Reference Architecture
Tanzu Kubernetes Grid with VMware Cloud Director Platform and Container Service Extension 4.0

vmware

Table of contents

List of Tables

List of Figures

Introduction

VMware Cloud Director (VCD) is a cloud service delivery platform that enables Cloud Service Providers-Cloud Builder (CSP-Cloud Builder) and enterprises to build, manage, and operate secure and scalable multi-tenant cloud environments. As a key component of VMware's cloud infrastructure suite, Cloud Director provides a comprehensive set of capabilities for creating and managing virtual data centers, networks, and storage resources. It empowers organizations to deliver Infrastructure-as-a-Service (IaaS) and other cloud services to their customers and end-users, while maintaining control, security, and compliance. Some of the core capabilities of VMware Cloud Director are as follows:

1. Multi-Tenancy: Cloud Director enables CSP-Cloud Builder to create isolated virtual data centers for multiple tenants, each with its own set of resources, policies, and permissions while leveraging vSphere infrastructure, NSX Networking and NSX Advanced Load balancing stack.
2. Resource Pooling and Allocation: CSP-Cloud Builders can aggregate compute, storage, and network resources from one or more vSphere Infrastructure and NSX Networking stack into resource pools and allocate them to tenants based on their needs.
3. Self-Service Portal: Cloud Director provides a self-service tenant portal that allows tenants to provision and manage their own VM and/or Container workloads, networks, and storage resources and Load Balancing services.
4. Automation and Orchestration: VCD automates the provisioning and management of cloud resources, reducing manual effort and improving operational efficiency by providing API, and integration with Terraform.
5. Extensibility: VCD offers a rich set of APIs and integration points, allowing CSP-Cloud Builders to extend and customize the platform to meet their specific requirements.
6. Value added Extensions: VCD offers value added self-service extensions such as App launchpad, to launch VM and Container applications, Object Storage Extension to provide S3 Object Storage and Kubernetes cluster backup services, VMware Data Solution Extension to provide On-demand messaging and database services on CSE clusters using VMware Data Solutions to the tenants.

As CSP Builders' customers undergo digital transformation, modern workloads such as containers and microservices are becoming increasingly important. It is important for CSP builders to offer greater agility, scalability, and portability, enabling organizations to rapidly develop and deploy applications in response to changing business needs. Kubernetes, an open-source container orchestration platform, has emerged as the de facto standard for managing containerized applications at scale. To help organizations fully leverage the benefits of Kubernetes, VMware offers Tanzu Kubernetes Grid (TKG), an enterprise-ready Kubernetes runtime that streamlines the deployment and management of Kubernetes clusters across on-premises, public cloud, and edge environments. TKG is part of the VMware Tanzu portfolio, a suite of products and services designed to empower organizations to build, run, and manage modern applications on Kubernetes. Following are some of the differentiators for CSP-Cloud Builders to consider when offering Kubernetes as a Service with TKG to their customers:

1. Consistency: TKG provides a consistent Kubernetes runtime across different infrastructure platforms, including vSphere, VMware Cloud on AWS, and public clouds such as AWS, Azure, and Google Cloud. This consistency simplifies operations and ensures a uniform experience for developers and operators.

2. Lifecycle Management: TKG includes integrated lifecycle management capabilities that automate the provisioning, upgrading, scaling, and patching of Kubernetes clusters. This reduces the operational overhead and ensures that clusters are always running with the latest security patches and updates.

3. Integrated Networking and Security: TKG integrates with VMware NSX for advanced networking and security features, including micro-segmentation, load balancing, and ingress control. This integration enhances the security posture of Kubernetes clusters and enables fine-grained network policies.

4. Conformance: TKG delivers fully conformant Kubernetes clusters that adhere to the standards defined by the Cloud Native Computing Foundation (CNCF). This ensures compatibility with the broader Kubernetes ecosystem and allows organizations to leverage a wide range of Kubernetes tools and extensions such as cluster API, Cluster class, and more.

To support the growing demand for containerized workloads, VCD has evolved to include native support for Tanzu Kubernetes Grid using Container Service Extension (CSE). This integration allows CSP-Cloud Builders to offer Kubernetes-as-a-Service (KaaS) to their tenants, enabling them to deploy and manage Kubernetes clusters alongside traditional virtual machines.

