skyline are also available as VMware Cloud Services. Organizations with existing VMware Aria licenses can similarly be exchanged for VMware Aria Cloud services using the VMware Aria Universal program.

For guest OS and application licensing, the following options may be available: bring your own licenses (BYOL) or a service provider's and independent software vendor's (ISV) licensing program. For BYOL, verify the terms of the license agreement for applicability in a VMware Cloud based environment. For additional assistance, please contact your VMware sales representative for more information.

Terms of Service

Most VMware Cloud Providers have a Service Level Agreement (SLA). Carefully review the Terms of Service for the selected VMware Cloud Provider(s).

For a link to the SLA documentation, please see the section below.

Spend Management

The roles and responsibilities for spend management may introduce organizational challenges that differs from typical on-premises server and/or IT services procurement.

Finance and IT teams are encouraged to review and potentially adapt their existing approval frameworks to support cloud resource commitments.

Here are some questions to consider for that framework:

- Who are the personas involved in spend management? For example: Finance manager, Cloud Admin, IT director.
- Which individuals or roles decide on when a subscription is needed? Is capacity trending analysis required?
- Who is authorized to commit funds? Does the same individual have permission to provision additional capacity and/or services?
- Does the procurement workflow require any processes for authorization? (Jira, ServiceNow, Remedy, Email/Spreadsheets)?
- Are there tools being used to provide spend management reporting and tracking?

Note VMware Cloud Health or other third party can be used to analyze and manage cloud cost, usage, auditing, and governance.

VMware Cloud on AWS

This section briefly describes the Seller of Record (SoR), available payment methods, licensing, billing, and terms of service.

- SoR - There are two available SoRs: VMware and AWS
- Payment Method - VMware provides SPP credits and AWS provides Enterprise Agreements.
  - The Spp credits is provided as a as a currency with discounting based overall VMware Cloud services and subscription spending. For details, go to https://my.vmware.com/web/vmware/spp-landing.
  - AWS provides Enterprise Agreements (EDP) is provided with discounting based on aggregate VMware Cloud on AWS and native AWS services spending. For details, go to https://aws.amazon.com/pricing/enterprise/.
- Licensing -
  - SPLA
  - Allows BYOL where permitted
  - Custom core counts are supported for all instance types
    - Feature brief
    - AWS Pricing
- Billing - Each SoR has its own billing portal, which can be one of the following:
  - VMware Console
  - AWS Console
- Terms of Service - The SLA services across both providers may be found here .

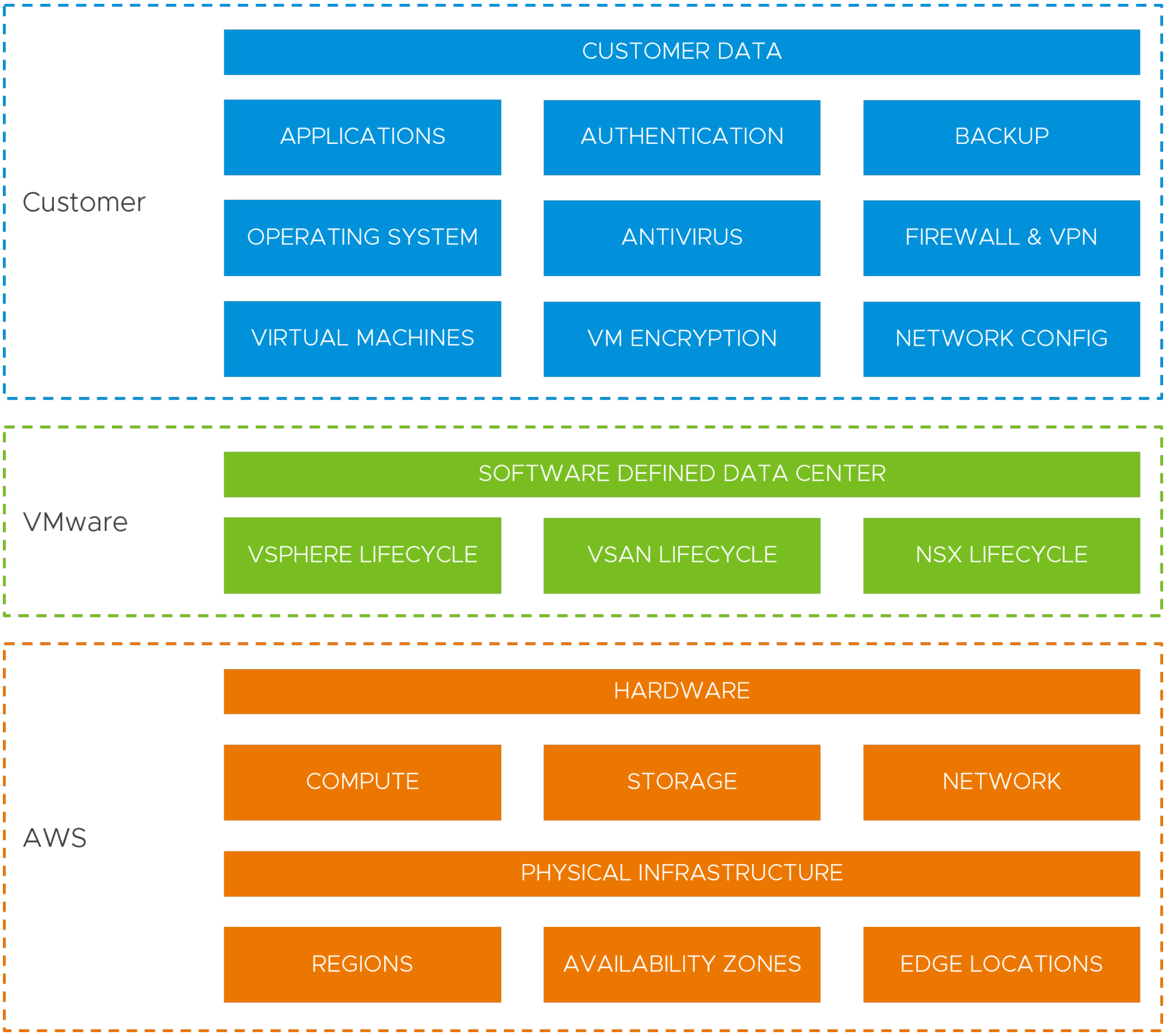
Shared Responsibility Model

A shared responsibility model is common among the different VMware Cloud Providers, which defines distinct roles and responsibilities between the VMware Cloud Infrastructure Services provider and an organization consuming the service.

Disclaimer: The intent of this document is to provide guidance and best practices for VMware Cloud Providers regarding the shared responsibilities of the service.

VMware Cloud on AWS

VMware Cloud on AWS implements a shared responsibility model that defines distinct roles and responsibilities for the three parties involved in the offering: Customer, VMware, and Amazon Web Services.



Responsibilities

Responsibilities are shared and the customer, VMware, and AWS are all responsible for the security in the cloud.

Customer Responsibility: Security in the Cloud

Customers are responsible for the deployment and ongoing configuration of their SDDC, virtual machines, and data that reside therein.

In addition to determining the network firewall and VPN configuration, customers are responsible for managing virtual machines (including in guest security and encryption) and using VMware Cloud on AWS User Roles and Permissions along with vCenter Roles and Permissions to apply the appropriate controls for users.

VMware Responsibility: Security of the Cloud

VMware is responsible for protecting the software and systems that make up the VMware Cloud on AWS service.

This software infrastructure is composed of the compute, storage, and networking software comprising the SDDC, along with the service consoles used to provision VMware Cloud on AWS.

AWS Responsibility: Security of the Infrastructure

AWS is responsible for the physical facilities, physical security, infrastructure, and hardware underlying the entire service.

Details on the shared responsibility model employed by VMware Cloud on AWS can be found in the table below. You can see that a great deal of low-level operational work is handled by the VMware Cloud on AWS Site Reliability Engineering team leaving the customer to focus on managing their workloads.

Shared Responsibility Matrix

For a detailed description of the roles and responsibilities for VMware Cloud on AWS, please refer to the Service Description .

Entity	Responsibility/Activity
Customer	■ Deploying Software Defined Data Centers (SDDCs) ■ Host Type ■ Host Count ■ Connected AWS Account ■ Configuring SDDC Network & Security (NSX) ■ Management Gateway Firewall ■ Management Gateway IPsec VPN ■ Compute Gateway Firewall ■ Compute Gateway IPSec VPN ■ Compute Gateway NAT ■ Public IP Addresses ■ Network Segments ■ Distributed Firewall ■ Deploying Virtual Machines ■ Installing Operating Systems ■ Patching Operating Systems ■ Installing Antivirus Software ■ Installing Backup Software ■ Installing Configuration Management Software ■ Migrating Virtual Machines ■ Live vMotion ■ Cold Migration ■ Content Library Sync ■ Managing Virtual Machines ■ Installing software ■ Implementing backup solution ■ Implementing Antivirus solution
VMware	■ SDDC Lifecyle ■ ESXi patch and upgrade ■ vCenter Server patch and upgrade ■ NSX patch and upgrade ■ vSAN patch and upgrade ■ SDDC Backup/Restore ■ Backup and Restore vCenter Server ■ Backup and Restore NSX Manager ■ SDDC Health ■ Replace failed hosts ■ Add hosts to maintain adequate “slack space” ■ SDDC Provisioning ■ Operate vmc.vmware.com 24x7x365 ■ Manage “Shadow” VPC holding customer SDDC
Amazon Web Services	■ Physical Infrastructure ■ AWS Regions ■ AWS Availability Zones

Entity	Responsibility/Activity
	■ Compute / Network / Storage ■ Rack and Power Bare Metal Hosts ■ Rack and Power Network Equipment

References

Go to the AWS site to learn how to meet your security and compliance goals using AWS infrastructure and services.

Amazon Web Services: Overview of Security Processes

In the next section, learn about the different considerations for managing infrastructure and application services.

Infrastructure and Application Observability and Services

When planning a cloud migration, it is critical to understand the various infrastructure services that are in use in the current environment and how those services will be consumed and/or rearchitected for use in the cloud.

Infrastructure and Application Assessment and Native Services

Infrastructure services are those critical functions on which all other workloads depend upon. Some examples include DNS, DHCP, NTP, authentication and authorization services, logging receivers, and monitoring solutions.

Planning for Infrastructure Services in the Cloud

Planning for infrastructure services is dependent on the location of the running workloads, with a recommendation of placing services in close proximity.

In a hybrid cloud environment, where a mix of on-premises and VMware Cloud based workloads exist, organizations may run some or all infrastructure services from their on-premises datacenter and extend these services to VMware Cloud-based workloads.

For organizations planning on exclusively running within a VMware Cloud SDDC, where an on-premises datacenter is no longer available, cloud-native infrastructure services is needed primarily to facilitate these core services. The use of multi-cloud can also add additional challenges, such as using multiple similar services across various infrastructure service providers or simply extending infrastructure services from an on-premises datacenter to a VMware Cloud-based environment. Security and access policies can also play a large role in the architecture of infrastructure services; hence it is essential to include your security team in the planning phase.

Retain Infrastructure Services Locally

There are good reasons to keep some infrastructure services running in an on-premises environment. If there are existing DDI (DNS, DHCP, and IPAM) solutions, organizations may want to continue to leverage their existing investment. There may also be security and compliance requirements that mandate infrastructure services continue to run in an on-premises location. Workload placement and data gravity can also affect the placement of infrastructure services, not just in the near term but also in the desired future state.

For common networking services such as DHCP and DNS, organizations can take advantage of VMware NSX networking capabilities included in a VMware Cloud SDDC. VMware NSX can forward both DHCP requests and DNS queries from a VMware Cloud SDDC to an organization's data center, allowing these services to remain on-premises. Similarly, log data can be forwarded to an on-premises logging receiver environment.

An important consideration before configuring infrastructure services is the transmission of data. Note that the amount of data being transferred from a VMware Cloud SDDC is the cost of egress traffic. Costs can vary per region as well as Infrastructure Services Provider. Without proper planning, egress traffic fees can be unpredictable and contribute to an organization's bill. It is important to understand the utilization for existing infrastructure services and accurately forecast egress utilization and plan accordingly.

When retaining infrastructure services in an on-premises environment, it Is essential to review any new or additional requirements before providing these services to an organization's VMware Cloud SDDC. Testing, documentation, and a security review are highly recommended before exposing a service to VMware Cloud based workloads.

Migrate to Cloud-Native Services

As an organization transitions to a VMware Cloud-based SDDC, they should assess and evaluate the benefits of cloud-native infrastructure service options including their SLAs. An assessment must be made of an organization's on-premises infrastructure services based on criteria such as manageability, availability, and total cost of ownership (TCO) which outweighs the benefits of a managed cloud-native service.

Automation and Application Programing Interfaces (APIs) are generally first-class citizens for cloud-native infrastructure services. This enables an organization to leverage modern automation and "infrastructure as code" tools to create and request new cloud resources in a much shorter period of time. This can possibly take weeks or months to build in traditional data centers. The combination of VMware Cloud and the ability to automate cloud-native infrastructure services can deliver a true Software-Defined Datacenter.

Another benefit to cloud-native infrastructure services is the level of availability that is offered, health monitoring, and scalability as a managed service. For undifferentiated infrastructure services, organizations can leverage cloud-native infrastructure services, which are priced based on consumption. It is important to have understand the utilization of existing infrastructure services before migrating them to a cloud-native service. Most infrastructure services providers offer tools to help predict the estimated cost of consuming their services. To proactively monitor and prevent unexpected costs, billing alerts and notifications should be configured.

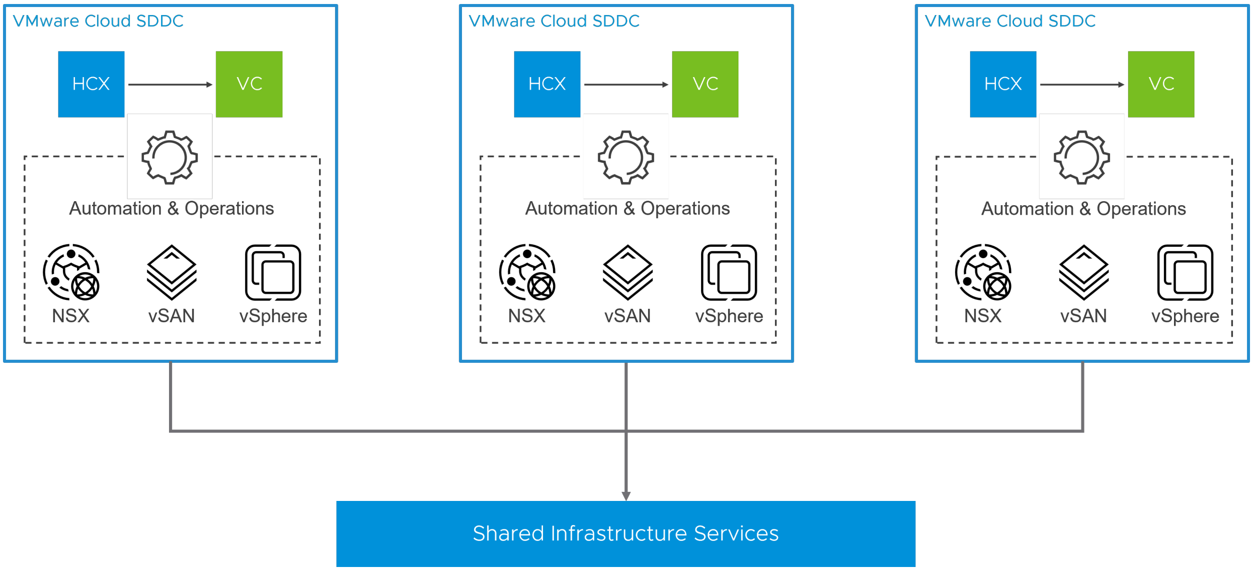
Multi-Cloud can also present a unique challenge when determining an organization's strategy for consuming infrastructure services. For example, if an organization plans to extend their existing on-premises infrastructure services to multiple infrastructure service providers, different types of connectivity must be configured and managed, which can bring additional complexities from an operational standpoint. In contrast, organizations may consume cloud-native infrastructure services from individual infrastructure service providers, which can also bring its own complexity as each solution will require unique skillsets to configure and manage.

Centralized/Shared Infrastructure Services

An alternative design for organizations that have a need to manage and control access to infrastructure services from a centralized location can be to leverage a shared infrastructure services model.

This implementation will result in one or more VMware Cloud-based SDDC terminating into a single and centralized infrastructure service endpoint. The configuration of network and security policies can now easily be managed and operated with all ingress and egress traffic terminating to this endpoint. This design can be applicable to both cloud-native services as well as infrastructure services running within an on-premises data center.

Figure 1-2. Centralized/Shared Infrastructure Services



Assessing Existing Infrastructure Services

An organization should first identify and assess all on-premises services that can be classified as an infrastructure service.

During the assessment, an analysis should be performed that includes the health, scale and configuration as an example to determine if the existing infrastructure service can support workloads running in a VMware Cloud SDDC. If an existing infrastructure service is deemed insufficient, then an organization should strongly consider rearchitecting the service or replacing it with a cloud-native service.

Upon completing the assessment, an organization can determine whether a given service should be replaced with a cloud-native service, assuming a feasible alternative exists. Common services such as DHCP and DNS are generally available in all infrastructure service providers, other services must be evaluated based on its utilization, criticality, and cost as an example to determine the best available option.

There are many tools and methodologies for cataloging and assessing your existing environment. Internal documentation can provide an initial understanding of how existing infrastructure services are configured and managed. Monitoring can be another source of information, especially if they are monitoring the utilization and health of specific infrastructure services (e.g., the number of DNS queries per second). With the availability of metrics, an organization can appropriately forecast infrastructure service utilization and cost.

NetFlow data can also provide insights into the different protocols running within an organization’s network. Example traffic types can include DNS, DHCP, NTP, Syslog, authentication, and client/server communication. NetFlow data can also be used to understand application and service dependency mapping, which will assist in migrating workloads to a VMware Cloud SDDC.

Note VMware Aria Operations for Networks can be used to analyze network traffic to help understand the different types of applications and/or services running within an organizations network.

Securing Infrastructure Services

Each infrastructure service has its own security and best practices. These should continue to be followed when migrating to VMware Cloud, but also in evaluating any update and new security capabilities.

When operating in a hybrid cloud model, it is important to assess existing firewall rules and access control lists to determine if additional or new configurations are required for connectivity to and/from an on-premises environment.

Cloud-native services also have their own security and best practices. Infrastructure service providers also offer network-based access-control lists (ACLs) and granular role-based access control (RBAC) for their services and in some cases, the ability to control access down to an individual API call. Most infrastructure services will have predefined roles, which typically map to different personas within an organization. Organizations should also assess the different personas that will be managing the infrastructure services as well as the workloads that will be consuming these services. Using this information, fine grain access control and least privilege accounts should be implemented. Audit logs can be configured to track all changes and access to infrastructure services for both compliance and troubleshooting purposes.

Authentication and Directory Services

Each infrastructure service provider will have a solution for managing authentication, users, and permissions which includes the ability to federate authentication with any SAML2-based identify provider.

For organizations that have a need for identify federation, additional planning and technical analysis is required before selecting a specific infrastructure service provider.

Similar to other infrastructure services, an organization must decide if they want to operate their own directory service or consume it as a managed service. Depending on the requirements for an organization's directory service, which may include security and compliance regulations, a managed directory service could be an option. Organizations subject to PCI, HIPAA, or other regulations will have to take this into account. Many infrastructure service providers provide yearly audits to attest to PCI or HIPAA compliance, among a range of other standards and compliance frameworks, organizations should verify the selected infrastructure service provider has all required certifications.

As organizations makes their transition to a VMware Cloud based environment, there is an opportunity to re-evaluate their strategy for infrastructure services. For undifferentiated infrastructure services, organizations can simplify their infrastructure management and operations by considering cloud-native services.

Build Pillar

2

This chapter includes the following topics:

- Deploying VMware Cloud on AWS
- Identity and Access Management Services for VMware Cloud on AWS
- Infrastructure Services for VMware Cloud on AWS
- Automation of Infrastructure, Workload, and Security Services for VMware Cloud on AWS
- Backup and Disaster Recovery for VMware Cloud on AWS
- Best Practices for Sustainability & Carbon Reduction for VMware Cloud on AWS

Deploying VMware Cloud on AWS

There are many considerations and high-level logical decisions to be made before you can deploy the VMware Cloud on AWS solution. You should be familiar with the decisions that must be made from the information provided in the planning stage, as they will affect the choices you must make.

The following highlights some of the important actions that should be considered prior to deployment.

1. Have a complete profile entered in my.vmware.com for the Fund User/Owner account.
2. Identify or create a customer-owned AWS account. This is required as a means of providing the SDDC with access to AWS services.
3. Review the CloudFormation Template used for account linking. This may be required depending on the security policies of your organization. Details of this template may be found in the official user guide.
4. Identify the correct AWS region for SDDC deployment.
5. Identify or create a VPC within above region which is to be used for SDDC cross-linking.
6. Identify or create a dedicated subnet in the desired availability zone within the VPC. This is for SDDC Cross-Account ENIs. Ths needs to be a dedicated /26 subnet as a minimum.

7. Identify SDDC Management IP subnet. A /23 subnet scales to 27 hosts, whereas a /20 subnet scales to 251 hosts. The SDDC Management subnet is exclusively for management and may not be carved up or otherwise used by compute workloads.
8. Identify SDDC Compute network IP address range(s). This is for network segments in the compute network. Network ranges must be at a minimum of a /30 subnet or a maximum /22 per. This is not required to deploy the SDDC but is required in order to deploy workloads.
9. Identify a strategy for integrating custom DNS servers with the SDDC (public DNS is used by default). This step is needed if your workloads need name resolution for IP address space which is private to your organization.
10. Identify a strategy for connectivity to the SDDC (IPSec VPN, Direct Connect, etc.).
11. Determine minimum network security policies to permit administrative access to the SDDC.

Deployment of a VMware Cloud on AWS SDDC may be performed one of two ways:

1. **Traditional deployment via the Cloud Service Portal interface** - this is the most commonly used option for initial service onboarding as the user interface provides input choices with examples, recommendations, and allows you to view the options when there is a specific selection to be made. This can be helpful when familiarity with the service boundaries and requirements are needed.
2. **API Deployment** - For customers that deploy VMware Cloud on AWS services on a regular and/or high scale pace, VMware Cloud on AWS SDDC's and associated functions can be called via API's. This allows scripted deployment and configuration of SDDC's rapidly without user interface interaction and can be a desired method when repeatedly deploying SDDC's or making common configuration changes quickly.

Identity and Access Management Services for VMware Cloud on AWS

VMware Cloud on AWS service has two identity and access methods that work together to provide access to the service, the VMware Cloud Services portal and private SDDC authentication.

Identity & Access Management

The Cloud Services Portal is a public website that can be accessed directly from the vmware.com main webpage. Private authentication to the SDDC itself is needed once the SDDC has been deployed.

Cloud Services Portal Roles

The VMware Cloud on AWS service utilizes the vmware.com Cloud Service Portal for identity, billing, and access service commonly referred to as the Cloud Service Portal. The vmware.com accounts are associated when a subscription is activated. The account will be tied to a specific email addresses provided to VMware during service activation.

Cloud Service Portal (CSP) roles available by default are Org Owner, Org Member, and Support User. Description and use of the CSP Roles is documented here: https://kb.vmware.com/s/article/2151069

The first account associated with a subscription will be an Org Owner. This first account is used to establish additional Org Owners, Org Members, or Support users. The CSP Portal allows an Org Owner to selectively assign roles to specific services in the VMware Cloud Portal, such as the VMware Cloud on AWS service. This method allows organizations to grant only the minimal permissions needed to run the service alone.

Granular Permissions for the VMware Cloud on AWS Service

As identified in the Plan pillar of the Well Architected Framework, organizations should have all stakeholders, roles, and functions defined from the planning exercise. There should be a documented permissions model to implement once the SDDC is provisioned.

For example, if an organization has designated a network engineer or a group to manage all SDDC networking components, they may implement a CSP account that has the Org Member Role, permissions to the VMware Cloud on AWS service, and assign the VMware Cloud NSX Cloud Admin role.

Accounts can be easily created, modified, or removed either manually using the user interface or automated through APIs.

SDDC Permissions

The second layer of identity and access management is the traditional SDDC roles and permissions that exist within a vSphere environment, commonly referred to as vCenter permissions or private authentication.

This management method is familiar to many organizations that already use vSphere to manage their virtual environments. This management and permissions structure is a common framework throughout all of VMware's cloud platforms and enables organizations to manage their virtual infrastructure in a consistent and familiar way. Managing and utilizing this permissions model is covered under the existing vSphere platform documentation .

VMware Cloud on AWS operates on a shared permissions model. This means that VMware maintains full admin rights to the SDDC infrastructure and the customer is provided with a restricted administrator role. This gives customers privileges to manage their workloads but not complete access to the entire SDDC. The roles and permissions defined by VMware Cloud on AWS are documented here: VMC on AWS product documentation.

Within vCenter, identity sources and policies may be customized for the vmc.local domain. Many organizations choose to use an identity source that already holds their user base, logical business groups, and has their functional roles described and documented. By customizing an identity source, customers may streamline their processes, documentation, and business functions.

Infrastructure Services for VMware Cloud on AWS

When a VMware SDDC is deployed in the cloud, default configurations related to infrastructure services are implemented at a rudimentary level to support the VMware components that make up the SDDC.

Introduction

There is no requirement to establish infrastructure services in the cloud before deploying a VMware SDDC because the Service Provider has accounted for these services as part of the initial deployment.

DNS

Upon SDDC creation, public DNS servers are configured for both Management and Compute Gateway zones.

The default DNS servers and gateway zones can be modified to suit an organizations DNS strategy. Read about it on VMware Cloud on AWS DNS Strategies.

DHCP

Network segments within the compute network may be configured to provide basic DHCP services. If your design requires a more advanced DHCP feature set, then the SDDC may be configured to provide DHCP relay services.

Authentication Services

VMware Cloud on AWS supports either a standalone service with separate authentication methods or an integrated authentication and access method with both the VMware Cloud Portal and the VMware Cloud on AWS SDDC environment.

Organizations can configure the VMware Cloud Portal authentication to leverage an organization's primary identity authentication method and grant access to VMware Cloud services. Additional information can be accessed at Enterprise Federation information.

In addition, an SDDC can be configured to use an organization's existing identity management solution similar to how vSphere on premise is connected to an authentication source. This is accomplished through the SDDC's CloudAdmin default administration account. Read detailed information at Configuring Hybrid-linked mode.

Automation of Infrastructure, Workload, and Security Services for VMware Cloud on AWS

Automation for VMware Cloud on AWS

Operations on VMware Cloud on AWS infrastructure and resources can be automated by leveraging the VMware Cloud on AWS APIs.

The VMware Cloud on AWS REST APIs are built on standard web protocols, HTTP and HTTPS, and network ports, 80 and 443 respectively.

To use VMware Cloud on AWS APIs, an API token must be generated from a VMware Cloud Services account to programmatically call the APIs. For more details on authentication and authorization of VMware Cloud on AWS API programming, please refer to the official VMware developer documentation at https://developer.vmware.com/apis/vmc/latest/.

One of the easiest ways to automate the management of the VMware Cloud on AWS Software-Defined Data Center (SDDC) infrastructure is to use PowerCLI. PowerCLI is a command-line interface tool based on PowerShell, specifically built to interact with VMware products. The PowerCLI cmdlets use APIs under the covers to perform desired management and operational activities. Therefore, any operations you can complete with VMware Cloud on AWS APIs can generally also be performed using PowerCLI cmdlets.

To automate the deployment and management of workloads that reside in a VMware Cloud on AWS SDDC, organizations can also leverage VMware Aria Automation. The VMware Cloud on AWS environment is added as a cloud account in VMware Aria Automation. Then with appropriate VMware Aria Automation configurations, workloads can be deployed to the VMware Cloud on AWS SDDC from VMware Aria Automation templates. Additionally, VMware Aria Automation can perform day 2 actions to the workloads, such as adding storage or creating snapshots. Security components, such as NSX-T distributed firewall rules and security groups, can be created using VMware Aria Automation templates as well.

Leveraging automation can help organizations consume VMware Cloud on AWS at an even faster rate and accelerate their cloud adoption journey.

Backup and Disaster Recovery for VMware Cloud on AWS

Having a disaster recovery plan for the VMware Cloud on AWS SDDC is important if there are critical, production workloads running in a VMware Cloud on AWS SDDC.

Backup and Disaster Recovery

Disaster recovery is a crucial part of a data center design. VMware Cloud on AWS provides flexibility and agility for organizations to quickly implement a working disaster recovery solution and protect their on-premises or cloud environments.

There are two main solutions that organizations can leverage for disaster recovery with VMware Cloud on AWS: VMware Site Recovery Manager (SRM) and VMware Cloud Disaster Recovery (VCDR).

VMware Site Recovery Manager (SRM)

Organizations can use VMware Site Recovery Manager (SRM) along with vSphere Replication. This on-demand Disaster Recovery as a Service (DRaaS) solution can be enabled as an add-on from the VMware Cloud on AWS UI.

Once the add-on service is enabled, a virtual appliance must be deployed in the on-premises environment for the sites to be paired and replication of virtual machines to begin. With this solution, organizations can protect their on-premises environment and utilize VMware Cloud on AWS as the disaster recovery site instead of having to invest in a separate, physical data center for disaster recovery.

VMware Cloud Disaster Recovery (VCDR)

VMware Cloud Disaster Recovery (VCDR) is an on-demand disaster recovery service delivered as a SaaS solution. A virtual appliance (DRaaS connector) is deployed in the protected site to replicate data to a cloud-based Scale-Out Cloud File System (SCFS). When a disaster occurs, a VMware Cloud on AWS SDDC is deployed and virtual machines are recovered to the SDDC.

One of the main advantages to using this solution is that organizations do not have to create and pay for VMware Cloud on AWS SDDCs upfront. This, in turn, reduces overall cost for organizations to operate their data centers without having to worry about not being able to recover from an unexpected scenario.

One of the simplest ways to enhance resiliency of the VMware Cloud on AWS SDDC is to use VMware Stretched Clusters. VMware Stretched Clusters span across two different AWS availability zones within an AWS region. Therefore, if there is an outage in an AWS availability zone, vSphere High Availability will automatically move the workloads to the other AWS availability zone. If there are more demanding recoverability requirements, another site may be needed to serve as a disaster recovery site for the VMware Cloud on AWS SDDC. The site can be an on-premises environment or another VMware Cloud on AWS SDDC. VMware Site Recovery can be enabled and leveraged to orchestrate the disaster recovery procedures between the sites and regions.

Best Practices for Sustainability & Carbon Reduction for VMware Cloud on AWS

VMware Cloud on AWS can help organizations quickly adopt environmental sustainability and carbon reduction best practices in their operations.

Sustainability Practices

By moving workloads to a VMware Cloud on AWS Software-Defined Data Center (SDDC), organizations can reduce the net environmental impact of data centers without making a dedicated investment for sustainability and carbon reduction.

VMware and AWS take on this responsibility by continuously working to make data centers more energy efficient. VMware implements sustainable practices to develop and support the VMware Software-Defined Data Center, such as using 100% renewable energy and supporting low-carbon sustainable development projects. AWS is also committed to sustainability in the cloud as they continue their path to powering their operations with 100% renewable energy by 2025. AWS is continuously making their data centers more energy-efficient with initiatives such as developing efficient water-use strategy and accelerating sustainability research and innovation.

Not only are the data centers supporting VMware Cloud on AWS energy efficient, but they also provide flexibility in VMware Cloud infrastructure. Organizations can scale their VMware Cloud on AWS environment up and down as necessary to meet the fluctuating workload demands and availability requirements. This can be automated as well by leveraging Elastic Distributed Resource Scheduler (EDRS). EDRS monitors cluster resource utilization and automatically scales a cluster to allow for randomness in the cluster utilization while maintaining desired CPU, memory, and storage performance. By using VMware Cloud on AWS infrastructure, organizations are supporting environmental sustainability and carbon reduction by eliminating waste of resources in data centers.

In addition to eliminating actual physical resources waste, organizations can further reduce waste by monitoring deployed services and right-sizing workloads in their VMware Cloud on AWS SDDC. By using VMware Cloud on AWS, organizations no longer need to spend time on taking care of the physical or logical infrastructure of their data centers. Instead, organizations can focus on improving their workloads and making them more energy efficient.

VMware Cloud on AWS enables organizations to accelerate not just their cloud migration and adoption, but also their contribution to environmental sustainability and carbon reduction in their operations.

Modernize Pillar

3

This chapter includes the following topics:

- Modernize Introduction
- Modernization Through Rehosting
- Modernization Through Re-platforming
- Modernization Through Refactoring
- Modernization Operations

Modernize Introduction

Application modernization can be achieved through various methods, including using new tools or processes, or even migrating the application in its current state. However, before embarking on an application modernization journey, organizations and teams must ensure certain prerequisites are met and understood.

Introduction

Most applications can be modernized in some way due to continuous technolgogy evolution. It could simply be improving the deployment process or it could be a larger effort to redesign the entire software architecture.

There are prerequisite activities that organizations should complete prior to completing actual application modernization practices.

- Assess the infrastructure and application environments.
- Complete an accurate inventory of their workloads. Ensure this includes how they are related and where applications are dependent on another.

Modernizing your data center can be broken down into discrete decisions about the applications that power your organization. With the application and workload inventory properly cataloged and dependencies identified, your organization now must decide the speed at which you will address your modernization efforts. The time needed does not always equal the allotted time to complete a modernization project. These business constraints must be identified and will influence which method of modernization you can take.

Let's revisit the different application modernization methods:

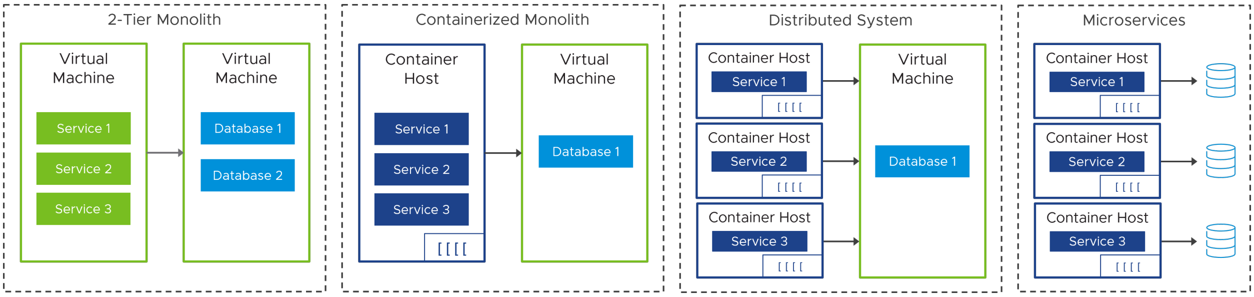
- **Retain** means leaving workloads in a private cloud environment.
- **Retire** means decommissioning workloads and/or converting to SaaS.
- **Rehost / Migrate** involves either changing the hypervisor. (e.g., migrate applications from one virtualized environment to another) which is known as Rehost or moving an application without changing the underlying hypervisor or application at a source code level (e.g., migrate VMs from one virtualized environment to another without requiring changes) which is known as Relocate.
- **Refactor / Build** involves changing the application at the source code level. Typically, applications are rewritten to take advantage of cloud microservices architecture and incorporate new services such as IoT, machine learning, and others.
- **Replatform** involves changing the operating system, such as going from Windows to Linux, modifying the application middleware, such as going from a self-managed database to a cloud provider-managed database or from a virtual machine to a container image.

The Modernization Continuum

Modernization does not have to occur as an abrupt change to an application but instead can be a gradual process that occurs over a more extended period of time.

In this way, application modernization can be considered a continuum with no end state. The idea of a "modernized app" means that it uses modern infrastructure, tools, or processes compared to how it was initially built. In this way, there is no end to this modernization process, but it is instead repeated on applications until the business decides it's not worth the cost to continue.

Applications may undergo a series of modernization stages before reaching their intended architecture. Businesses may decide to improve the application's underlying infrastructure by moving to new hardware or a cloud platform. Companies might improve an application's functionality by re-platforming or refactoring the code, or businesses might begin by migrating to the cloud and then re-platforming. Each application in your portfolio might go through different paths in this modernization continuum.



This continuum allows application teams to lower risks by making smaller changes to the application instead of an entire code rewrite. At some point in the continuum, after sufficiently modifying an application, there is a diminishing return on the amount of value received from updating that app further. Because of this, the application should be continuously assessed after each stage to determine whether it should continue to be optimized. You should continue modernizing an application as long as it provides a positive return on investment to the business or the opportunity costs of updating the app are too high.

Future Application State

A vital part of an application modernization strategy will be defining how each application is expected to operate and perform in its future state.

It is crucial to align application function, performance, availability, and recoverability characteristics in the cloud to its business objectives. An application's business objectives or requirements will not change with its deployment in a cloud platform.

How you meet these applications' business objectives will vary with your organization's cloud use case.

For applications being moved or operating in hybrid mode, its future state may still be subject to existing organizational parameters.

For example, operating in the cloud will remove some operational tasks like infrastructure upgrades that staff previously had to spend a lot of time managing. However, a migrated application will still likely need application updates and upgrades. These activities will still need to be completed, scheduled, tested, and planned for.

As applications are moved to the cloud, some common related enterprise functions and responsibilities will require review and re-evaluation. The following are some functions that should be considered:

- Application-level patching.
- Application-level upgrades.
- Whether the authentication services will be local, on prem, or in the cloud.
- Whether backup services will be required and used.
- Notification and reporting services needed in the cloud that were previously addressed through redundant hardware on-premises.
- What network services will be needed in the cloud that were previously provided on site to applications. For example, load balancing services needed vs the load balancing solution deployed on site.
- Recovery plans to another service or another geographic region.
- What security services, appliances, processes, and methods will replace security services leveraged on prem. For example, firewall functions.

With the exception of data center patching and upgrading, your organization may have these and many more functions or dependencies of applications that need to be well defined to meet business requirements in the cloud. This is in addition to documenting an application's performance in the cloud to understand if it is operating effectively. It is important to understand and document how the application is expected to be managed alongside all its related dependencies in order to fully modernize an application for the cloud.

For existing workloads, these environmental factors govern a lot of what the future state of a migrated or hybrid application may look like. Together, these functions, characteristics, and requirements will make up the entire state of an application in the cloud.

Use Known Patterns

As part of the modernization effort, you will begin to find patterns that work for your organization. Many applications will follow similar architectural designs, and it is wise to re-use these known patterns.

It is challenging to gain operational benefits if your applications are vastly different. Grouping applications into application patterns can provide some known solutions that can be recreated many times. Using a set of reproducible architectural patterns offers several benefits to the business.