This document proposes a generic reference architecture to CSP-Cloud Builders to get started with KaaS with Tanzu Kubernetes Grid on VCD. The target audience for the document is Infrastructure architects, Cloud provider administrator, Cloud Security Specialists, Network, Storage administrators, and personas such as DevOps Engineers.

Scope and Bill of Material

The scope of this proposed architecture is for CSP-Cloud Builders to introduce KaaS on existing IaaS which as following components already enabled:

VMware SDDC with VMware recommended best practices.

NSX Data Center integrated with VMware SDDC Infrastructure

NSX Advanced Load balancer configured and integrated with VCD.

With above assumptions in consideration, CSP-Cloud Builders can begin with the following Stack to offer KaaS with CSE 4.0.

BILLS OF MATERIAL WITH CONTAINER SERVICE EXTENSION 4		
IAAS COMPONENT	VERSION	LICENSE (INLCUDED AS PART OF FLEX-CORE BUNDLE
VMware Cloud Director Platform	VCD 10.3.1 and above	Included
VMware NSX®	3.1.3	VMware NSX SP Base
NSX Advanced Load Balancer	21.1.1	Basic
Container Service Extension	4.0.3	Value added service with Tanzu Kubernetes Grid
VMware App Launchpad	2.1.2	Value added service
VMware Chargeback	8.10	
VMware Aria Operations for Logs	8.10	

Table 1 Bill of Materials for CSE 4.0 based KaaS solution on VCD

Container Service Extension

Container Service Extension 4.0 is an add-on to VCD that simplifies the deployment and management of Tanzu Kubernetes Clusters. CSE 4.0 provides a seamless experience for tenants to create, scale, and operate TKG clusters directly from the VCD Tenant portal. It also offers enhanced security, monitoring, and networking features as part of TKG runtime, ensuring that the TKG clusters are secure, performant, and fully integrated with the underlying cloud infrastructure.

Starting CSE 4.0 release, CSE is responsible for creating and deleting the Tanzu Kubernetes Clusters, making all other Cluster operations to be performed on Cluster itself. Update operations such as Scale control plane, worker plane, node pool, cluster Upgrade and changing settings of the cluster operations are performed directly on the cluster. Figure 1 showcases Create/Delete and Update operations flow on the tenant portal and CSE 4.0.

Another enhancement decouples CSE Interaction with tenant's TKG clusters. CSE 4.0 is no longer client to server request flow. Tenant's users use VCD API to initiate cluster creation and deletion operations, CSE listens to the operations responds to the client over VCD APIs. Thus, if VCD API is available, cluster operations requests can be made by the user. Figure 1 also showcases this feature enhancement, where CSE and TKG cluster author interacts with VCD APIs, without interacting directly with each other.

Another enhancement also introduces out of the box support for Cluster API for VMware Cloud Director (CAPVCD), which enables lifecycle operations such as upgrade, update, scale, etc. on Tanzu Kubernetes Clusters making them 'Self-managing' clusters.