- Shared institutional knowledge – Any time a new application pattern is created, it requires further understanding of maintaining the application. Using familiar patterns across the organization provides a shared language between teams. This shared institutional knowledge also allows developers to work across teams or projects with a working knowledge of the application patterns in use.
- Shared Operational Processes – Operations teams are often responsible for many applications, and modern application patterns often require new strategies to operate them. A standard set of operational processes can reduce the burden on shared operations teams who must monitor performance, security events, and outages.
- Licensing – There are financial benefits to standardizing on specific patterns. Software vendors often provide discounts for bulk purchases. Standardizing on specific software stacks may limit the flexibility of new application decisions but provide a reduced price for software if shared across the enterprise.
- Optimized Developer Time – Developers who have previously created applications in a specific pattern can re-use what they're learned in subsequent applications. A smaller set of known architecture patterns or deployment solutions limits the number of tools a developer needs working knowledge of. It also reduces the number of decisions that need to be made when starting a new application, leading to a faster start.

Consider Portability

When rebuilding applications for a cloud environment, you're presented with many options for architecting those apps. When building these applications, consider the entire lifecycle of these applications. Will the application always live in the cloud it was initially deployed in? Will there be a need to move it to another cloud or duplicate the deployment in another cloud?

These considerations may determine what form factor the application's final form will be built as. For example, applications depending on native public cloud services may be pinned to that cloud unless there is an equivalent service in another. For these reasons, customers often choose to stick with a VMware virtual machine form factor that is compatible with any VMware-based infrastructure or a container form factor that can be run on any Kubernetes platform.

Start Fast and Learn

The first applications that are modernized will almost always present unforeseen challenges. You may find operational challenges with the deployment of the apps, recognize previously unidentified dependencies, or find that new services don't work as expected. These challenges experienced are unlikely to be identified in a planning phase even with an abundance of analysis.

Since a planning phase won't identify all issues with a modernization project, it's essential to get started and let the work inform and tune the modernization process. Issues with creating your first modernized applications should not be seen as failures but rather an expected part of your modernization process, which is used to improve the process further.

Modernization Through Rehosting

Keeping your infrastructure consistent between your data center and public cloud allows the flexibility to seamlessly migrate to any cloud provider, giving you agility in changing times.

Rehost/Migrate

For many applications and businesses, it is more practical to rehost/migrate an application in its current state to the VMware Cloud than to modify it as it currently runs.

Time may become the most constrained resource when considering which method to migrate an app when combined with the current staff's daily pressure of running infrastructure. This is when it makes the most sense to move an application and its dependencies entirely to an instance of VMware Cloud without modifying its functionality. By doing so, you will be able to continue to provide the application functionality to the business as it currently is with minimal change while gaining the efficiencies of VMware Cloud.

The VMware Cloud provides a common abstraction layer across different infrastructure providers and the tools to migrate workloads from one location to another.

Network Strategy

Prior to moving workloads, a proper networking strategy for all workloads both on-premises and in the cloud needs to be identified and determined.

The critical networking decisions related to migrating workloads consist of the overall IP strategy for workloads that are being migrated and the actual network traffic flows that impact the associated workloads and their downstream consumers.

Determine an IP Addressing Strategy

A critical decision to make early during planning, is an IP addressing strategy for workloads to be moved. This comes down to a decision between two possible scenarios.

Scenario 1: Workloads change their IP addresses

In the case where workloads will change their IP addresses, there are certain additional planning activities to be performed.

- A new IP addressing scheme must be determined for the migrated workloads.
- Application owners must be made aware of the new IP addressing schemes and must be prepared to update their applications accordingly.
- System administrators must make plans for changing workload IP addresses post-migration.
- Updates to DNS, Firewalls, Load Balancers, Certificates, and other infrastructure services must be coordinated to reflect changes in IP assignments.

Scenario 2: Workloads keep their existing IP addresses

A common goal of a migration project is to minimize disruption to business. However, IP address changes tend to be very disruptive. For this reason, most migration projects will require that workloads keep their IP addresses post-migration.

In order to accomplish this, there are two choices:

- Migrate entire subnets worth of workloads at a time.
- Utilize a layer-2 network extension during the migration.

Since workloads that house applications aren't always arranged neatly within the bounds of a given subnet, it is often impractical to plan migrations around subnet boundaries. For this reason, the most common strategy is to utilize a Layer 2 network extension during a migration. Network extensions provide a great deal of flexibility in how a migration is performed and allow entire applications to be migrated regardless of the layout of the underlying network addressing scheme.

Transition Plan

Each application's business value will determine whether an organization continues to evaluate its architecture for future functionality and investment in its modernization continuum as previously discussed.

If an application is key to the organization and its business, the organization may consider different styles of deployment even after migration. This is normal for an application to be continually re-evaluated for best performance and functionality during its life cycle.

Maintain

After a migration, an application may very well be in the perfect state for its business objective and provide full functionality and performance. As such, the application may not warrant further review or architecture investment.

Often, an organization may simply achieve their cloud strategy objectives by moving the application and addressing the surrounding business dependencies like backup and load balancing services.

Hybrid Transition

Applications can often be moved in their entirety to the cloud to achieve business objectives like moving to an OpEx (Operating Expenditure) model and speed cloud adoption.

An application can further warrant investment in different architectures once in the cloud to further gain application functionality like application-level availability and redundancy. A large application with several functions in its architecture will often require careful planning and transition plans to minimize disruptions. The application components could be separated and transitioned one by one. As the transition occurs over time, an application will be in a state of transition.

For other applications, a portion of the service may be transitioned while maintaining the remainder in its current form. For example, if an application has a database component and an organization has plans to migrate the database component to a different architecture where they are taking advantage of native cloud database services or licensing benefits. This split application architecture may be the final design to minimize service disruptions, maximize cloud benefits, and maximize database features.

Thus, an applications actual modernization journey may continue in a progressive fashion in the modernization continuum.

Automate Workload Migration

Moving workloads to the cloud with VMware can be significantly less impactful for businesses since VMware is consistent through the different platforms and you can utilize VMware tools to facilitate the migration of existing workloads.

By leveraging VMware tools, your migration methods can be streamlined with thorough planning in migration waves.

Migration Waves and Events

A migration wave is comprised of a group of workloads that will be migrated concurrently in a single migration event. There can be one or multiple miration waves in a single migration event. Migration events also contain other steps such as table-top activities, go/no-go checks, infrastructure updates (DNS, load balancers, etc), application teams testing, and operations hand-over.

The general recommendations for planning migration waves are as follows:

- Plan migration waves around applications whenever possible. In other words, attempt to migrate all workloads of a given application in as few waves as possible. This approach helps to keep intra-application traffic local to the SDDC.
- Plan to isolate larger/complex workloads to dedicated waves. Larger/complex workloads are typically databases or those VMs which have a high rate of I/O writes. These types of workloads tend to generate excessive amounts of delta data replication and may negatively impact other migrations.
- Choose the most appropriate migration option which fits the needs of each workload group. If a workload can afford to be powered off, then use a cold migration. Bulk migration is the preferred migration method for most workloads. Most workloads can tolerate the relatively small amount of cutover time imposed by a bulk migration. Additionally, VMware HCX bulk migration provides the opportunity to perform certain updates such as VM machine version and vmtools upgrades; another benefit is that it keeps a shutdown copy at the source, so it can be restored quickly if needed.
- Plan for contingency and recovery of the business activities even in the case some of the workload of the wave will have issue during the scheduled time.
- Make sure the operations, monitoring, and backup teams are aware of the Migration Events and disable/suspend specific automated tasks (i.e. taking snapshots, starting filesystem backups, databases exports etc.) which will impact the migration.
- Automate the initiation of migration waves. Often migration tools offer application programming interfaces (APIs) that allow writing scripts to batch the procedure. Automation not only simplify the migration activities but reduce the human errors.

VMware Migration Solutions

With migration wave and events in mind, we can take advantage of several VMware based tools.

- VMware Content Libraries
- VMware vSphere Replication(embedded in SRM, HCX)
- VMware HCX
- VMware Cross vCenter vMotion
- API’s & PowerCLI

VMware Content Libraries

The Content Library service provides simple and effective management of OVF templates, ISO images, and scripts for vSphere administrators. The Content Library service lets you synchronize content across vCenter Server instances.

Content libraries are container objects for VM and vApp templates and other types of files, such as ISO images, text files, and so on. To deploy virtual machines and vApps in the vSphere inventory, you can use the templates in the library. You can also use content libraries to share content across vCenter Server instances in the same or different locations. Sharing templates and files results in consistency, compliance, efficiency, and automation in deploying workloads at scale.

By using Content Libraries, organizations can use vCenter to subscribe to their existing Content Libraries to move content up into a VMware Cloud platform without using third party tools or software.

VMware vSphere Replication

VMware vSphere Replication is an extension to VMware vCenter Server that provides a hypervisor-based virtual machine replication and recovery. It is also a part of VMware Site Recovery Manager, VMware Site Recovery Service, and VMware HCX.

vSphere Replication is an alternative to storage-based replication. It protects virtual machines from partial or complete site failures by replicating the virtual machines between the following sites:

- From a source site to a target site
- Within a single site from one cluster to another
- From multiple source sites to a shared remote target site

Another automated and supported method of moving VMware based workloads up to the cloud is by using vSphere Replication between vCenters on premise and in the cloud. This method is well understood and used by many organizations today. It is included in existing VMware vCenter licensing, VMware Site Recovery Manager, VMware Site Recovery Service, and VMware HCX.

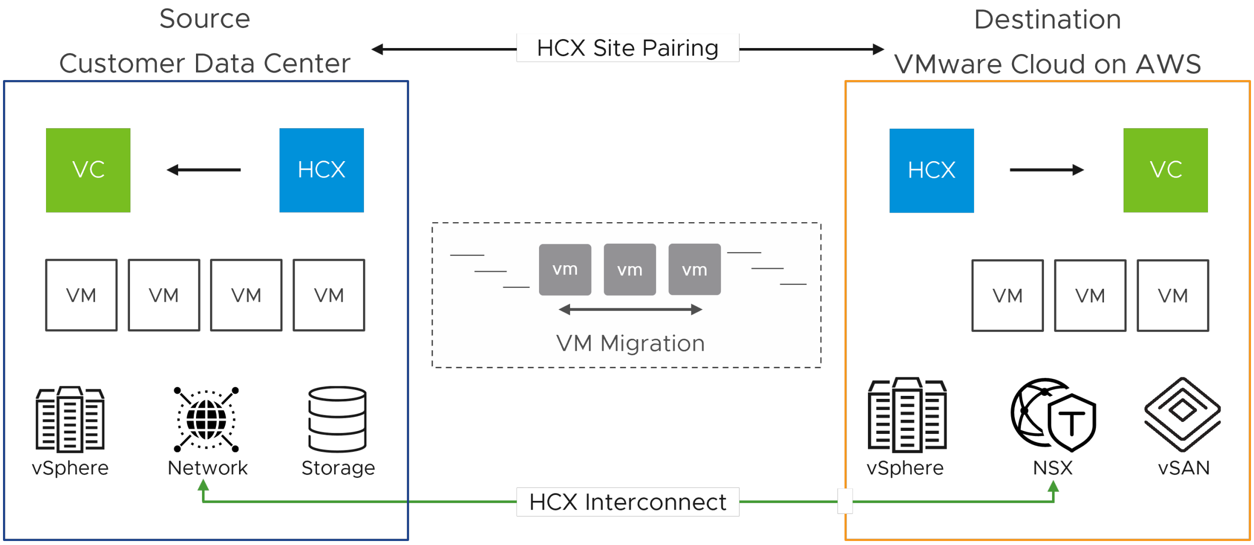
VMware HCX

VMware HCX™ is an application mobility platform designed for simplifying application migration, workload rebalancing and business continuity across data centers and clouds.

The HCX platform provides a hybrid interconnect to enable simple, secure and scalable application migration and mobility within and across data centers and clouds.

VMware HCX abstracts vSphere-based on-premises and cloud resources and presents them to the applications as one continuous resource. At the core of this is a secure, encrypted, high throughput, WAN optimized, load balanced, traffic-engineered hybrid interconnect that automates the creation of a network extension. This allows support for hybrid services, such as application mobility, on top of it. With HCX hybrid interconnect in place, applications can reside anywhere, independent of the hardware and software underneath.

The following figure illustrates the conceptual design for the HCX solution between two sites.



HCX Migration Types

Virtual Machines can be moved to and from VMware HCX-enabled data centers using multiple migration technologies. The migration type depends largely on the use-case, number and size of migrated virtual machines, and network bandwidth. The migration type is also restricted by the HCX license available.

The following sections describe the available migration options.

VMware HCX Bulk Migration

This migration method uses the VMware vSphere Replication protocols to move the virtual machines to a destination site.

- The Bulk migration option is designed for moving virtual machines in parallel.
- This migration type can set to complete on a pre-defined schedule.
- The virtual machine runs at the source site until the failover begins. The service interruption with the bulk migration is equivalent to a reboot.

Note For the most up-to-date list of requirements and restrictions when using HCX Bulk Migration, please refer to the Understanding VMware HCX Bulk Migration section in the VMware HCX online documentation.

VMware HCX vMotion

This migration method uses the VMware vMotion protocol to move a virtual machine to a remote site.

- The vMotion migration option is designed for moving single virtual machine at a time.
- VM is migrated while powered on. There is no service interruption during the VMware HCX vMotion migration.

- The vm data is encrypted by HCX.

Note For the most up-to-date list of requirements and restrictions when using HCX vMotion, please refer to the Understanding VMware HCX vMotion and Cold Migration section in the VMware HCX online documentation.

VMware HCX Cold Migration

This migration method uses the VMware NFC protocol. It is automatically selected when the source virtual machine is powered off.

Note For the most up-to-date list of requirements and restrictions when using HCX Cold Migration, please refer to the Understanding VMware HCX vMotion and Cold Migration section in the VMware HCX online documentation.

VMware HCX Replication Assisted vMotion

VMware HCX Replication Assisted vMotion (RAV) combines advantages from VMware HCX Bulk Migration (parallel operations, resiliency, and scheduling) with VMware HCX vMotion (zero downtime virtual machine state migration).

Note For the most up-to-date list of requirements and restrictions when using HCX Replication Assisted vMotion, please refer to the Understanding VMware HCX Replication Assisted vMotion section in the VMware HCX online documentation.

VMware HCX OS Assisted Migration

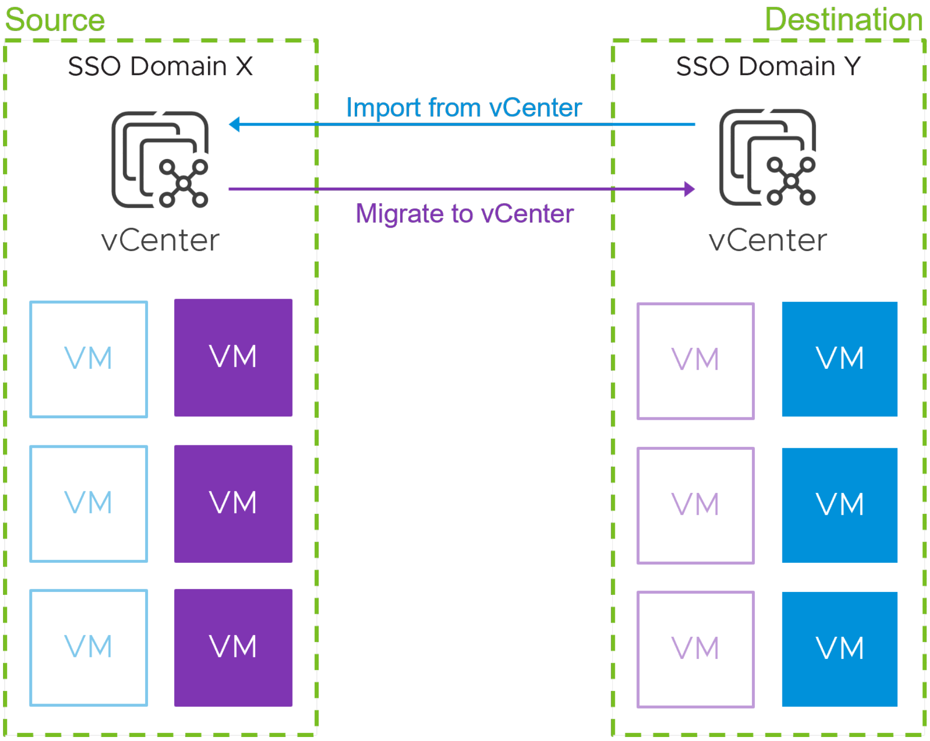
This migration method provides for the bulk migration of guest (non-vSphere) virtual machines using OS Assisted Migration to VMware vSphere on-premises or cloud-based data centers.

Note For the most up-to-date list of requirements and restrictions when using HCX OS Assisted Migration, please refer to the Understanding VMware HCX OS Assisted Migration section in the VMware HCX online documentation.

TYPE	DOWNTIME	NO DOWNTIME	MIN DOWNTIME	CONVERSION DOWNTIME	SUMMARY
COLD MIGRATION	✓				■ NFC Protocol
HCX vMotion		✓			■ Serialized
Bulk Migration			✓		■ Parallel ■ Large Scale
Replication Assisted vMotion		✓			■ Parallel ■ Large Scale
OS Assisted Migration				✓	■ Hyper-V ■ KVM

VMware Advanced Cross vCenter vMotion

The Advanced Cross vCenter Server vMotion (XVM) helps to migrate virtual workloads between vCenter Server instances, without the requirement for Enhanced Linked Mode (ELM) or Hybrid Linked Mode (HLM). This means it's possible to migrate virtual machines (VMs) between vCenter Servers that are in different Single Sign-On (SSO) domains. The XVM capability is embedded into the vSphere Client with the vSphere 7 Update 1c release.



From within the vSphere Client, two workflows are available to migrate workloads between vCenter Servers. Either using the 'import VMs' option in the Hosts and Cluster view to import VMs from a target vCenter Server Appliance (VCSA), or by selecting VMs and opting for 'Migrate' in the menu.

For detailed information, visit Advanced Cross vCenter Server vMotion Capability .

Modernization Through Re-platforming

In contrast to rehosting, re-platforming more aggressively focuses on resources closer to the application than rehosting.

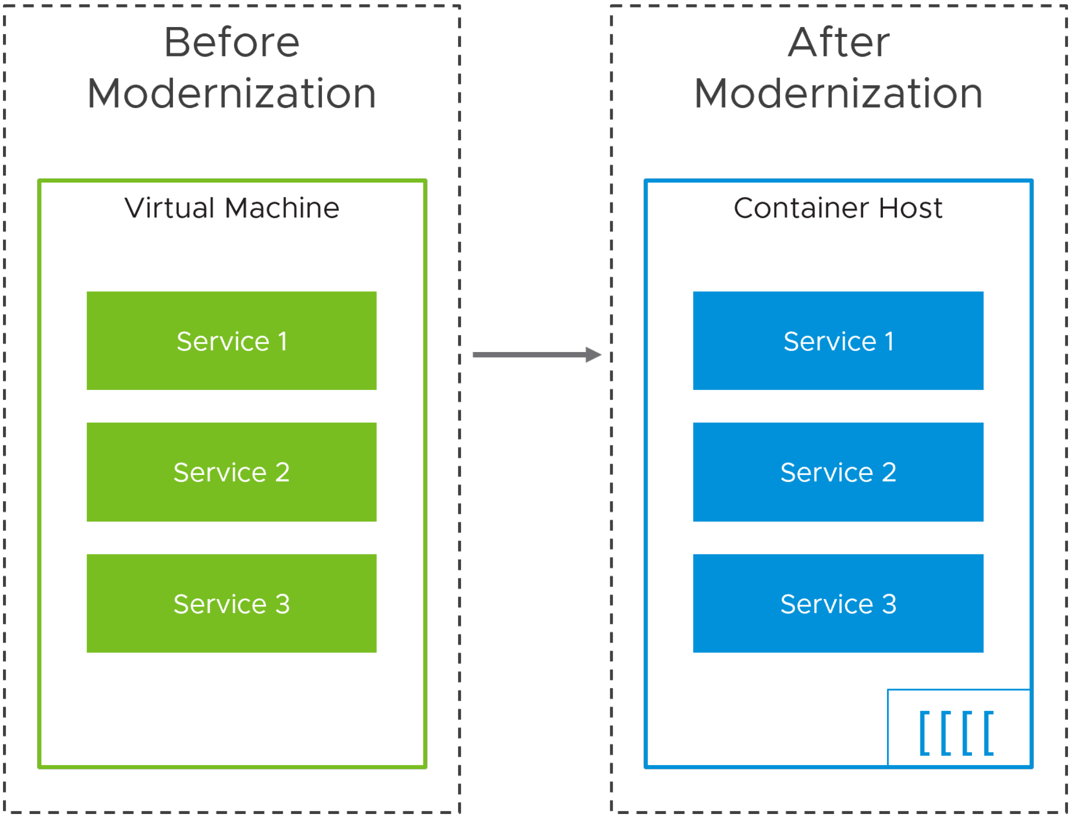
Deliver Modernization Through Re-Platforming

While rehosting typically consists of a virtual machine to virtual machine type migration, re-platforming replaces part or all of an application with a virtual machine alternative.