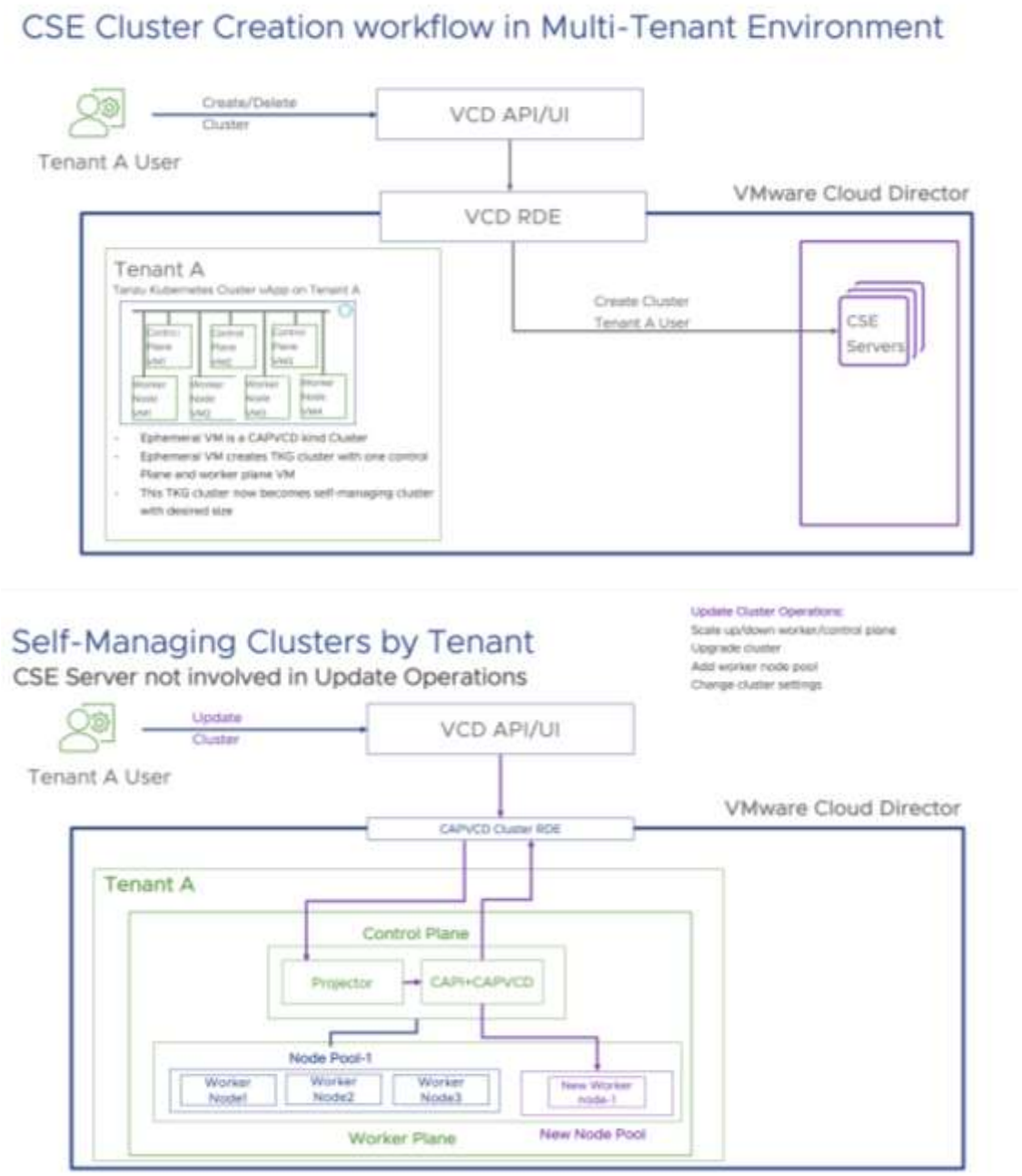


Figure 1 Container Service Extension Architecture and workflows with VMware Cloud Director

Starting CSE 4.0 release customers can create TKG clusters with multiple control plane VMs. The following sub-section describes CSE components.

Container Service Extension Components

CSE has simplified deliverables in two components. Also, CSE 4.0+ supports Tanzu Kubernetes Clusters to provide TKG runtime on VCD platform, dropping support for Opensource Kubernetes templates. The simplified architecture offers UI management for CSP admin, added management, security capabilities such as global proxy support, Log forwarding as described below.

CSE Server OVA:

CSE Server runs as system service where it listens for TKG cluster creation and deletion requests from VCD API. Cloud provider admin or KaaS admin user role can instantiate OVA with VCD infrastructure details and start CSE server. To achieve high availability, cloud provider admin/system admin can instantiate multiple CSE server vApps (recommended 3 vApps). By creating

multiple CSE server vApps, when there is an incident on one of the CSE servers, workloads and configurations shifts to remaining healthy CSE server vApps. CSE Server is a stateless application, meaning system admin won't have to perform backup/restore of the CSE server state. Upon node failure, system admin can re-create CSE server for their relevant VCD infrastructure.

CSE UI Plugin

The CSE UI plugin also called as "Kubernetes Container UI plugin" is offers services to two user personas: system admin and tenants. CSE Management tab guides system admins to onboard and manage CSE Service by importing TKG templates, CSE Server OVA, setup pre-requisites, setup Kubernetes cluster rights bundles, user-role for tenant organizations, onboard tenants and lastly start CSE service. The tenant users can user UI plugin to perform Lifecycle management operations and view TKG cluster events. Please refer to VMware CSE documentation for complete guide *here*.