Re-platforming does not always require modifying the application code, but it does assume that some of the components of the application are not virtual machines. This re-platforming process may result in the end state for an application, or it may be an intermediate step within the modernization continuum. There are several common re-platforming patterns easily used in VMware Cloud environments.

Containerize the Monolith

One way to modernize a monolithic application running within a virtual machine is to use the same code base but change the form factor to a container. A containerized application is managed much differently from a virtual machine and comes with its own strengths and weaknesses. A monolithic application moved to a container can take advantage of the portability that containers provide such as portability, immutability, and start up time but requires a change to the operational processes to manage the application.



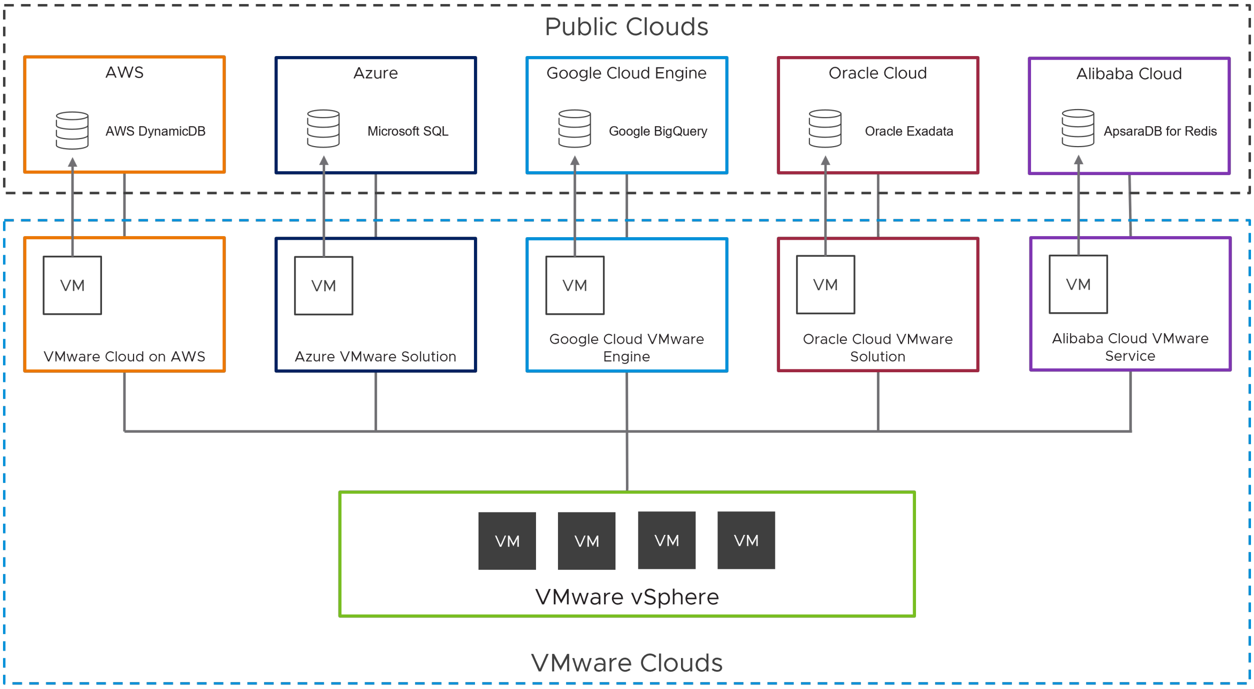
Converting a monolithic application into a container can be accomplished in a couple of ways. The first is to repackage the application as a container using tools such as the Tanzu Build Service, Cloud Native Buildpacks, or creating a Dockerfile for the application. This method would be analogous to building a new application as a container after the application code has been written.

An alternative to building a container from the source code would be to take a running virtual machine and convert it to a container. VMware Cloud customers have access to Application Transformer for VMware Tanzu, which can be used to convert running virtual machines into a containerized image, that can later be deployed. This is particularly effective for WebSphere and Tomcat applications due to Application Transformer's capabilities.

Adjacent Cloud Services

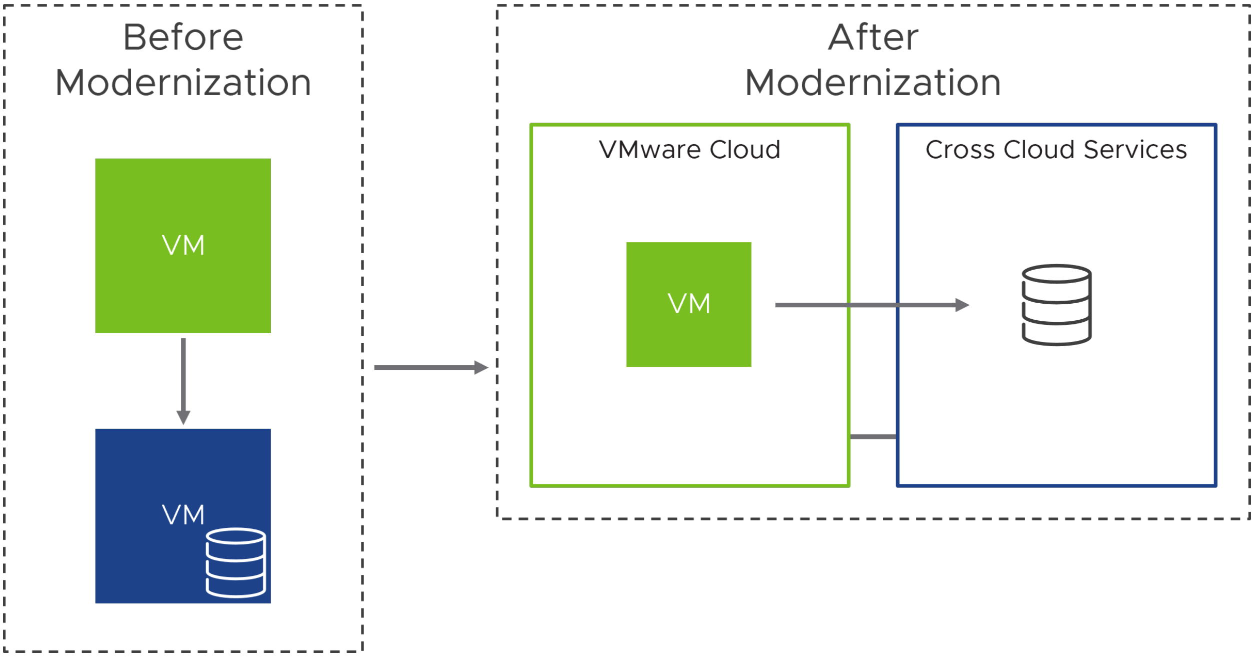
The public cloud offers a wealth of options for redesigning your applications. The public cloud providers offer a variety of managed services that can be used with your applications. These managed services come at a cost but can reduce the amount of time operators spend on maintaining systems. The managed services come in a variety of types such as managed SQL or NoSQL databases, event busses, Internet of Things (IoT), content delivery, AI/ML services and many more.

VMware Clouds are paired with public cloud infrastructure allowing application architects to choose which services should be utilized as part of an application. For example, instead of needing to provide a virtual machine with a database installed within it, application architects can pick a database service provided by a public cloud provider, negating the need for in-house maintenance tasks such as backups and patching.

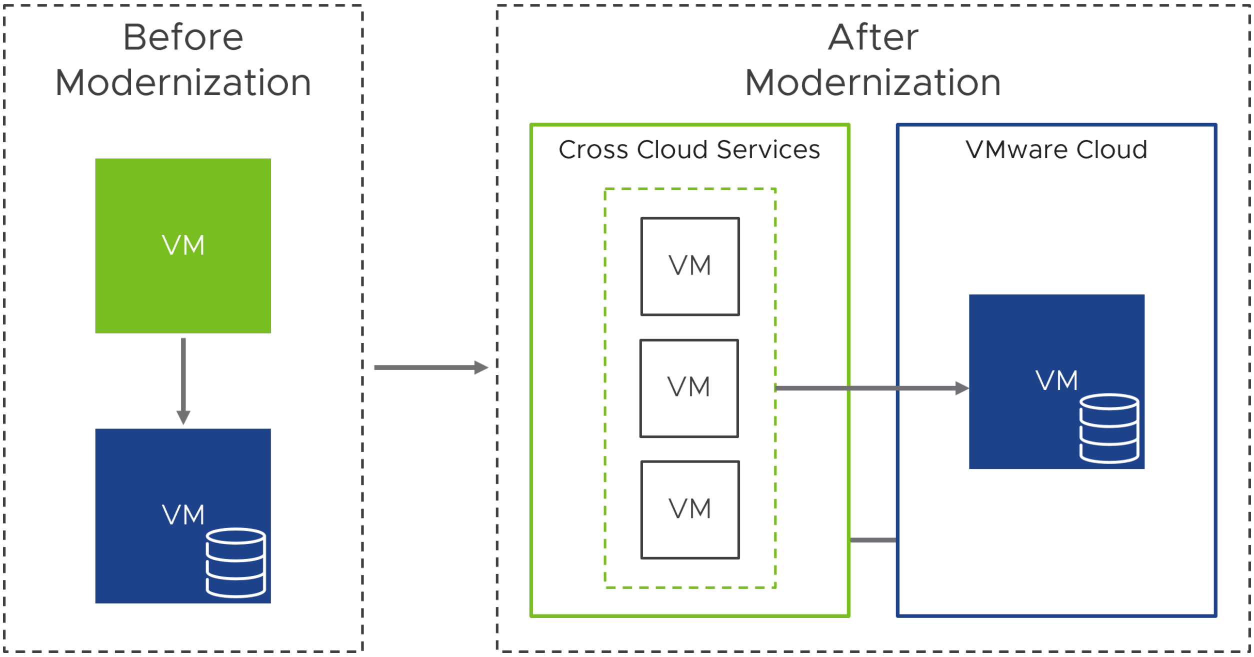


Adjacent cloud services can be an extremely effective option for VMware Cloud customers. Each VMware cloud can be thought of as having specific capabilities depending on which public cloud it is paired with. These capabilities can be used in your planning process to identify which landing zones will be the right fit for your modern applications.

In the example below, there is a typical two-tier application running in virtual machines. This application could be moved to a VMware based cloud and the database virtual machine can be replaced with a public cloud database service.



To look at a similar example, the database virtual machine could remain in your VMware based cloud, while the application virtual machine can be replaced by an autoscaling group of virtual machines, providing better horizontal scaling characteristics.



Modernization Through Refactoring

Refactoring applications requires modifying the code used to run the applications. When refactoring applications there is an opportunity to make sweeping changes to the code to reduce technical debt, increase performance, distribute components, or even change the coding language.

Deliver Modernization through Refactoring

The primary goal of refactoring is not to add new functionality to an application, but rather make it more efficient and easier to maintain.

Reducing technical debt and making it more extensible can improve the speed at which new features are added in the future and lower the cost to maintain the application.

Deconstruct the Monolith

Monolithic applications are difficult for multiple teams to work on together. Code changes made by one team are committed to the central code base and can often interfere with other teams working on the same application.

Monolithic applications can struggle with scale if not designed specifically with this in mind. If parts of the application require additional physical resources, the entire monolith must be scaled to accommodate the needs for the single service within the app. Because of these reasons, application architects often break down the monolithic application into pieces so that they can be worked on individually and can scale independently from each other.

The Twelve Factor App is a modern application methodology for building software specifically designed for cloud platforms and is a helpful guide when refactoring. As the monolith is deconstructed into smaller components, consider the following from the 12 Factor App Framework.

Backing Services

Backing services consist of any service the application consumes over the network for normal operations. This could be a third-party API service, an email system for notifications, a file store, or a database. Backing services often store state for a distributed application.

Carefully plan which backing services should be used with your modernized application. Since backing services store state data, they are more difficult to move to different clouds. Consider using backing services that are portable across cloud platforms to provide multi-cloud flexibility. For example, using a traditional cloud managed database offering might not be portable to a different vendor cloud’s database offering.

Processes

An application is executed as a series of one-to-many processes that share nothing and are loosely coupled. Stateful data should not be stored within these processes, including session state.

All state should be stored within a backing service to ensure portability and scaling capabilities won't affect the application.

Concurrency

As processes are separated, they should support concurrency to allow for horizontal scaling. In contrast to monoliths which generally scale vertically by increasing CPU or memory, twelve-factor apps can scale horizontally for just the processes needing more resources.

Monoliths require more resources granted to the entire application, whereas a twelve-factor app can scale part of the application without affecting the rest of it. Concurrency is critical to a distributed systems' reliability. The ability to horizontally scale also provides availability since each process can be deployed in pairs to protect against outages.

Disposability

Processes in a twelve-factor app can be started and stopped at any time. Disposability of our processes means that these modern applications should be built with an expectation of the processes failing.

Disposable processes should be optimized to reduce startup time which provides the ability to quickly scale more services as the system needs.

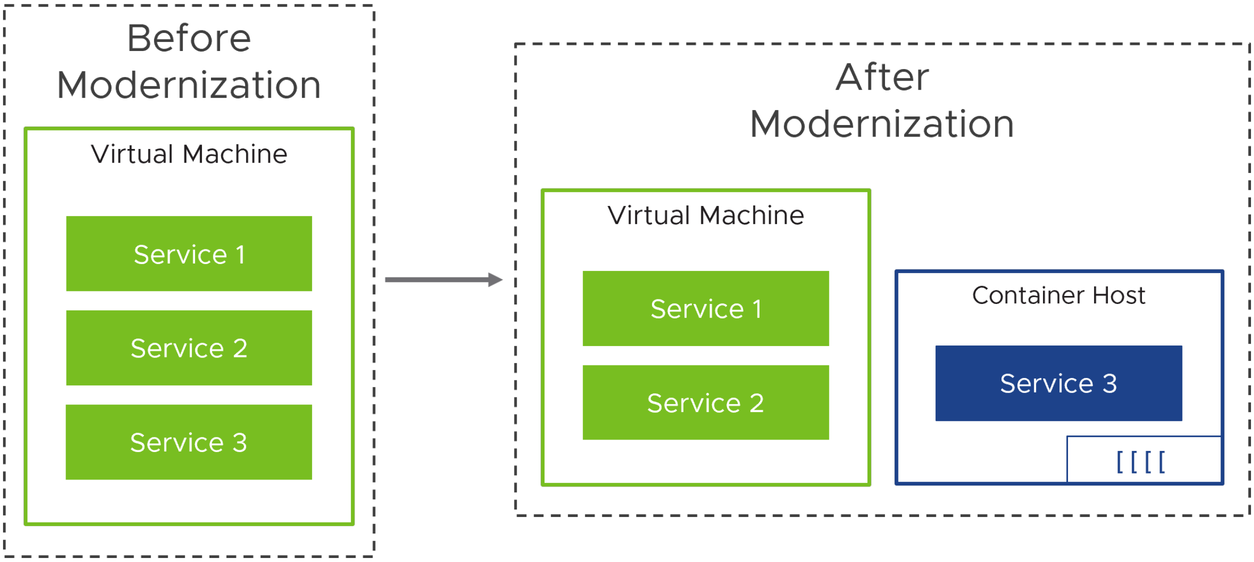
Event Streams

Distributed systems may need log correlation to determine issues. Event streams make it possible for each individual component to log messages in a similar method and later aggregated and correlated with a log management solution.

Dependencies

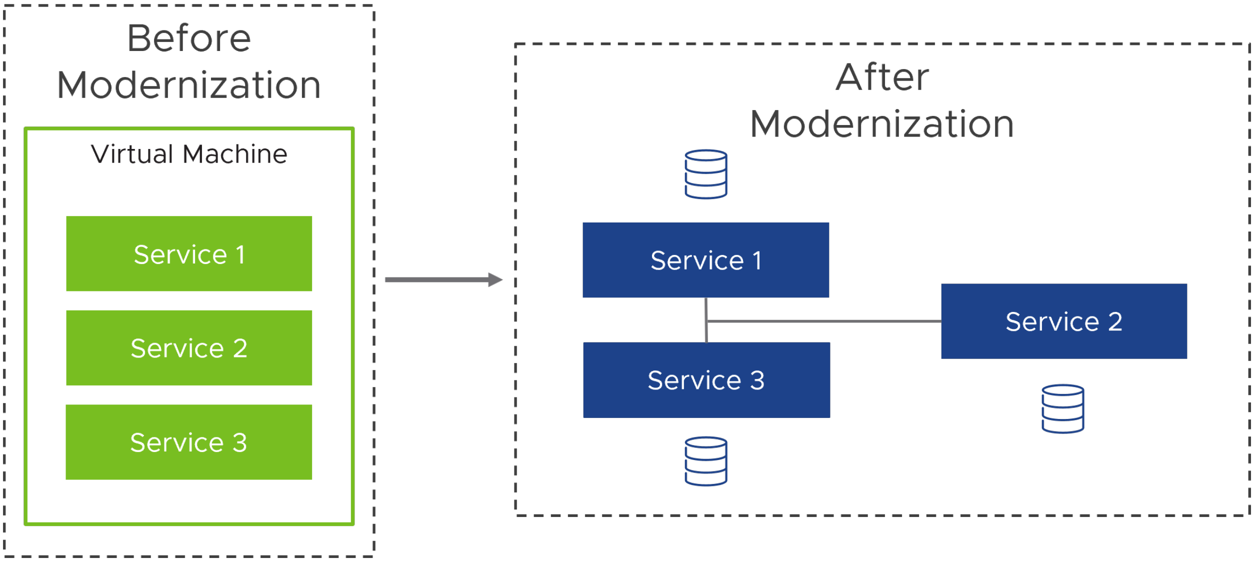
Application dependencies should never be implicitly inherited from the system they run on. Each application should explicitly define the dependencies and versions used to build the application. This removes the tight coupling that may have been needed in legacy apps where they host system must have the right dependencies installed prior to the application be deployed.

When decomposing a monolithic application, containers are often used instead of a virtual machine to reduce the resource overhead. Containers also are easily restarted, aiding with disposability, and are easily scalable.



Build Microservices

On the extreme end of the modernization continuum are microservices. Microservices are a architectural approach where the applications are composed of independent services separated by a contract such as an API. Each microservice has a specialized function and may adhere to the twelve-factor application framework including having their own backing services.



Microservices bring additional challenges to the operations of an application including orchestrating the deployment of many components across fault domains and providing service discovery between components. Kubernetes provides a container orchestration solution capable of managing containers across hosts and providing a cloud native platform with service discovery. VMware provides Tanzu Kubernetes Grid to all VMware Cloud customers to run their container-based workloads in a vSphere environment. Tanzu Kubernetes Grid can also be used across native cloud environments to keep a consistent Kubernetes runtime across clouds.

Modernization Operations

Operate in the Cloud

Workloads built or moved to the cloud still need a plan for operations. Applications that keep the same form factor during a migration such as a VMware virtual machine may need to be added again to logging platforms, monitoring tools, and security patching strategies since they are now located in a different data center.

Applications that change form factors such as a re-platformed application may require entirely new operations routines to ensure they are monitored, secured, audited, etc.

Verifying Application Functionality

An organization's applications team are the key stakeholders to verify application functionality after a migration or transition. As such, they should have been a part of the migration event planning activities and received advanced notification.

Some of the key joint planning tasks to address post migration can include:

- The validation time can range from few minutes to some hours depending on the tasks agreed for validation.
- Ensure the application team is using the application functionality tests and has confidence of the expected timeline.
- Conclude with an approved or unapproved validation answer from application team, which in some rare circumstances may trigger a revert of the workloads to the sources.

Verifying Application Performance

Application performance can be a subjective topic during and after a migration. Interpretation of performance profile and metrics will need to be definitively agreed upon prior to migration. This will allow all parties to reach consensus based on agreed upon metrics quickly.

Ensure identification and collection of proper KPI's during planning. Proper KPI metrics should be monitored with enough time to provide adequate data for comparison of application performance before and after migration to minimize performance evaluation and narrow a scope of investigation if the application is not performing once migrated.

VMware Wavefront or VMware Aria Operations (SaaS) could be helpful in monitoring application KPI's.

Post-Migration Cutover Events

Virtual machines migrated from a data center into the cloud, still require the same operational processes as virtual machines on-premises. Moving virtual machines to the cloud may provide additional options not available to you previously with an on-premises virtual machine.

Network Connectivity - The post-migration cutover refers to making the target SDDC the new "home" for a given set of migrated virtual machines. The exact process depends on the IP addressing strategy that is chosen for migrated workloads. If you've opted to change workload IP addresses, then a cutover would mean ensuring that the new address space is reachable by end-users and that all DNS has been updated to reflect the change. If instead, you've opted for workloads to keep their IP addresses, then a cutover would involve disconnecting any layer-2 extensions for migrated networks and configuring routing such that the migrated networks are reachable via a routed path to the SDDC (VPN, other specific Cloud connectivity).

Asset Tracking – Virtual machine objects being moved between data centers may have an impact on security, licensing, and costs. As virtual machines are moved to a different location, such as the cloud, these asset or configuration databases should be updated to reflect the new state of your applications after a migration.

Monitoring – Many times, ownership of virtual machines by Infrastructure teams may change if the virtual machine exists in the cloud instead of on-premises. It is important that monitoring is appropriately configured and alerts go to the right teams as workloads are moved to different data centers.

Disaster Recovery – Add your virtual machines to a new disaster recovery routine. You may be using a second cloud location to protect against a regional disaster or you may start using your on-premises locations as your fail over location. It is vital to ensure that your disaster recovery routines are still viable after migrating your virtual machines to cloud infrastructure.

Post-Refactor or Replatform Events

Modern applications often require different operational tasks to maintain them. As monolithic applications are broken down into multiple components, how the application needs to be maintained has changed.

Consider the following operational tasks on a modernized application that doesn't consist of a virtual machine.

Recoverability - If an application is made up of a set of disposable microservices, there shouldn't be state data within services, so they don't require backups. However, the backing services for a monolith might've been a single source, meaning a single item to backup. In a distributed system, multiple backing services may need to be protected from failure, and subsequently, more backup routines are needed.

Deployment - Monolithic applications result in a virtual machine being provisioned. Microservices may require for many services to be deployed. If your corporate deployment process involves opening and closing tickets to deploy a service, the number of tickets being created will dramatically increase. A continuous deployment methodology may be a better fit for a modern application architecture.

Observability - Within a virtual machine, it is common to have an agent installed to gather performance metrics or logs. With microservices-based architectures, logs are distributed, meaning that to correlate logs between services, you must first aggregate them. Performance metrics in a container-based solution are often published by the application and scraped by a solution such as Prometheus.

Security Updates - Virtual machines running a monolith are often left in place for long periods and patched regularly to mitigate vulnerabilities. Container-based workloads aren't patched in place but are rebuilt and redeployed with a fresh image. Be prepared to rebuild microservices as part of a regular security mitigation process.

Network Security - In a simple two-tiered application between a virtual machine and a database, there is one network path to secure. When using microservices, there may be many network paths to secure. A service mesh solution such as Tanzu Service Mesh might be used to secure these connections with mutual transport layer security (mTLS).

Secure Pillar

4

This chapter includes the following topics:

- Introduction to Security
- Compliance
- Design Considerations
- Administrative Access
- Endpoint Security
- Conclusion

Introduction to Security

Information security is, at its core, defined by three main concepts: confidentiality, integrity, and availability. Organizations often think about "security" primarily in terms of confidentiality, but omitting the other concepts leaves out many essential design ideas that help supply resilience to systems, during incidents ranging from a security breach to a natural disaster. The concepts also interrelate. For example, designing for availability often means that resolving vulnerabilities is easier, helping to restore confidentiality and integrity to systems. Thinking about security in the cloud means thinking about, and designing for, all three data security concepts.

What is Security?

Security is a process that belongs to everyone in IT and technology, not just the "security team" or the chief information security officer (CISO).

This is especially true in small and medium-sized businesses that may not have a separate security team. Regardless of an organization's size, cloud and virtual infrastructure administrators are often the primary implementers of security controls in an organization, while the security team sets policy, audits for compliance, and conducts incident response. Everyone within the organization participates in the process in some way. As such, the information presented here is intended to aid everyone in gaining an understanding of security design principles, and how they apply to operations in the VMware Cloud.

Information Security Concepts

An understanding of information security concepts enables efficient communication within organizations, promotes understanding among different groups within an organization, and improves system design by highlighting areas of consideration.

Authentication

The ability to prove that a person or application is genuine, verifying the identity of that person or application. Authentication uses one or more of three primary methods, or factors: what you know, what you are, and what you have.

“What you know” encompasses passwords, personal identification numbers (PINs), passphrases, and other secrets. This type of authentication is not strong on its own and is typically paired with another authentication factor.

“What you are” involves biometric authentication methods, such as retinal scans, fingerprints, voice or signature recognition, and so on. These factors cannot be easily changed if compromised.

“What you have” entails objects or applications running on objects that you physically possess. Traditionally this involved keys, but modern forms may also involve USB tokens, smart cards, and one-time password applications on devices. This factor requires possession of the object at the time of use and may be hindered by intentional or unintentional loss of, or damage to, the object.

Multi-Factor Authentication is a method that uses authentication techniques from more than one factor. For example, combining a password with a one-time password application, or a facial scan with a PIN. This approach helps mitigate weaknesses in the use of each factor. Use of two techniques from the same factor, such as two passwords or two physical keys, is not considered multi-factor.

Authorization

The act of determining whether a user or application has the right to conduct particular activities in a system, relying on authentication to prove the identification of the user or application.

Availability

Ensuring that data is available to authorized parties when needed.

Compensating Control

Security and privacy controls implemented as an alternate solution to a requirement that is not workable for an organization to implement in its original form. The sum of the compensating controls must meet the intent and requirements of the original security control.

Confidentiality

Ensuring that data is protected from access by unauthorized parties.

Data Breach

An incident where data is accessed, copied, transmitted, viewed, or stolen by an unauthorized party. This term does not indicate intent; other terms such as “data leak” and “information leakage” help convey whether a data breach was intentional or not.

Identification

The ability to uniquely prove who a user of a system or application is, to enforce access control and establish accountability.

Incident

The attempted or successful unauthorized access, use, disclosure, modification, or destruction of information or interference with system operations. Note that this is not limited to people, nor does it indicate intent; natural phenomena, disasters, and animals can also cause incidents, for example.

Integrity

Ensuring that data is protected against unauthorized modification.

Lateral Movement

A method of describing the techniques used by attackers, after breaching an endpoint or system, to “pivot” and extend access to other systems and applications in their target organization. This moves the attacker closer to their goals, such as accessing, changing, exfiltrating, or destroying sensitive information.

Least Privilege

Only assigning the minimum access rights that are necessary for staff or systems to perform their authorized tasks, for the minimum duration necessary.

Non-repudiation

The ability to associate messages, actions, and/or authentications with an individual in a way that cannot be denied by that individual.

Recovery Point Objective (RPO)

The largest amount of data that is acceptable to lose after recovering from an incident. This is measured in time, e.g. “one hour of customer data.”

Recovery Time Objective (RTO)

The largest amount of time that is acceptable for data to be unavailable due to an incident.

Security Control

A safeguard or countermeasure designed to protect the confidentiality, integrity, and availability of data.

Separation of Duties

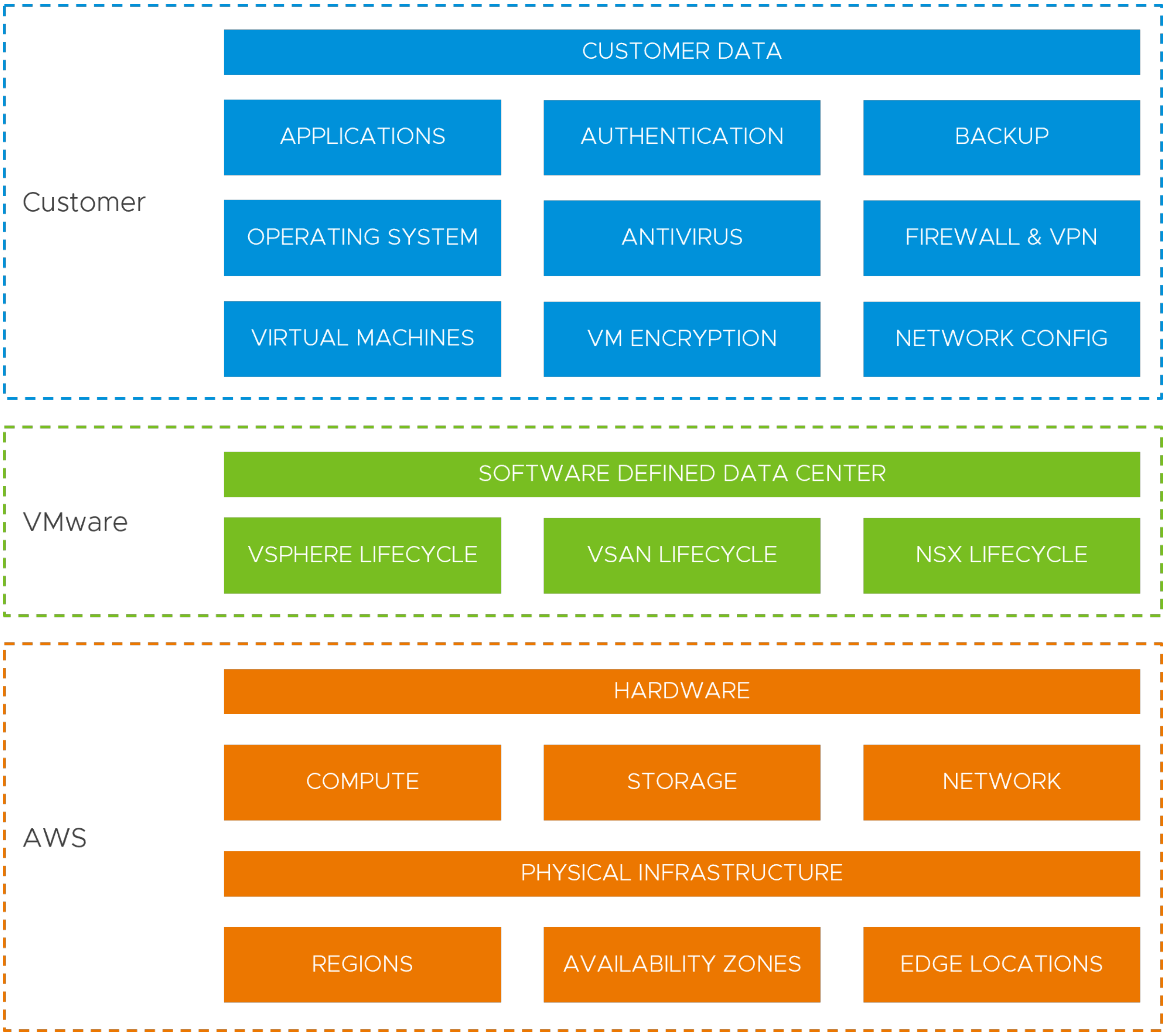
Dividing critical functions among different staff to help ensure that no individual has enough information or access to conduct fraud.

Vulnerability

A weakness in an information security system, system security procedures, security controls, or implementations that could be exploited by a threat actor.

Shared Responsibility Model

Whenever there are multiple groups working together on a system it is helpful to define roles and responsibilities. The Shared Responsibility Model does that for VMware Cloud, helping make clear who supports what components of a deployed SDDC.



Customer

CUSTOMER DATA

APPLICATIONS | AUTHENTICATION | BACKUP

OPERATING SYSTEM | ANTIVIRUS | FIREWALL & VPN

VIRTUAL MACHINES | VM ENCRYPTION | NETWORK CONFIG

VMware

SOFTWARE DEFINED DATA CENTER

VSPHERE LIFECYCLE | VSAN LIFECYCLE | NSX LIFECYCLE

AWS

HARDWARE

COMPUTE | STORAGE | NETWORK

PHYSICAL INFRASTRUCTURE

REGIONS | AVAILABILITY ZONES | EDGE LOCATIONS

Working from the bottom up, the public cloud provider handles the design, implementation, and operation of the physical computing environment, including servers, networking, power, and other obligations within their data center facilities. This also includes regulatory compliance certification of those components, where required.

From there, VMware manages the configurations of the physical equipment that the cloud provider supplies, as well as the relationship to the cloud provider themselves. VMware installs, maintains, secures, patches, and upgrades ESXi, vCenter Server, NSX, vSAN, and other infrastructure management components that are required for an SDDC. VMware does this seamlessly by using classic vSphere resiliency features, like DRS, HA, vMotion, and the like. Where outages may be noticed, such as with an update to vCenter Server, notifications will happen via email to SDDC administrators.

Customers handle the workloads, data, choices about availability, and the methods of accessing workloads and data over the network. VPN configurations, public IP addresses, and network security are managed by customers as well.

Compliance

Regulatory Compliance

Regulatory compliance is an important consideration for many organizations, enabling them to participate in specific industries that require minimum standards for organizational processes and technology.

Compliance vs. Security

Regulatory compliance is a business requirement driven by the need to perform regulated tasks like accepting credit cards as payment, conducting health care activities, running energy production facilities, and more. In contrast, security is driven by the need to protect an organization's assets from constant threat.

Both activities often deal with security controls, but regulatory compliance is only assessed periodically through an audit. At the end of the audit an organization is granted an "Authority to Operate" wherein they can begin or continue the regulated activity.

How Compliance is Achieved

Regulatory compliance is assessed on implementations of systems and products, not on the products themselves.

An auditor does not deal with hypothetical situations, system designs, or product capabilities. They want to see how the system is built and operated. While a VMware Cloud-based SDDC has hundreds of security features and is validated for use in the world's most sensitive environments, it is still possible to make implementation decisions that provide opportunities for attackers and disasters. An auditor seeks to find those problems and shine a light on them.

VMware Cloud Trust Center

Because auditors assess implementations of systems, and VMware Cloud is an implementation of VMware's infrastructure products, VMware can certify the environments against common regulatory frameworks.

VMware keeps a record of these certifications at the VMware Cloud Trust Center:

https://www.vmware.com/products/trust-center.html

Given the Shared Responsibility Model, certifications of infrastructure do not carry over to workloads themselves but do help organizations conduct audits faster because third-party auditors already certify the infrastructure, with the results posted in the VMware Cloud Trust Center or available under non-disclosure agreement via account teams.

Design Considerations

Security always depends on context, and in most cases, the context is influenced by how an organization intends to use its cloud presence.

Infrastructure Dependencies & Design Considerations

Access control, roles, permissions, network connectivity, and other security controls will differ between a VMware Cloud SDDC intended only as a disaster recovery site and a VMware Cloud SDDC running active production workloads.

Determining the use case of your VMware Cloud SDDC and then documenting it helps teams and organizations make decisions about security, risk, and availability, as well as helping decide what needs to change if the scope of the deployment changes.

DNS Availability and Records

The Domain Name System (DNS) is crucially important to both workloads and the users of those workloads.

DNS provides an abstraction layer so that IP addresses don't need to be tracked and used directly, and more human-friendly names can be used. TLS certificates are also correlated with DNS, helping establish trust between systems within virtual infrastructures. As such, most organizations depend heavily on DNS, and require that their DNS servers be made highly available. Your organization's intentions for an SDDC will determine the methods you use to ensure DNS availability.

VMware Cloud SDDCs provision resolvers and DNS records during deployment for use by the infrastructure. This ensures that the infrastructure is always reachable internally through the DNS names.

DNS is a powerful system, and the hierarchy in it can be used to help track other information, such as where a workload is running, which makes IT operations easier. However, as workloads move between clouds other systems (like a Configuration Management Database) may need to be kept up to date, too. Location information in DNS can also leak information to attackers about where your organization's facilities are. Security is always a tradeoff; ensure your organization is comfortable with the risks.

Ideas to consider:

- DNS records help determine how network traffic flows. If a domain name resolves to an IP address internal to your organization then traffic will flow on internal links and VPN connections between sites. If a domain name resolves to a public IP address then traffic will flow across the public Internet.
- Latency of DNS resolution has a profound impact on the performance of workloads and services, from response time to overall system load. Keeping DNS resolution as close to a workload as possible ensures the best performance as well as site resiliency if network links become unavailable.
- Do the DNS servers you use supply authoritative name services for customer-facing and external services? If those systems are unavailable will your customers be unable to reach you? Many cloud providers have DNS services that can be used across local clouds and global public cloud regions. This can add resilience while simplifying management & workload migrations with hybrid deployments.
- "Split-brain" DNS methods, where one view of an environment is available to some clients, and another view is available to others, can be a useful tool for organizations. It also can be very confusing and lead to errors if there are multiple sources of authoritative information. The phrase "security through obscurity" was coined many years ago to describe the act of hiding things to secure them. This is not a legitimate approach to security.
- The method in which your organization implements DNS will determine how it can be made available in a hybrid or disaster recovery scenarios. Some DNS server software is fine with being cloned & replicated. Other software is not. One example of that is Microsoft Active Directory, where the best practices for deployments state that systems supporting an Active Directory should not be cloned, but instead installed as fresh deployments.
- How will your workloads reach the DNS server? Does your organization employ "service IPs" for DNS that are highly available, moving between DNS servers, or an anycast routing scheme to direct traffic to the nearest DNS resolver?
- How will workloads that move between an on-premises deployment and a VMware Cloud SDDC find their DNS resolvers? Will moving a workload require reconfiguring the network settings? How will reconfiguring many workloads during an incident affect your RTO?

- Do you have your authoritative DNS servers and DNS recursive resolvers separated, according to best practices for DNS operations? If so, you may need to employ different availability and security methods for each type of server. In general, authoritative DNS servers should not answer recursive DNS requests from clients, and recursive DNS resolvers should not be accessible publicly.
- Do your systems permit access based on domain names? For example, many Linux systems are configured with hosts.allow and hosts.deny files that can contain either IP address ranges or domain names that are permitted. Similar configurations can be achieved with guest OS firewall rules, too. During an outage where DNS is potentially affected will authorized administrators be able to connect to systems to repair them?