Following are newly introduced features in CSE Management tab for providers in container service extension:

Global proxy settings

Container Service extension supports configuring global proxy settings for all tenant's TKG cluster's outbound traffic if they have specific security requirements.

CSE server configuration and upgrade

The cse admin or system admin can select supported and desired components versions such as CAPVCD, CPI, CSI. System admin can also perform CSE server upgrade from UI plugin.

Syslog forwarding

Cloud provider admins can leverage their existing syslog collector to collect CSE server logs. Please refer to VMware Cloud Director best practices to use recommended Syslog collector such as VMware Aria Operations for Logs.

The system admin can update component versions, proxy settings and syslog settings from UI. Note that the server configuration changes are not applicable to clusters that already exist. It is necessary to manually update the server configuration associated with existing clusters. It is also required to restart the existing CSE Server vApp to apply the updated configuration.

The following table describes user profiles for CSP-Cloud Builders involved in Container Service Extension.

USER PROFILE	ROLE DESCRIPTION
Cloud Administrator	• Cloud Admins handle setting up VCD, CSE server using CSE management tab in UI plugin, Tanzu Kubernetes templates. Cloud administrators are expected to possess this role and be experienced in VCD administration
CSE admin	• This role comes with all the VMware Cloud Director rights that CSE needs to function. The CSE Admin Role allows a user to perform administrative tasks in VMware Cloud Director Container Service Extension. You can use these user credentials as OVA deployment parameters when you start the VMware Cloud Director Container Service Extension server. Cloud administrators are expected to possess this role and be experienced in VCD administration

Table 2 CSE User Personas for cloud provider

The following table describes user personas for Customer organization when working with Tanzu Kubernetes Clusters through CSE

USER PROFILE	ROLE DESCRIPTION
Tenant Org admin	• The tenant administrator is responsible for creating user roles for the developers and can also create and manage the Kubernetes clusters. Tenant users who manage Kubernetes clusters are expected to understand VCD org administration principles. They should have accounts with privileges required to create vApps and manage them. Finally, such users should understand Kubernetes cluster management including setting up user access and defining persistent volumes. • To perform cluster management functions, you must hold the rights of the Kubernetes Cluster Author role. Additionally, you require Administrator View: VMWARE: CAPVCDCLUSTER to view all the clusters in your organization. If you cannot assign these rights to yourself, contact your service provider. • As an organization administrator, you must reassign the existing Kubernetes users in your organization who manage their own clusters to the new Kubernetes Cluster Author role.
Kubernetes Cluster Author	• Kubernetes Cluster Author role contains all rights required for the tenant users to manage the lifecycle of their Tanzu Kubernetes clusters. This global role will be automatically published to all tenants. Users without this role, or the equivalent rights, cannot manage Tanks Kubernetes Grid clusters. This role must be assigned by Organization Administrators to required users in the Organization.
DevOps, Developers and other Kubernetes Users	• Develop and deploy applications on the Tanzu Kubernetes cluster using kubectl. Tanzu Kubernetes clusters work like any other Kubernetes cluster implementation. No special knowledge of VMware Cloud Director or VMware Cloud Director Container Service Extension administration is required. You do not require a VMware Cloud Director account.

Table 3 User personas for Customer Organization for CSE

Integration with VMware Cloud Director:

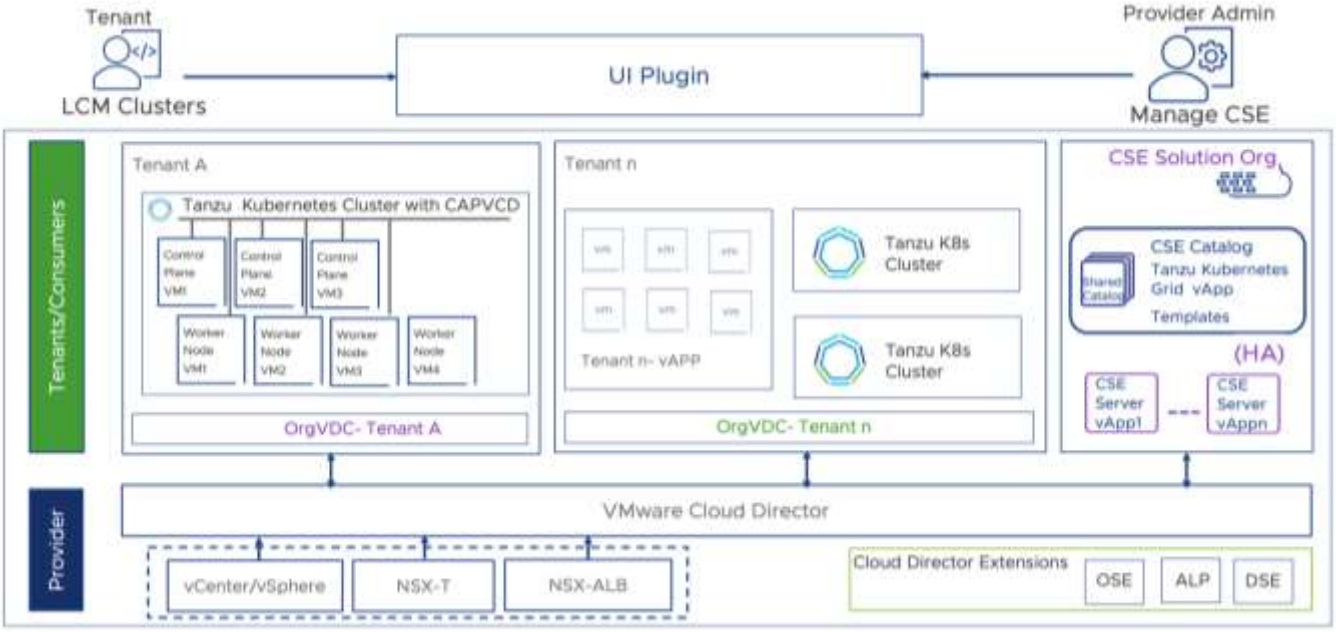


Figure 2 VMware Cloud Director and Container Service Extension Platform Architecture

CSE's integration with VCD provides an effective solution for managing containers and virtual machines in a single platform, by offering standardized TKG KaaS offerings. Figure 2 describes high level platform architecture and Container service extension's placement in VCD. Below integration points makes the solution complimenting VCD's offerings.

vApp and VM Management

When user sends a request to create a cluster in tenant portal, CSE server deploys the Tanzu Kubernetes Cluster as a vApp in the customer's relevant tenant portal. This allows the TKG cluster to utilize all VCD infrastructure features such as VM placement policies, GPU Policies, Sizing templates, the tenant users' roles, user groups, IDP, API token and more.

Network and NSX Adv Load Balancer Integration

CSE 4.0+ releases require that the system/network admin has created VMware NSX® (NSX) based routed networks. TKG clusters on VCD require use of NSX routed network topology. All VMs in the TKG clusters are connected through tenant's routed network. All TKG clusters' control plane VMs are serviced by VMware NSX® Advanced Load Balancer™ – Basic Edition (NSX Advanced Load Balancer) and consumes external network IP address. Figure 3 showcases high-level Load balancing for TKG clusters for tenants using NSX routed networks, NSX Advanced Load Balancer for control plane.