IP Addressing and Management

Organizations that migrate to the cloud find themselves with more complex IP address management. Decisions about addressing help determine the complexity of many other activities, especially failover planning, migrations, and firewalling. Working to simplify IP address allocations pays dividends in operational efficiency later.

Ideas to consider:

A VMware Cloud SDDC requires specific network allocations for the management components. Does your IP allocation strategy for the cloud assume that your organization will always only have a specific number of SDDCs, a specific number of sites, or that the sites are all in the same region?

Separating and isolating infrastructure management interfaces is an important step towards making it hard for attackers inside your environment. Does your IP addressing system allow for infrastructure management interfaces to be isolated from clients, workloads, and other infrastructure systems?

Network Address Translation

Network Address Translation, or NAT, is a technique where multiple network addresses can be mapped to a single IP address. In most cases it is used as a way to circumvent IPv4 address exhaustion on the public Internet. Many network providers only supply a single IP address to their customers, thereby requiring the use of NAT. These IP addresses may also be dynamic, changing periodically as the provider reprovisions equipment and networks.

There are two types of NAT: source NAT and destination NAT. Source NAT is the type of NAT most users are familiar with, as it translates internal IP addresses to a single public IP. By default, source NAT is applied to outbound network traffic on workload network segments. Destination NAT is the inverse of that and will translate public IPs to a single private IP address. Destination NAT is often used in conjunction with port forwarding to enable application access.

NAT is not a security control by itself. There are several ways to deanonymize network traffic that is obfuscated by NAT, and there are situations where unsolicited network traffic can be transmitted back through the NAT device, to probe networks and initiate attacks. Use of NAT should also be accompanied by use of firewalling technologies and rules. In VMware Cloud that is the case, with NSX protecting traffic in all directions from the management and compute gateways.

Ideas to consider:

Organizations with a distributed workforce and/or distributed branch offices may need to consider the impact NAT has on access control? Does your organization authorize people or systems based on IP addresses subject to NAT?

Storage Availability & Security

VMware Cloud SDDCs are built by default using VMware vSAN, using storage supplied by the cloud provider.

Additional options are available, including the Amazon FSx for NetApp ONTAP storage. When vSAN storage is configured it has vSAN Data-at-Rest Encryption and compression enabled. Key rotation for the vSAN datastore can be done through a VMware Cloud support request. vSAN datastores in VMware Cloud SDDCs offer choices for storage availability, including across multiple hosts, stretched clusters, and the number of hosts. These choices are done at SDDC provisioning.

Inside the cluster, the VM Storage Policies can be customized. VMware vSAN allows customers to choose the affinity and disaster tolerance in stretched clusters, as well as host failures to tolerate (from none to three). This allows an organization to balance capacity, performance, and space efficiency against their tolerance for risk, and their budget.

Elastic DRS (EDRS) is a tool within an SDDC to provision and deprovision SDDC ESXi hosts based on performance and capacity. For clusters with three or more hosts EDRS can be configured flexibly to optimize for best performance, lowest cost, or rapid scale-out. Minimum and maximum cluster sizes can be configured as well.

Ideas to consider:

- Changes in vSAN storage policies require a resynchronization process, which is not instantaneous. If your organization customizes storage availability policies to improve capacity will there be a time, perhaps as part of a failover process, where storage availability will need to change? Is that process documented? Is the risk during the resynchronization process acceptable? Are the relevant policies preconfigured to ensure they are correct, saving time and avoiding errors during an already stressful incident?
- Elastic DRS has configurations, such as with two-host clusters, where when a particular capacity is reached it will permanently extend the cluster, thereby changing the capacity but also the economics of the deployment. Does your organization monitor and alarm on storage capacity? Will an action by EDRS negatively impact your organization?

Templates and Container Registry

VMware Cloud SDDCs support template management through the vSphere Content Library. The Content Library makes storing, replicating, using, and updating templates, ISO images, and other system artifacts easy.

Customers migrating to VMware Cloud, or running in a hybrid design model, can configure their on-premises Content Library as a replication source, and configure their SDDC's Content Library as a consumer of that content. Not only does this enable day-to-day operations in the cloud, but also provides resilience for guest OS boot media and other recovery tools during an incident.

Virtual machine templates provide a straightforward way to deploy new VM-based workloads. There are many ways to manage templates, from completely manual processes to heavy reliance on automation tools like SaltStack. Automated methods of configuring a new virtual machine save time by guaranteeing consistency for system configuration, including software updates and security controls. In turn, this speeds audits and makes improving security easier. It also makes template management easier, because templates can be generic, customized at deployment time based on the current patch levels and system configurations in use.

Container based workloads rely on a container image to run. These container images are immutable, meaning that they cannot be changed and can be considered static. The nature of these immutable images means that there is no need to patch the images in place, but new container images should be built as new security vulnerabilities are identified with the individual components of the container image. Since the container images are rebuilt frequently, it is important to have a secure supply chain to ensure that no new vulnerabilities are unexpectedly introduced into an immutable image used by your container workloads.

Ideas to consider:

- Does your organization store installation and recovery media for all guest OSes in the VMware vSphere Content Library? Is that library replicated to all the sites that might depend on it for incident response?
- Does your organization have a process to regularly update content library content, templates, and other content stored as part of disaster recovery and business continuity processes? Are old template images removed in order to prevent redeployment of outdated and potentially insecure configurations?
- Does your organization use a configuration management and auditing tool such as SaltStack to ensure consistency of deployed virtual machines, and make building new workloads easier and faster? Configuration management tools reduce the complexity of templates and container images by managing the configurations once a workload is deployed.

Backup and Restore

Backup systems are the last line of defense against incidents, especially security breaches involving ransomware. Incidents can also encompass less dramatic situations, such as a failed application upgrade or human error. Being able to roll back a workload to a known-good state is a powerful protection.

Breaches involving ransomware can be quite long, measuring hundreds of days from the initial breach to the containment of the breach. Attackers are patient and will work to ensure that an organization must pay the ransom. This often entails disabling or corrupting backups. Organizations must make it difficult or impossible for attackers to access backup systems.

Recovery Point Objective (RPO) and Recovery Time Objective (RTO) are important considerations for determining backup frequency and scope, as well as whether workloads should also be protected with other means, such as with replication.

Ideas to consider:

- An attack occurring over a long period of time will cause your backup systems to capture the results of the attack on affected systems. This is an important consideration, because restoring a backup may also restore infected systems, and/or restore systems to a vulnerable state. How would you recover workloads in a questionable state, as well as how would you assess the reliability of such workloads?
- Are your backup systems isolated from corporate authentication and authorization systems, such as a centralized Active Directory? If an attacker gains administrative access to the central directory what security controls will stop them from accessing, deleting, and corrupting backups and replicated copies of workloads?
- Are workloads configured to separate operating systems, applications, and application data, so that if malware is found it might be possible to independently restore the data, remounting or reattaching it to a fresh installation of the application?
- Have you documented the restore procedure for workloads? Do you rehearse it regularly to ensure that it works, and that staff understand it? Are all components and tools for the restore available if the original SDDC is not available?
- Is it possible that you will need to restore your backups to a different availability zone or SDDC, following the loss of an SDDC or loss of access to an availability zone? Are your workloads able to have their public IP addresses renumbered? Do DNS entries have Time-To-Live values suitable for your desired RTO?
- Would you be able to recreate NSX network segments to restore internal connectivity to applications? Do you have backups of firewall and NSX network segment configurations?

Management Interface Availability

Access to management interfaces and cloud console interfaces can be incredibly paradoxical. Organizations need to limit access to them, but at the same time allow access to authorized staff, possibly from unexpected but legitimate IP addresses and locations if an organization's primary site is unavailable. This is where modern zero trust methods of authentication and access control are very helpful. The VMware Cloud Console and cloud management interfaces support multi-factor authentication, helping to ensure that only authorized users gain access.

Many organizations employ bastion hosts or “jump boxes” to help control access to management interfaces. Additionally, some organizations, including VMware internally, use dedicated VMware Horizon VDI deployments to provide secure & trusted access to systems management tools and interfaces. Staff connect to these systems, then can interact with infrastructure from a known & trusted management workstation image.

Ideas to consider:

- How will IT staff access cloud consoles and management interfaces to conduct recovery operations during an incident if the primary site is offline and potentially unrecoverable?
- Is multi-factor authentication enabled for all users of the VMware Cloud Console, and part of the login sequence to access infrastructure management interfaces?
- Are bastion hosts, jump boxes, and/or dedicated VDI instances patched quickly and proactively, to ensure that attackers cannot exploit new vulnerabilities to gain access?
- Do management interfaces rely on authentication and authorization provided by a central directory, such as Microsoft Active Directory, that may be unavailable during an incident? Is that directory considered “in scope” for compliance audits? How does your organization protect against unauthorized changes by administrators of those systems, potentially allowing privileged administrator access to infrastructure systems?

Incident Response and Business Continuity

Organizations that make the shift to assuming that a breach will happen are the organizations that tend to be the most prepared if it happens. Making this assumption is an important change of mindset an organization needs, to combat ransomware and other types of attacks. Ensure that attacks are covered in your organization’s disaster recovery & business continuity planning. Plan for an “everything down” scenario.

Ideas to consider:

- Does your organization have its own security response team, or has your organization proactively engaged a security consultancy that specializes in incident response? Incident response is a separate function from business continuity planning, but crucial for understanding how an attack happened and how to recover in a way that preserves evidence and prevents the reoccurrence of an attack. Does your incident response team have a plan for response?
- Ensure that contact information and roles & responsibilities documents are stored in a place that will be accessible if IT systems are offline. Many organizations, with otherwise terrific business continuity plans, have found themselves hampered because their plans were stored on systems that were inaccessible because of the outage.

Administrative Access

Protecting the management interfaces of infrastructure is critical, as virtual and cloud administrators have enormous power over workloads and data.

Managing Administrative Access

Core information security practices such as least privilege, separation of duties, and defense-in-depth are important to deny attackers access to environments.

Cloud Console Account Management

The VMware Cloud Console is the central management portal for VMware Cloud Services, and provides the ability to deploy, manage, and deprovision SDDCs, subscriptions, network connectivity, and other services like NSX Advanced Firewalling, VMware Aria products, and Tanzu Mission Control. By default, the organization's owner's Customer Connect account is granted access as part of the onboarding process.

Customer Connect accounts are managed by VMware and support multi-factor authentication through the use of a time-based one-time password (TOTP) application, such as Google Authenticator. An organization can also configure Enterprise Federation, allowing a SAML 2.0 Identity Provider (IdP) or a connection method supported by VMware Workspace ONE Access to handle authentication and authorization in the Cloud Console. This allows an organization to control access through existing account management processes. Additionally, any multi-factor authentication solution supported by the IdP can be used seamlessly.

API tokens can be generated by Cloud Console users, giving the token an equivalent level of access to their own user account. Organization-level applications can be defined by organization owners without connecting them to a user account.

Ideas to consider:

- Use Enterprise Federation to support Single Sign-on through an enterprise IdP.
- Require multi-factor authentication for all accounts with access to the VMware Cloud organizations. Carefully consider the use of source IP address restrictions in context of incident response and access. Consider using a "break glass" native VMware Customer Connect account with multi-factor authentication enabled in case of a loss of connectivity to the configured IdP, or a loss of access to the network that the access is restricted to.
- Consider using dedicated administrative accounts that are different from what the cloud infrastructure administrators use on their desktops. This helps prevent immediate lateral movement by attackers when an administrator's workstation has been compromised.
- Configure the allowed domains for Cloud Console accounts to prevent the addition of external users, either accidentally or maliciously, to the Cloud organization.
- Define policies for API token management that include token lifetime and key storage requirements. Regularly enable, review, and revoke OAuth Apps violations reports for tokens that do not meet the defined policies.
- Use organization-level application IDs for services connecting via API, to avoid sharing accounts and help enforce least privilege.

Role-Based Access Control (RBAC)

VMware Cloud Infrastructure products, from VMware Cloud down to the core vSphere, contain a robust set of permissions that can be configured as part of roles that users are assigned to. These permissions allow granular access to capabilities inside the VMware Cloud SDDC. The VMware Cloud Console also allows users to be assigned roles and permissions to manage their organization's assets.

Ideas to consider:

- Define groups for each role and grant access based on those groups.
- Follow a least-privilege model when assigning permissions to roles. Only assign the minimum permissions necessary for that user or system to do its job.

Virtual Private Network (VPN)

VMware Cloud VPN functions provide an encrypted end-to-end path over untrusted networks using IPsec. It can be used for connections across the open Internet, but also across a Direct Connect. Security is always a tradeoff, and IPsec VPNs trade security for performance, limited by available CPU and network capacity inside the SDDC.

IPsec VPNs rely on Path MTU Discovery, which in turn may require relevant ICMP protocol messages (IPv4 type 3, IPv6 type 2) to be permitted. This is a general best practice for networks, as blocking all ICMP messages to disable ICMP echo ("ping") causes the collateral loss of other important network messages like Fragmentation Needed, Time Exceeded, and more. Path MTU Discovery is important for automatic network optimization of most modern operating systems. Workarounds such as MSS Clamping add complexity and rigidity to an environment and may not be the best solution.

Deploying a VPN to connect to an SDDC involves other decisions about network topology and will depend on the network capabilities and topologies of the SDDC and other sites. Route-based VPNs use the BGP routing protocol to exchange information about networks between sites. This adds both complexity and flexibility, and the design of these networks is beyond the scope of this document. With simpler IP addressing schemes and network deployments the Policy-Based VPN options are possible. Layer 2 VPN connectivity allows for migrations into the cloud without re-addressing a workload, by extending an on-premises network, but requires the NSX Autonomous Edge appliance to be deployed in the local cloud.

VPNs between sites with dynamic addresses may require additional design considerations or operational process work. If the dynamic address changes then the VPN connection will not be functional until the SDDC is updated for the remote site's new public IP address.

Ideas to consider:

- Use IKEv2 with a GCM-based cipher with as high a bitrate as can support the required performance levels.
- Use Diffie Hellman Elliptical Curve groups (19, 20 or 21), with the highest group number of those that can support the required performance (generally based on the total number of tunnels).

- Enable Perfect Forward Secrecy where supported on both sides of the VPN connection. Enabling it on one side only may initially work but will disconnect after a preset amount of time.
- Use a long, randomly generated pre-shared key, or if available, certificate-based authentication.
- If the BGP endpoint is on a different device from the IPSec VPN, or there is a possibility of access to the BGP network being used, then a BGP Secret should be configured on both endpoints to prevent route hijacking.

Private Network Links

Direct Connect is an AWS solution where a network port on AWS's network is made available for customers to connect to.

In most cases, the port will be in a Point of Presence (PoP) datacenter facility where the end customer will order an MPLS WAN connection from their preferred carrier, who will assist with cross-connecting it to the port provided by AWS. Other configurations are possible, such as a Hosted Connection (a VLAN on a shared port) a Hosted VIF (a single virtual interface on a shared connection), and in some cases customers may collocate space in the PoP and run the cross-connect directly from their own equipment. All of these options provide different features, bandwidth, and cost models. Dedicated ports provide the most capability and highest bandwidth, including the possibility of using MACSEC to provide Layer-2 encryption between the AWS router and the customer router. Note that this can provide protection for a portion of the path but will require additional MACSEC or other encryption methods to provide end-to-end protection.

Ideas to consider:

- In order to minimize latency, select an AWS point-of-presence that your WAN provider can support, and is as close as possible to the sites that will be communicating with the SDDCs.
- Deploy multiple Direct Connect circuits to different points-of-presence for redundancy, that terminate in the same AWS account so that AWS knows they are for redundancy and will provision them on independent paths. Ensure that they have fully independent paths to the enterprise network.
- If multiple regions are being used for SDDCs, and latency tolerance is acceptable, consider deploying Direct Connects to different regions, and mapping them to a DX Gateway attached to an SDDC Group to provide redundancy against wider-area events while simultaneously providing connectivity to multiple regions.
- If possible, use MACsec encryption on the Direct Connect link to prevent packet interception on the wire.
- Use BGP secrets on all BGP sessions to avoid route hijacking.

Connected Accounts and Virtual Private Clouds (VPCs)

VMware Cloud configures native connections to the public cloud provider's networks and accounts to enable fast and secure access to public cloud services.

Every SDDC in VMware Cloud on AWS is connected to a VPC in a native AWS account owned by the customer. This connection is made by running a CloudFormation template provided by VMware that creates the necessary IAM roles in the customer account. Once those roles are in place, VMware will create and update the VPC, ENI, and route tables to establish and maintain connectivity. These IAM roles are necessary for proper SDDC operation, but there are other security controls that can also help manage the connectivity between the SDDC and connected AWS account.

Ideas to consider:

- Ensure only one CloudFormation Template (CFT) is used for each linked AWS account. Only the last successfully run CFT will be tracked by the VMware Cloud organization, and that will be used for any SDDCs deployed within that Organization and linked to that AWS account. However, once deployed, the SDDC will reference the AWS IAM roles, VPC, subnet, and main route table from that point in time. It will not automatically update them if a new CFT is run in that AWS account and Organization, which can result in different IAM roles being used by different SDDCs.
- The Lambda function created by the CFT is only used for the initial template deployment. It can be deleted once the linking is successful. Do not delete the entire CFT as it will remove the IAM roles as well, which are required for the operation of SDDCs.
- SDDCs will create Elastic Network Interfaces (ENIs) in the selected VPC & subnet upon their deployment. In some cases additional ENIs will be created afterwards, such as if the SDDC's Cluster-1 ever grows beyond 16 hosts. These ENIs will have the VPC's default security group (SG) attached to them. This security group operates as though the entire SDDC was an EC2 instance with that security group attached. For example, Outbound rules refer to traffic originating within the SDDC and going to native AWS service, and Inbound rules refer to traffic originating within the native AWS account and going to the SDDC. By default, this security group will allow all traffic from the SDDC, but traffic going to the SDDC must be manually added. Since the Compute Gateway firewall in the SDDC provides the same protection (using the Services Interface under its Applied To field), it is a viable option to allow all traffic through the security group and enforce protection through the compute gateway firewall alone. Both firewalls can be configured for the reduced traffic set, but this can make it operationally challenging to keep them in sync and does not provide meaningful security improvements in most cases.
- It is also possible to replace the default security group with a custom security group for all 17 ENIs created by the SDDC. This may cause operational challenges in case a new ENI is ever added, as the customer will need to monitor for that scenario and apply the desired security group immediately to avoid disruptions to network traffic.

Network Perimeter Controls

VMware Cloud has multiple network boundaries and perimeters that should be secured. The primary boundary is at the SDDC itself, consisting of dedicated sets of network segments for management and workloads. These network segments are separated from the network uplinks by an NSX Edge Gateway firewall. This firewall implements two different network gateways, one for SDDC management components, another for workloads and compute.

The VMware Cloud also employs the concept of an SDDC Group, which can extend the security perimeter to include multiple SDDCs, native VPCs, and Direct Connect Gateways, across multiple regions. The SDDC Group itself does not implement firewalling directly, relying on the individual SDDC gateway firewalls, VPC network ACLs, cloud provider security groups, and on-premises devices terminating connections from Direct Connect circuits. However, it can be considered an isolated zone, and should be treated as a managed network service, like an MPLS WAN.

Management Gateway

The gateway firewall is divided into two different policies, one which protects the management appliances in the SDDC (vCenter, NSX Manager, add-on service managers, etc.). It does not affect customer workload VMs and has a limited set of rules that only allow specific services through to each management appliance. It also allows creation of outbound rules from the management appliances, which always allow any service. The source or destination of every rule on the management gateway must be one of the management appliances. Arbitrary rule definitions are not permitted, nor are inbound rules with “any” as the source.

Access to management appliances can be via the private IP, allocated from the management network that was supplied during the SDDC provisioning process. Some appliances, such as the vCenter Server and HCX Manager, also have a public IP address automatically configured with destination NAT. These appliances register public DNS Fully-Qualified Domain Names that can be configured to resolve to either the private or the public IP address. Additionally, access to the NSX Manager can be through a reverse proxy accessible through links from the Cloud Console, allowing firewall rules to be managed “out-of-band.” This helps if an errant firewall rule denies access to the SDDC.

Ideas to consider:

- Ensure all rules allowing inbound access are restricted to the most specific set of source IP addresses and services required.
- Use private DNS resolution (and therefore access only over private connections) for the connections that offer public or private.
- Consider that the DNS resolution only changes the IP returned by DNS. It does not impact IP connectivity, and the public IP and NAT will always be in place, regardless of the DNS setting for vCenter Server, HCX Manager, and NSX Manager. Therefore, the source IPs for the firewall should still be configured to the minimal set of private IPs required, even when DNS is set to private resolution.
- Outbound traffic from the management appliances will follow the SDDC’s routing table, so if a default route is advertised, then outbound traffic will go through the connection advertising that route rather than using the SDDC’s native Internet connection, and therefore public IP. You will not be able to use the SDDC-assigned public IP when a default route is being advertised to the SDDC.

- Only groups can be referenced for the source and destination fields in firewall rules. Groups can only consist of IP addresses/CIDRs in the Management Gateway, and groups created here are separate from groups used by the compute gateway or DFW. Defining groups so that they are named clearly to represent the purpose and members is important to ensuring that the desired access is being defined by the rules.
- There is one exception to the above: traffic to/from ESXi hosts will NOT pass through the gateway firewall when a Direct Connect (DX) Private VIF (PVIF) is connected to the SDDC, or if the SDDC is a member of an SDDC group. In these specific scenarios, this traffic will always follow the DX PVIF or SDDC Group/vTGW path, regardless of the SDDC's route table and management gateway firewall rules.
- Logging should be enabled on rules necessary to track access, or attempted access. By default, logging is not enabled.

Management Appliance Access and Authentication

A deployed SDDC will have a number of appliances that manage different aspects of the infrastructure. These appliances are managed by VMware as part of the Shared Responsibility Model, and include vCenter Server, NSX Manager, and NSX Edge appliances by default. If enabled there may also be HCX Manager, Site Recovery Manager, Tanzu Kubernetes Grid Supervisor cluster, and vSphere Replication appliances.

All appliances are joined to the SDDC's Single Sign-on (SSO) domain, vmc.local. This SSO domain is local to the deployed SDDC and customers cannot create additional users in the SSO domain. Instead, they are provided with a single administrative account, cloudadmin@vmc.local . The cloudadmin account has restricted management permissions as part of the Shared Responsibility Model and is allowed to perform operations in support of workloads. Full administrative control of the SDDC is reserved for VMware itself.

The initial credentials for cloudadmin@vmc.local are displayed in the VMware Cloud Console. The password for this account can be changed through vCenter Server or its APIs/automation tools. Once the password is changed the Cloud Console will no longer display the correct password (it does not update from vCenter Server), so care should be taken to save it. The password is not recoverable, though VMware Cloud support can reset it via a support request.

A vCenter Server allows for the integration of an LDAP-based identity source which allows customers to use existing directories and authentication sources. Additionally, the vSphere Cloud Gateway Appliance can be deployed which allows linking the vmc.local domain to a local cloud vCenter Server's SSO domain. This permits the use of the on-premises SSO domain for access to the VMware Cloud infrastructure.

Ideas to consider:

- Use private DNS resolution for vCenter & HCX Manager so that these appliances are accessed from the on-premises network. SRM, vSphere Replication & NSX Manager only support private DNS and private IP connectivity, although NSX Manager can be accessed through the VMware Cloud console as well.

- Link an on-premises identity source to vCenter using either the Cloud Gateway appliance or an LDAP connection, to use existing accounts for access to vCenter.
- Adding individual user accounts to the Administrators group, rather than importing an Active Directory group, helps separate authorization from authentication, reducing attack vectors in case of Active Directory compromise.
- Use tiered access models where everyday tasks can be handled by regular accounts/group access, but any privileged access should use a separate account, individually added to the vCenter group.
- Reset the cloudamin@vmc.local account password using a PowerCLI script that automatically stores the password in your credential store, and only use it as a break-glass account when required for configuring new services that do not support service accounts (e.g. HCX) or when needed to make changes that other accounts to not have access. Rotate this password according to your password policy.
- If HCX has been enabled on the SDDC, remove any unused Public IPs (for example if HCX is being connected over a Direct Connect).
- Access to management components should not depend solely on IP address restrictions, as the compromise of an administrator's desktop often also includes the compromise of the administrator’s credentials. A bastion host or “jump box” solution may be implemented with multi-factor authentication. The Management Gateway firewall should then have appropriate restrictions on management services, allowing only the bastion host access. Appropriate hardening and monitoring should be applied to bastion hosts, including considerations for the compromise of an organization's central Active Directory or authentication source. Using separate administrator accounts is also recommended to help identify the presence of attackers. The compromise of an administrator’s regular desktop account would not automatically lead to the compromise of infrastructure and may force the attacker to generate login failures which can be monitored.
- Limit connectivity to the SDDC’s ESXi hosts for destinations using the services required:
- vMotion can be proxied through HCX for a controlled, secure channel.
- VM Remote Console access is proxied through vCenter Server. Direct access to ESXi hosts by VM administrators is not required nor desired. Workload administrators should access guest OSes using Remote Desktop console functionality, or through direct SSH to the guest OS. This helps simplify firewall rulesets and access control for both the workload and the infrastructure.
- IPFix data will originate from SDDC ESXi hosts, and traffic should be restricted through the on-premises firewall to only the IPFix collectors.
- Port Mirroring traffic also originates from the SDDC ESXi hosts in a GRE tunnel, and traffic should be restricted through the on-premises firewall to only the necessary ERSPAN destinations.

- vSphere Replication traffic will originate from the SDDC ESXi hosts and traffic should be restricted through the on-premises (or destination SDDC Management gateway) firewall to only the necessary vSphere Replication appliances where VMs are being protected.

Endpoint Security

VMware Cloud provides numerous ways in which workloads can be made resilient to security and other types of incidents.

Endpoint and Workload Security

Consider the ideas listed in this section.

- Ensure all rules allowing inbound access are restricted to the most specific set of source IP addresses and services required and avoid using "ANY" as the source or destination IP or services.
- Provide more general outbound rules at the perimeter but enforce specific outbound rules at the DFW.
- Use groups with dynamic membership and/or tag-based membership to simplify management.
- Include top-line rules to drop any traffic that should NEVER be allowed (for example traffic from a public IP source).
- Logging should be enabled on rules necessary to track access, or attempted access. By default, logging is not enabled.
- Always limit the Applied To field to the specific uplink the traffic is expected on, and avoid using "All Uplinks"
- Block traffic closest to the source (e.g. outbound traffic with the DFW, on-premises outbound traffic at the on-premises FW)
- When using NAT, ensure that only the ports required for the NAT are included in the NAT rule, and ensure the NAT matching criteria is set correctly for the use case. If the NAT matching criteria is set to private IP, then it will not be possible to differentiate between traffic that has been NATted and traffic that originated internally.
- If a NAT rule is configured for ANY services, then that NAT rule will be also be used for outbound (SNAT) traffic for Internet traffic from the private IP specified.
- NAT will not match unless the traffic is routed out the SDDC's native internet (e.g. NAT cannot be used when a default route is advertised from one of the uplinks to the SDDC).
- Traffic between Management VMs and customer VMs in the same SDDC do not require Compute Gateway rules, but still must be allowed by the Management GW firewall.

Microsegmentation

The NSX Distributed Firewall is included with every VMware Cloud on AWS Software Defined Datacenter (SDDC). This firewall provides microsegmentation capabilities by inspecting and controlling traffic at the VM network interface. Unlike a traditional firewall, this allows control of network traffic between workloads on the same network segment, as well as from other sources.

The Distributed Firewall can be configured using a variety of rule types, from traditional rules to dynamic groups that allow policies to be applied based on tags, VM names, or other workload properties. Rules can also be applied to specific objects allowing for scoped policies, including default rules that only apply to specific VMs. Limiting traffic between VMs makes it much harder for attackers to move laterally, and the flexibility of rule definitions means that rules can be very specific but also easily updated when environments change.

Ideas to consider:

- The Distributed Firewall is IP-based, so dynamic objects are translated into IP addresses, using the IP addresses detected by VMware Tools or through traffic snooping. Dynamic membership cannot be “Applied To” IP addresses.
- Much as with traditional firewalls, the complexity and scope of rules impact performance as each packet is evaluated against each rule, though the Distributed Firewall can take advantage of additional CPU as clusters scale out. It is recommended that rules be limited in scope, such as to a particular network segment, and global rules be considered carefully before implementing.
- Where possible use groups to define rules so that changes are easier and updates less susceptible to human error. Create nested groups to aggregate similar rules. For example, rather than having one group for all cloud administrators, consider creating a group for each cloud administrator, then aggregating that into a “supergroup.” If an administrator leaves your organization it is easier to find and remove their access group.
- Consider defining and documenting a naming strategy for groups so that similar items are grouped together, and data is sortable and easily filtered.
- Consider defining and documenting a firewalling strategy, such as default allow, default deny, or a documented combination using “Applied To.” Do the same for rules, such as per-application or per-service, so that there is consistency.
- Enable and implement Distributed Firewalling, so that workloads must have specific distributed firewall rules at all times. Limit generic rules implemented at the gateways.
- Restrict outbound traffic using DFW to meet the standard approach of blocking traffic closest to the source.
- Apply IDS/IPS to outbound connections to look for known Command & Control access and common attack signatures.

- Consider using a proxy server for filtering & inspecting web traffic if virtual desktops are being hosted in the SDDC. Using this model, Internet access would be blocked for endpoints, and only the proxy would be permitted to access the Internet, providing additional URL filtering based on real-time updated lists and/or other identifying capabilities such as geo-location, site categorization, etc.
- NSX L7 firewall from the Advanced Security add-on can be used to ensure SSL/TLS connections are using encryption methods that meet minimum standards to avoid known attack vectors.

Restrict DNS traffic destined for Internet-based DNS servers and require all workloads to use internal DNS servers that are managed and patched. Log all queries and block or check for requests of known C&C or malicious domains using lists that are updated frequently.

VMware Tools

VMware Tools are an important component for virtual machines, supplying drivers for paravirtual devices like the vmxnet3 network interface and the pvscsi virtual SCSI controller, as well as a communications channel between ESXi and the guest operating system. That communications channel is important, as it can ensure that guest operating systems and workload applications shut down gracefully when needed. It will also help the infrastructure detect when virtual machines have booted correctly, as part of vSphere HA actions should a cloud host fail.

VMware Tools is a software package that, like all other software packages, requires updates and maintenance. Include it in your configuration management systems or enable “Check and upgrade VMware Tools before each power on” in the VM settings to have them automatically updated on Microsoft Windows guests.

The Windows drivers for virtual machine hardware have been added to the Windows Update repositories, so that Microsoft Windows operating systems will automatically download and install the latest versions if automatic driver updates is enabled. If your organization manages Microsoft Windows updates using Windows Server Update Services (WSUS) ensure that these driver updates are configured as part of what is presented to client systems.

Linux vmxnet3 and pvscsi drivers are incorporated into the upstream Linux kernel sources. The other components that manage the hypervisor-to-guest communications are part of the open-vm-tools package supplied by VMware to Linux distribution maintainers. The open-vm-tools package is updated when you patch and update your Linux guest operating system.

VMware Tools also allows an SDDC to automatically detect the IP address of a workload, for use in dynamic firewall rules.

In-Guest Controls

Security controls inside workloads are the responsibility of the customer in the Shared Responsibility Model. As discussed earlier in this document, we often suggest that organizations explore using configuration management tools like SaltStack to apply and audit configuration settings on workloads. This has benefits of saving time and ensuring security consistency, but also simplifying template management.

Ideas to Consider:

- Monitoring and configuration management systems are two examples of systems that have privileged access to an organization's workloads. High-profile attacks have demonstrated that these types of systems are targets for attackers to breach, allowing them to move laterally throughout the organization with ease.
- Are there sufficient controls protecting other guests from monitoring breaches?
- Does your organization use change control and source code control techniques to manage and test changes to configurations? How will you know if an attacker has gained access to that system?
- Workloads deployed in a Kubernetes environment require some additional considerations to prevent containers from gaining too many permissions on the node's operating system. Without protections in place, containers may be able to access host resources such as processes, volumes, or network access. To account for this, consider using Kubernetes admission controllers to prevent unwanted access to the host operating system.

In-Guest Data-at-Rest Protections

VMware Cloud uses vSAN Data-at-Rest encryption to store data in a public cloud provider's storage. It is possible to use in-guest encryption technologies like Microsoft BitLocker and Linux dm-crypt to protect workloads. This has performance impacts, given the double encryption (BitLocker plus vSAN Encryption), and defeats space efficiency processes like deduplication and compression (your virtual machine will consume its entire allocated disk space). In general, VMware does not suggest using in-guest encryption, but for some very sensitive workloads like Microsoft Active Directory it may be a suitable additional layer of defense. Use it sparingly due to the performance impacts and management overhead.

Third-party integrations, such as Amazon FSx for NetApp ONTAP, may have different encryption and performance considerations depending on the features and capabilities present in those solutions.

Virtual TPM

Workloads can benefit from the use of virtual Trusted Platform Module (vTPM) available in VMware Cloud on AWS 1.19 and newer. The addition of a vTPM presents a TPM 2.0-compliant device to the guest OS, for use by the guest OS and workloads as they see fit, just as the workload would use a physical TPM when running on physical hardware.

Virtual TPMs use VM Encryption to protect data on disk, encrypting just the VM "home" files, but not the entire VM. VM Encryption is enabled with Native Key Provider, a feature within VMware Cloud that manages encryption keys without requiring an external key management system (KMS).

Storage Policies

As discussed in the Infrastructure Design section, virtual machines can be assigned different vSAN storage policies which have an impact on performance and storage usage. Reviewing these policies and ensuring they match your organization's risk tolerance and use of VMware Cloud on AWS SDDCs is important.

These policies can be customized on a per-VMDK basis, but in general simpler is better. If complex policy setups are needed it is suggested that they be automated, for auditing and reconfiguration purposes.

Multicast

L3 Multicast is not supported (e.g. PIM, IGMP snooping). However, and L2 multicast traffic is treated as a broadcast and sent to all ports on the network segment.

This enables applications that use multicast to communicate in the same network segment but does not support the optimization of having the network send traffic only subscribed devices.

Workload Resilience

VMware Cloud offers many of the same resilience features found in local cloud versions of vSphere and Cloud Foundation.

This includes snapshots, clones, replication, as well as vSphere High Availability, vMotion, and the Distributed Resource Scheduler (DRS).

Ideas to consider:

- Ensure that workload applications start automatically when the virtual machine boots. This helps immensely for regular and automated patching, but also as part of incident response. For example, if a cloud host fails vSphere HA will restart the workloads on other cluster hosts. If the workloads automatically start the need for off-hours administration work is reduced, pushing it to normal working hours.
- Ensure that workloads spanning multiple virtual machines or containers are resilient to restarts on components, either because of patching or from a vSphere HA automated restart. Applications should employ techniques to retry connections periodically. Use of the NSX Advanced Load Balancer can help make internal application subcomponents more reliable, as well as detect application health and present customized outage pages to customers.
- Use DRS affinity and anti-affinity rules to separate clustered components from each other, reducing the impact of a host failure.

Conclusion

Security is a broad topic that impacts all parts of an organization, and is often best treated as a process that everyone participates in. Security is a source of tension in an organization, between the need to move forward with new work and techniques, and the need to pass compliance audits which verify the security controls in place.

Organizations that have traditional on-premises deployments have to rethink how they operate in the public cloud, as the public cloud is significantly different.

This is where VMware Cloud offerings shine, as they bring the elasticity and options of the public cloud, but in a way that operates just like traditional VMware environments. This familiarity makes it easy to maintain operations, compliance certifications, and security while your organization grows. Hopefully this guide has helped you consider changes to your architectures and methods to embrace these new models, giving your organization more options while making it more resilient to incidents, small or large.

Operate Pillar

5

This chapter includes the following topics:

- Operate Introduction
- Infrastructure Operations and Service Control
- Automation
- Cost Management
- Sustainability

Operate Introduction

The purpose of the Operate pillar is to outline the guidelines and considerations for operating VMware Cloud infrastructure.

Introduction

VMware Cloud brings VMware's enterprise-class Software-Defined Data Center (SDDC) software to the public cloud, enabling you to operate your production applications and workloads across VMware vSphere®-based private, public, and hybrid cloud environments. The solution integrates VMware's flagship compute, storage, and network virtualization products (vSphere, vSAN, and NSX) along with vCenter management. This solution is then optimized to run on elastic public cloud infrastructure.

With the same architecture and operational experience as on-premises and in the cloud, your IT teams can now quickly derive instant business benefits from use of the VMware Cloud and hybrid cloud experience.

Infrastructure Operations and Service Control

With the adoption of VMware Cloud, it's expected that several of the organization's existing operational processes will remain consistent and intact in the public cloud. This is part of the value proposition of the service. It is critical for an operations team to understand the VMware Cloud management domains in detail to cohesively evaluate the capabilities and identify any nuances or new areas that could introduce operational gaps.

Day-to-day Infrastructure Operations

The day-to-day management of VMware Cloud will typically focus on the following areas, Lifecycle Management, Policies and Processes and vCenter Management.

Lifecycle Management

Updates to the SDDC software are necessary to maintain the health and availability of the VMC service. The VMware Cloud provider will share notifications of upcoming SDDC lifecycle management activities. Customers will monitor, review, and manage the patching and upgrade schedule to the SDDC (review release notes and provide forward looking change notifications to internal teams). This includes vSphere ESXi, vSAN, NSX, and VMware management components.

Review Policies and Processes

The organization must review existing policies and processes for lifecycle management and how they can be adapted for VMware Cloud. This includes the scheduling of VMware Cloud patching and updates, version control between on-prem and VMware Cloud for compatibility, identification of components not included in the scheduled VMware Cloud patching processes and defining validation processes applied before and after patching has occurred.

vCenter Management

VMware Clouds are provisioned with VMware vCenter Server. After deployment customers are provided with credentials to manage the core vCenter Logical Constructs (Folder Structures, Alarms, Tags etc), Roles and Responsibilities, and the vCenter Content Library.

Considerations

Review existing processes for managing vCenter and how this can be adapted for VMware Cloud. This includes the management of Content Libraries, core vCenter logical constructs (Datastores, Alarms, etc) and vCenter integration between on-prem and VMware Cloud (if supported).

Control Plane Management

The control plane plays a primary role in acting as the interface between applications and service delivery. This area will include event data collection, resiliency & backup strategy, configuration and access management, as well as Day 2 operations.

Considerations

Review existing processes for control plane management and how they can be adapted for VMware Cloud. This should include core VMware based control plane technologies (vROPs, vRLI, vRNI, HXC, vRA) that will be integrated with the VMware Cloud SDDC. Consider the following associated processes; lifecycle management, control plane backups, resiliency, and event management.

Compute & Workload Management

This area focuses on the processes and policies for managing compute resources and deployed workloads. This includes VMs, Containers, Resource Pools, vApps and Compute Management policies.

Considerations

- Define VMware Cloud processes for managing vSphere clusters, including host provisioning, child resource policies (naming, shares, reservations, reservations, limits), and integration with core IT Service Management systems (If required). This should also include the decisions & process to trigger compute scaling.
- Review existing processes for workload management and how they could be adapted for VMware Cloud. This includes VM templates, Container images, snapshots, clones, OS patching, licensing compliance and considerations, antivirus, VM tool management and any migration considerations.