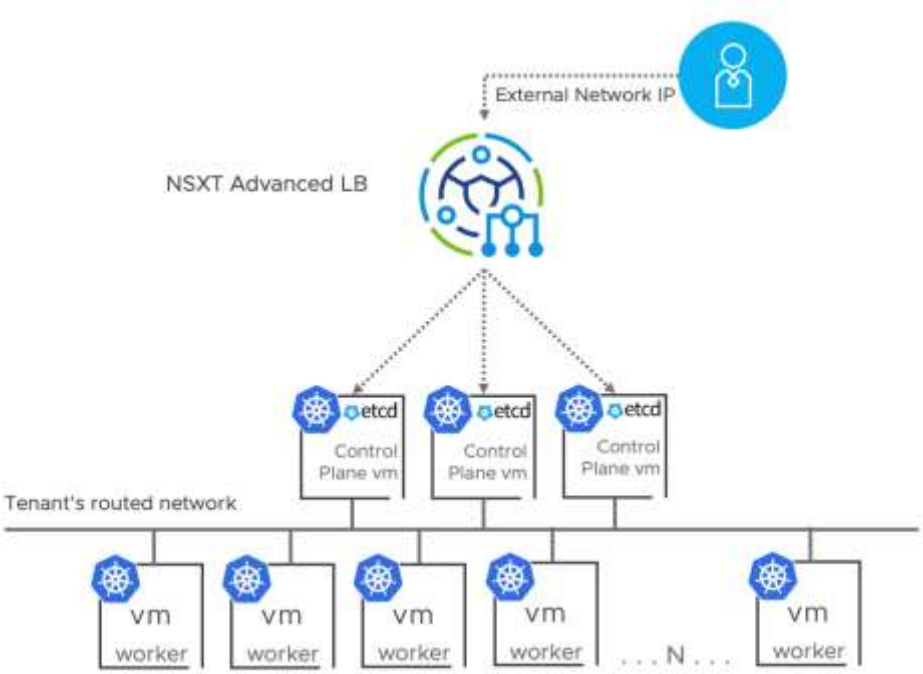


Figure 3 NSX Adv Load Balancer Service for Control Plane Access

Resource Allocation and Management with VCD

Upon starting CSE server, as a pre-requisite standardized TKG sizing templates gets created. The system admin then publishes desired sizing templates to the tenant organizations. These sizing templates defines CPU, Memory, and Storage values per VM in the cluster. Similarly, providers can create desired placement policies to place TKG clusters control plane and worker plane on various infrastructure clusters based on the desired outcome. Not to mention, each user creating cluster follows their allocated user quota to use the resources; same as VMs. Please refer to VMware Cloud Director best practices for allocation model, resource pool, Storage profiles, etc. for efficient resource management.

Please refer to Service Organization Recommendations for CSE in Provider Operations section in detail in following Reference Architecture section.

Reference Architecture and Provider Operations

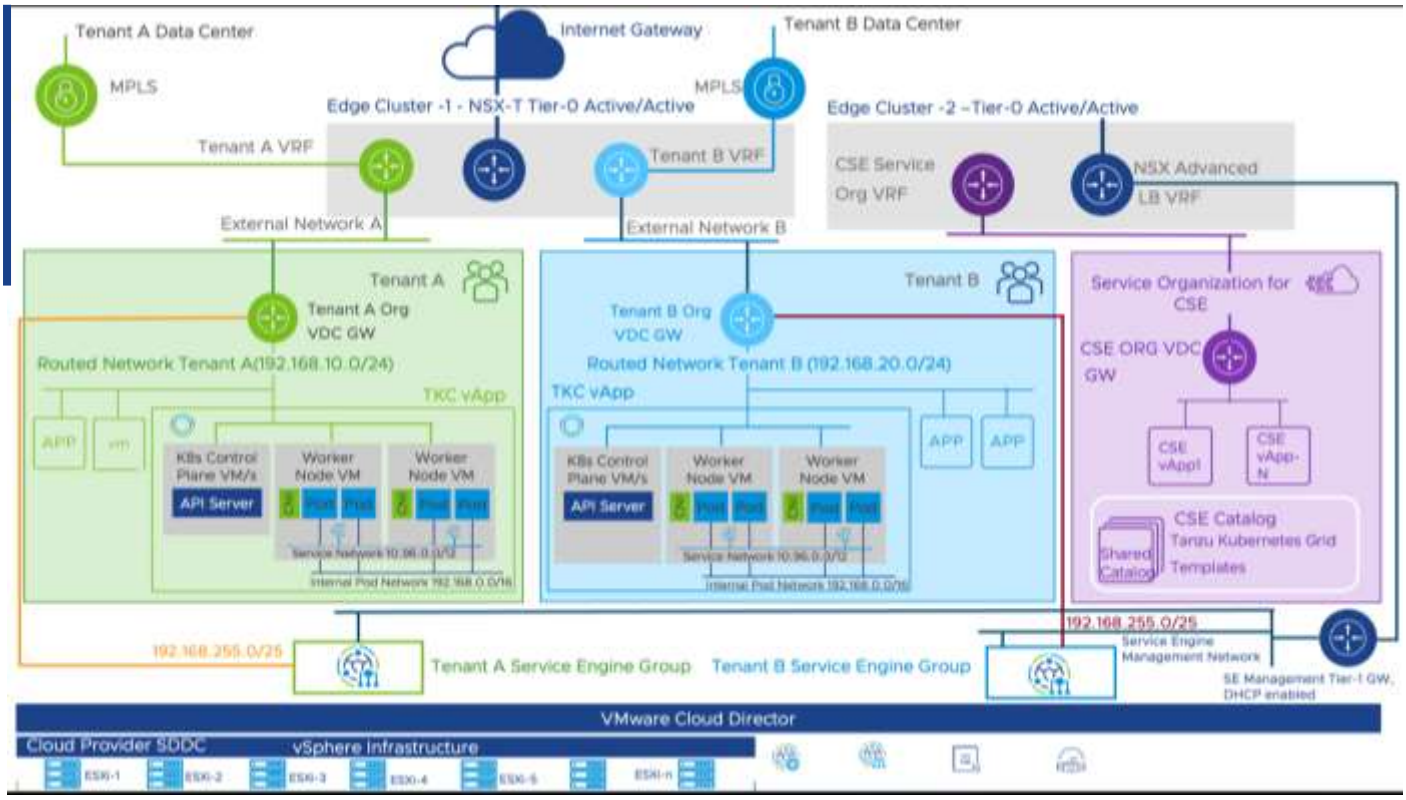


Figure 4 Detailed Recommended Architecture for KaaS for CSE with VCD

Figure 4 describes detailed reference architecture for CSP-Cloud Builders. You can see that proposed architecture has two edge clusters: Edge Cluster 1 - for workload domain where customer organizations relate to their respective VRFs. Providers allocates External network for each customer organization and, each customer organization has a Tier-1 Edge Gateway. The Tier-1 Edge gateway offers Gateway firewall, NAT Rules configuration, routed networks with static IP pools, and Load Balancer settings. For the scope of this document, we will focus on Routed network with static IP Pools, Gateway Firewall rules, required SNAT rules for outbound traffic and NSX Advanced Load balancer configuration for each customer organization. Second is Edge Cluster 2 – which connects CSE service Org and NSX Advanced LB Management domain with VCD endpoints.

In coming sections, we will cover recommendations for the service organization, NSX Adv Load balancer config, customer org network, NAT, Firewall rules and LB configuration settings.

Following Table 3 describes all required configuration on NSX Tier-1 Edge gateway for each customer organization:

PROVIDER CREATED TIER-1 EDGE GATEWAY CONFIGURATION	
COMPONENT	CONFIGURATION
External IP Pool (configured by provider)	At least 3 available static IPs for Kubernetes Cluster configuration
Routed network (configured by provider)	with Static IP pool and valid DNS configuration

PROVIDER CREATED TIER-1 EDGE GATEWAY CONFIGURATION					
COMPONENT	CONFIGURATION				
NAT Rules (configured and managed by customers network admin)	**Name**	**Type**	**External IP**	**Internal IP**	**External Port**
	Generic SNAT Rule for K8s outbound traffic	SNAT	<Available IP from External IP Pool>	<Org Routed Network CIDR>	Any
Gateway Firewall Rules (configured and managed by customers network admin)	**FW Rule name**	**Source**	**Destination**	**Service**	
	K8s Service (access Tanzu Binaries and VCD endpoint)	<Org Routed Network CIDR>	<Any>	TCP:443 TCP:6443	
	NTP Service	<Org Routed Network CIDR>	NTP Servers IPs	UDP:123	
	DNS Service	<Org Routed Network CIDR>	DNS Servers IPs	UDP:53	

Table 4 Networking configuration per customer organization

Securing VCD and CSE with Gateway firewall rules

Following table describes all required gateway firewall rules managed by Cloud provider related to successful Kubernetes as a Service Solution using CSE. Please refer to complete VMware Cloud Director Security Guide for security recommendations *here*.