Storage Management

This area focuses on the management of datastores, storage encryption, storage monitoring & optimization, storage performance & capacity management, and operationalizing vSAN within VMware Cloud.

Considerations

- Review existing processes for managing storage, including vSAN storage policies, encryption, storage performance approach, vSAN threshold requirements, and proactive capacity management.
- How will you monitor vSAN storage usage and capacity to comply with the service SLA?

Network Management

Management of NSX, configuration of network segments, IP address management, management of network security policy, and monitoring/managing network traffic levels.

Considerations

Review existing network and security processes and how they can be adapted for VMware Cloud. This includes network and security architecture/design, security models (current and cloud optimised), network stretching, security requirements, micro-segmentation, distributed firewall, IP address management, roles/responsibilities and lines of demarcation (between your teams and vendor ecosystem), end to end traffic flow monitoring, proactive bandwidth optimisation and load balancing requirements.

Performance & Capacity Management

Performance and Capacity are tightly linked as they both ensure your workloads get the necessary resources to perform optimally.

Consider the following topics for adopting a successful strategy:

- Identify key stakeholders and confirm the scope and metrics for capacity and performance (Compute, Storage, Network). This will allow you to then effectively monitor the utilization of your workloads and identify thresholds for scale-out expansion.
- Regularly collect performance data and periodically review the metrics. By identifying cyclical patterns, you can recommend cloud infrastructure scaling ahead of demand. This would not just lead to effective operations of your workloads but also to potential cost optimization as you build synergy between your workload elasticity and VMware Cloud purchasing strategy.
- Identify key tools/technologies that will be used to monitor and manage performance and capacity. Make sure that these tools are capable of operating with the permission set provided within the VMware Cloud deployment.
- Whenever possible, establish baseline performance before, during and after migration to VMware Cloud. This helps to isolate any performance issues that might arise post-migration.
- Establish roles and responsibilities for VMware Cloud Performance management ahead of any planned migrations.

Availability Management

Availability management plays a lead role in ensuring your services can perform their agreed functions to meet the needs of the business.

The availability of your services will depend on the percentage of time that your workload is available. This percentage (such as 99.99%) will be reflected over a period and is often a design goal for applications.

As you operate your VMware Cloud workloads, consider the following topics to ensure that you have established availability processes extended to the SDDC:

- Develop a plan for continuously balancing/distributing applications/workloads across available cloud resources. Utilize anti-affinity policies whenever necessary.
- Work with application teams to ensure that applications are architected with the availability model of the cloud in mind.
- Review VMware Cloud Availability Commitments to ensure that these are well understood by your operations teams.

The Hosting Reliable Applications on VMware Cloud whitepaper provides high level strategic guidance on the design of highly available and reliable VMware Cloud infrastructure.

Infrastructure Observability & System Health

The goal of observability is to understand a complex system's internal state by observing its external outputs.

Overview

Proper instrumentation enables you to aggregate metrics, traces, logs and events from a distributed system and correlate them across various application components and services, identifying complex interactions between elements and allowing you to troubleshoot performance issues, improve management, and optimize cloud native infrastructure and applications.

Underlying Hardware Infrastructure & VMware Control Plane

VMware Cloud was designed from the ground up to be simple to consume, allowing your operations teams to focus higher up the stack and away from the undifferentiated heavy lifting associated with hardware infrastructure.

This means, the management, health, and lifecycle of the underlying hardware infrastructure (compute, network, and storage) is the responsibility of the VMware Cloud Provider. This includes lifecycle management of the VMware component stack and the adding and removing of physical hosts for scaling and maintenance purposes.

From your perspective as a consumer of the platform, your operational responsibilities and overall observability capabilities now begin at the VMware software layer, which includes ESXi, NSX, and vSAN. Capturing the appropriate metrics from these components in addition to your enterprise workloads will form the building blocks for achieving full-stack observability.

Common metrics

When deciding on the metrics that need to be observed, it’s important to adopt a user centric approach that works backwards from application owners.

The goal should be to collect the minimum number of data points necessary to implement observability in the most efficient possible manner. Choosing more metrics than necessary and you could experience alert fatigue and lower attention towards the statistics that matter. In contrast, not selecting enough metrics would be counter intuitive as it leads to lack of visibility and overall inability to examine significant behaviours.

This section will outline key considerations when building an observability plan for your VMware Cloud infrastructure. It is advised to think about your observability plan when you are in the pilot or pre-production stage of your cloud journey. Consider the following high-level guidelines:

- Shift towards an SLO centric culture to observe your services based on critical end-user experience rather than system metrics. Ensure VMware Cloud monitoring and event metrics/thresholds are aligned to Service Level Requirements and SLOs that are documented in Service Level Agreements with the service consumers (i.e., LOBs)
- Define and select the appropriate Infrastructure and Application metrics to create SLIs that help you achieve better system observability.

- All key thresholds and metrics formally established and reviewed regularly. The review process is documented and formally established, and reviews are fully aligned to service level requirements, and they support business commitments. There is a well understood and documented understanding of the bidirectional impact of VMware Cloud in addition to the future planning of new KPIs/Metrics to drive further efficiencies and user experience.
- Review your existing processes and tools used for monitoring and event management and how they could adapt to VMware Cloud i.e., Predictive analytics, guided troubleshooting, root cause analysis as well as policy-based, automated remediation capabilities. This will proactively protect you against degradation of performance and capacity.

Workload Health

Workloads operating in the VMware Cloud need to consistently instrument the applications to emit metrics, logs, and traces so that the signals can be correlated to identify the root cause of any issue. These issues could relate to inaccessibility, operating system (OS) instability, application misconfiguration, or any number of other possibilities.

A well-designed system aims to have the right amount of observability that starts in its development phase. Don't wait until an application is in production before you start to observe it. This includes the setup of monitoring, alerting, and logging so that you can act based on the behaviour of your system.

Questions to consider when choosing instrumentation for VMware Cloud observability:

- What tooling will be used to monitor VMware Cloud and manage related events? Is this a new tool or will something existing be adapted?
- Do you require a system that supports multi-clouds, including on-premises?
- Is there an egress cost for sending data to the observability system?
- Should the system provide support for multiple regions?
- Should the system scale out on-demand for capacity?
- Should the system support multi-tenancy with separation of teams?
- Should the system include AI-powered intelligence to facilitate AIOps as you evolve your observability practice?
- Does the system need to support 3 rd party integrations such as PagerDuty, ServiceDesk, DataDog, Slack and VictorOps?
- Does the system need to provide immutability (data/logs/metrics), which cannot be modified, deleted, manipulated? Access control is required.
- Does the system support scraping metrics from modern apps, or does it require an agent to be installed?

Automation

VMware Cloud is an open, extensible platform that provides several ways to integrate, extend, and automate IT workloads across VMware products and services. As automation has become increasingly important for hybrid cloud infrastructure and the expectation is to have an easy and fast way to automate and learn about new features of hybrid infrastructure, VMware Cloud provides a seamless developer experience across the entire platform with developer tools and automation tools available at your fingertips.

Introduction

The automation of a process is based on series of criteria and logics, in a form of human-written codes or Machine Learning models, that consume metrics, events, and alerts as source of data for decision making.

Benefits in Automation of operational tasks

Automating a process comes with an initial cost in terms of time and resources, that should be analysed in terms of return of the investment when achieving some of the key benefits.

- Reduce time to perform an action
- Reduce risk in performing the action
- Increased human capacity for further innovation
- It increases productivity, reliability, and performance
- It makes auditing easier or even possible in first place
- Automation is work force multiplier

CI/CD Pipeline & Automation

Automation and continuous monitoring through all phases of app development are at the heart of DevOps and agile methods. A CI/CD pipeline is one of the most important assets for building, testing, and deploying modern applications. Continuous integration and continuous delivery forge a connection between a developer committing code and the delivery of new functionality to applications in production. Between the endpoints in a CI/CD pipeline, the code is built, verified with multiple tests, checked against audit and security controls, prepared for deployment, and, in some cases, automatically deployed to production.

Reliability and repeatability are key aspects of CI/CD that require automation.

We recommend to:

- Question every manual task when possible; Implementing automation from the start is easier than performing major revamps to clear technical debts across multiple components of a distributed system.
- Formulate an automation strategy that best suits the teams and the technology stack.
- Automate low-effort, high-value tasks first.

Continuous Assessment and Optimization

Workload rightsizing

Workloads on clouds are set to be as efficient as possible. However, common problems with the operation of virtual systems arise over time. Individual virtual machines (VMs) that run the workloads are supposed to have the right levels of processing power, memory, and disk space. Sometimes, though, they do not use what has been allocated to them, which results in inefficiency.

In workload rightsizing, cloud administrators assess the virtual CPU processing power allocated to a workload. They also look at the assigned RAM and virtual disk space, and what the workload uses.

Although administrators can use manual workload rightsizing, there are software tools available that provide this kind of administration. These tools can either suggest manual pre-provisioning of resources or, in some cases, change the provisioning automatically.

Rightsizing is important for a VM. Here are some benefits:

- The processes inside the Guest OS may experience less ping-pong. The Guest OS may not be aware of the NUMA nature of the physical motherboard, and think it has a uniform structure. It may move processes within its own CPUs, as it assumes it has no performance impact. If the vCPUs are spread into different NUMA node, example a 20 vCPU on a box with 2-socket and 20 cores, it can experience the ping-pong effect.
- Lower risk of NUMA effect. Lower risk that the RAM or CPU is spread over a single socket. Due to NUMA architecture, the performance will not be as good.
- Lower co-stop and ready time. Even if not all vCPU is used by the application, the Guest OS will still demand all the vCPU be provided by the hypervisor.
- Faster snapshot time, especially if memory snapshot is included.
- Faster boot time. If a VM does not have a reservation, vSphere will create a swap file the size of the configured RAM. This can impact the boot time if the storage subsystem is slow.
- Faster vMotion. Windows and Linux use memory as cache. The more it has, the more it uses, all else being equal.

Workload Optimization

Workload Optimization provides for moving virtual compute resources and their file systems dynamically across datastore clusters within a data center or custom data center.

Using Workload Optimization, you can rebalance virtual machines and storage across clusters, relieving demand on an overloaded individual cluster and maintaining or improving cluster performance. You can also set your automated rebalancing policies to emphasize VM consolidation, which potentially frees up hosts and reduces resource demand.

Workload Optimization further enables you potentially to automate a significant portion of your data center compute and storage optimization efforts. With properly defined policies determining the threshold at which resource contention automatically runs an action, a data center performs at optimum.

Capacity Planning

VMware Cloud Sizer is a complimentary VMware Cloud service that estimates the resources required to run various workloads within VMware Cloud.

In addition, the VMware Cloud Services Portal includes an integrated user interface for the sizer to make the process even easier to navigate.

VMware Cloud Sizer is responsible for estimating the resource utilization for any VMware Cloud deployment. The VMware Cloud Sizer currently supports VMware Cloud on AWS.

Estimating the resources required to host a given workload within VMware Cloud is a non-trivial task largely dependent on the presented data. The service can accurately calculate project utilization and requirements with the data although the data is not always available. Therefore, the service supports several different input methodologies and sizing workflows. The VMware Cloud Sizer tool has three different sizer options.

- Quick Sizer
- Advanced Sizer - Import
- Advanced Sizer - Manual

Each of the sizer options provides you with an accurate estimation that is required to run your workload. The quick sizer is helpful in providing a rough estimation whereas an Advanced Sizer - Manual provides advanced accurate information in detail for your new deployments. For more information, see Access VMware Cloud Advanced Sizer - Manual .

The VMware Cloud Sizer provides a mechanism to sign in to your organization and access the different sizer options to calculate your workload estimation. The tool also provides an additional feature of creating and saving your entries as a project for your reference. For more information on creating a new project after signing into the sizer tool, see Create and Save Your Sizer Project in VMware Cloud Sizer .

OS & Application Patch Management

Organizations today are expected to support thousands of workloads across a wide range of device types.

Identifying and patching security risks across different applications and operating systems is challenging, particularly when not using a unified platform.

Cost Management

Cloud cost management helps businesses control their spending on cloud services while also maximizing their resources.

Introduction

Most cloud providers offer basic cloud cost management tools to help achieve cost management, and there are also more specialized third-party solutions that provide additional visibility and insight into cloud costs.

By making cloud cost management a priority, an enterprise can control its costs and practice good governance while also ensuring that it has the cloud resources it needs to stay competitive.

In addition, cloud cost management best practices also support other business objectives and cloud best practices, such as security, visibility, organization, and accountability. Thus, cloud cost management is important for reasons beyond simple cost control. Good cloud cost management gives businesses the ability to plan, reduce waste, and forecast both their costs and their resource needs.

Defining Cloud Cost Management

Cloud cost management (also known as cloud cost optimization) is the organizational planning that allows an enterprise to understand and manage the costs and needs associated with its cloud technology. This means finding cost-effective ways to maximize cloud usage and efficiency.

As cloud infrastructure becomes more complex, cloud costs become more difficult to track. The “pay for what you use” model used by most public cloud providers adds to the difficulty: If usage is monitored and managed appropriately, this model can result in significant savings, but it’s also easy for costs to spiral out of control. This is especially true if decision making is decentralized across an organization, with individuals able to spin up instances (and accrue costs) with little or no accountability. Thus, it’s important for enterprises to employ a cloud cost management strategy to make the most of their infrastructure and keep costs down.

Advantages of cloud cost management:

- **Decreased costs:** This is the most obvious benefit of cloud cost management. Businesses that take a proactive approach to planning for cloud costs can ensure they don’t overspend on unused resources, and they’re able to take advantage of discounts based on volume or advance payment.
- **Predictability:** A business that properly forecasts its cloud computing needs won’t be surprised by a sudden increase in costs.
- **Efficient usage:** Taking a close look at spending also helps enterprises reduce waste and make the most of the resources they do pay for with techniques like automatic scaling and load balancing.
- **Better performance:** An important cloud cost management tactic is right-sizing or ensuring that the public cloud instances you choose are the right fit for your organization’s needs. Overprovisioning means overpaying; under provisioning can cause performance to suffer—but with careful planning, businesses can ensure smooth performance without increasing costs.

- **Visibility:** It's impossible to practice good cloud cost management without detailed visibility into your organization's usage and cloud architecture. Fortunately, this visibility also serves many other business needs besides cloud cost management, including governance and security.

Sustainability

Enterprises worldwide, across geographies and industries, are pursuing the imperative of digital transformation, which enables them to stay ahead of their competition through business agility and operational efficiency. This is becoming a non-negotiable imperative as organizations face relentless competitive pressure from both traditional rivals and digital upstarts.

Introduction

Successful digital transformation journeys are built on the foundations of modern infrastructure paradigms such as virtualization, cloud computing, containers, serverless infrastructure, and innovations like artificial intelligence (AI)/ML technologies.

Cloud and virtualization technologies have served as the bedrock on which organizations of all sizes have modernized their datacentre environments. By virtualizing compute, storage, and networks, organizations can transform to modern software-defined datacentres that employ a cloud operating model for better agility, flexibility, utilization, and scalability

The goal of cloud computing, and the virtualization that underpins it, is about minimizing the energy and carbon associated with running workloads. And with that adding to sustainability goals that enterprises have nowadays.

The goal of this subject is how the VMware Cloud Well-Architected Framework can help as a solution to become more sustainable. The following section outline strategies that can help to achieve the reduction of energy and carbon associated with running workloads on top of the platform.

Achieving Workload Energy Efficiency

Workload energy efficiency minimizes the energy required to run workloads hosted on IT infrastructure housed in datacentres.

There are four components to achieving workload energy efficiency:

1. Making energy visible
2. Maximizing productive host utilization
3. Designing compute-efficient applications

Making Energy Visible

For a host (server), energy is an intrinsic characteristic reflecting the extent of use by workloads of its compute resources, such as CPU, memory, and disk. Similar to improving the energy efficiency of our homes, making container and host energy visible enables benchmarking that we can act on. Adding that visibility informs strategies for management and optimization.

VMware Aria Operations has the ability to make energy visible:

https://blogs.vmware.com/management/2021/10/sustainability-dashboards-in-vrealize-operations-8-6.html

Maximizing Host Utilization

Before virtualization, the best practice was to run one application per physical server. In other words, servers typically ran at only 5-15% utilization. This gross underutilization translated into massive energy waste — incurring both financial and environmental impacts. Virtualization enables higher server utilization, which enables more consolidation. This drastically reduces global datacenter electricity consumption. However, because many servers today are running at only 20-25% utilization, there is still significant room for improvement.

Key opportunities for innovation include:

1. Enabling “cloud-sharing” that puts spare capacity to productive use by transient and non-time-sensitive workloads.
2. Recouping stranded capacity from oversized virtual machines, containers, and servers that no longer do useful work (sometimes called “zombies”).
3. Leveraging hybrid public cloud bursting to provide on-demand peak and backup capacity, enabling customers to reduce on-premises infrastructure and run it with higher utilization.

These innovations would produce productivity and sustainability improvements, while also meeting performance and availability requirements.

Designing Compute-Efficient Applications

Compute-efficient applications are a focus of an emerging practice of sustainable software engineering , in which applications are designed, architected, coded, and tested in a way that minimizes the use of CPU, memory, network, and storage. Mobile-phone applications are good examples of this.

Mobile phones have limited power, so the best-designed apps are built to minimize battery consumption . The Green Software Foundation has a working group to research and develop tools, code, libraries, and training for building compute-efficient applications. It also has a working group that’s developing a Software Carbon Intensity Specification to help users and developers make informed choices of their tools, approaches, and architectures.

Achieving Workload Carbon Efficiency

Integrating electricity carbon intensity as an optimization factor into workload management can significantly minimize system carbon emissions.

Workload Placement and Scheduling

A less-obvious component of workload carbon efficiency is placement and scheduling – when and where workloads are run.

A characteristic of the electricity that powers datacenter workloads is carbon intensity — the weighted average of the carbon emitted during the generation of that electricity across all generators on the grid. Carbon emissions can vary anywhere from near-zero for wind, solar, hydro, and nuclear power plants to very carbon-intensive for coal and natural gas power plants (e.g., 500 kg CO2/MWh). The mix of generators contributing electrons and the quantity generated on the local grid varies at any given time. Therefore, a grid's carbon intensity varies over time.

For workloads that are not latency-sensitive and/or geographically restricted, the management system may determine when and/or where to run these workloads based on when and/or where the electricity is cleanest. For example, the management system can delay running a workload or run the workload in an alternate datacenter. This idea isn't far-fetched. The share of renewables and low-carbon electricity reached almost 55% in 2019 for global electricity generation. In aggregate, workload placement and scheduling could help reduce demand for carbon-intensive electricity. In the longer term, managing datacenter workload demand could also improve the economics and stability of the electricity grid by facilitating the balance of electricity demand and supply.

Carbon-aware Workloads

Carbon-aware workloads are necessary for enabling workload placement and scheduling to optimize system carbon emissions.

Quality-of-service requirements such as latency, geographic restrictions, and mission-critical elements of these workloads can be communicated back to the management system. This enables the management system to identify and prioritize workloads that have the flexibility to alternate their scheduling and/or placement. The Green Software Foundation has a working group focused on developing an SDK to enable carbon-aware applications.

As we can see, there are pathways to zero-carbon clouds that can help accelerate the coming transition to a low-carbon economy. Innovations that maximize the productive use of cloud infrastructure will bring significant economic and environmental benefits. And managing workloads to use the cleanest energy can help stabilize the grid and provide lower-cost electricity. Some of these innovations can leverage existing capabilities. Others will require the maturation and adoption of emerging capabilities, such as hybrid cloud bursting to provide on-demand capacity for peak loads.