If would like to know more about the various ports across VMware products, then *click here*

FIREWALL REQUIREMENTS (GATEWAY FIREWALL RULES)				
COMPONENT	SOURCE	DESTINATION	PORT	DESCRIPTION
CSE Server in CSE service organization	CSE server	VMware Cloud Director Endpoint	TCP: 443	CSE server to access VCD public endpoint
	CSE Server	DNS Servers	UDP: 53	Allow CSE server to lookup DNS
(?)	NSX ALB Controllers and Cluster VIP	vCenter Server	TCP: 443	Allow AVI to discover vCenter Objects and deploy Service Engines (SE) as requested/required
	NSX ALB Controllers and Cluster VIP	NSX Manager (workload domain)	TCP: 443	NSX Cloud integration and discover NSX-Objects
	NSX ALB Controllers	ESXi Hosts	TCP:443	Management access for Service Engine Creation
	NSX ALB Controllers	DNS Servers	UDP: 53	Allow DNS lookup
	NSX ALB Controllers	NTP	UDP: 123	Allow time sync
	NSX ALB Service Engine mgmt. network	NSX ALB Controllers	TCP: 8443 TCP: 22	Secure channel for key exchange Secure channel for communication between NSX ALB components for configuration sync, metrics and logs transfer, heartbeats, and other management processes
	NSX ALB Service Engine Management Network	NSX Adv LB Controller	UDP: 123	Allow time sync

Table 5 Firewall Requirements to setup for CSP Builder admins for CSE service org and NSX Adv LB

Service Organization for Container Service Extension Server

As shown in Figure 4, the detailed reference architecture, there are two take aways:

1. Starting CSE 4.0 we recommend dedicating a service organization to host CSE Server vApp/s and TKG templates in a shared catalog. CSE server interacts only with VCD API for create/delete cluster operations. When the TKG cluster is created, the cluster creation requires outbound internet access to pull images for CNI, CPI, CSI, CAPVCD and Tanzu core packages. The NAT and Firewall rules reflect this requirement in detail for each customer's Tier-1 Gateway configuration.
2. Connect this service organization to Edge Cluster-2 which hosts NSX Adv Load balancer and CSE Server vApps. The next Table 5 describes all network design considerations for the service organization.

This recommendation offers many advantages to cloud provider admins. First, it **Enhances Security posture** of CSE and TKG templates from other customer orgs by preventing unauthorized access to CSE server vApps. The system admins can **improve operational efficiency**, as they can simplify administration of the org, logging of events on CSE server, etc. from single org. The system admin can improve resource allocation for management workloads. Also, system admins can experience consistency in operation, customer onboarding and overall predictable and streamlined CSE service offerings. Which in turn can provide them competitive advantage over other providers. The following table 5 provides VCD Org sizing specification for the service organization. This recommendation also calls for a mention of Solution Landing Zone introduced in VCD 10.4.1 release. CSP admin can now leverage preconfigured environment to support deployment and management of add-ons for VCD. It serves as dedicated and isolated area where CSP admins can introduce and manage new add-ons in CSP admin-controlled environment. Solution

landing zone offers UI to manage resources and life cycle of solutions extending functionalities to VCD. Please refer to Solution landing zone concepts, and admin guide and overall documentation *here* for more information.

SERVICE ORGANIZATION VDC CONFIGURATION	
CONFIGURATION	**SPECIFICATION**
Allocation Model	Flex
CPU Allocation	10 GHz/5 cores when each core is 2GHz, Adjust GHz value based on cores. For example, 12 GHz/4 cores when each core is 3 GHz.
Memory Allocation	16 GB
Guaranteed CPU and Memory	100%
Storage Allocation	100GB
Maximum Provisioned Networks	10
Catalogs	Two: - Shared with desired customer orgs; To host TKG templates. - Not shared; To host CSE Server vApp

Table 6 CSE Service Org Configuration on VCD

To achieve the reference architecture depicted in Figure 4, following design decisions are expected from CSP-Cloud Builders, which also provides justification and implication of each decision. The recommendation is to deploy and power on 3 CSE server vApps to achieve High Availability and improved CSE performance.

Tier-0 and Network configuration for Service Organization

DESIGN DECISIONS/RECOMMENDATION FOR TIER-0 AND NETWORK CONFIGURATION		
DESIGN DECISION/RECOMMENDATION	JUSTIFICATION	IMPLICATION
Create a dedicated Tier-0 VRF gateway to handle the Container Service Extension Server traffic	Dedicated VRF allows the separation of segments and tier-1 gateways. With a dedicated Tier-0 VRF gateway, the Routed Organization network under the Service organization VDC can use overlapping IP addresses without any interference with other existing networks in the infrastructure	Tier-0 VRF Gateway for NSX ALB SE Management and Container Service Extension Server will share the same Parent Tier-0 Gateway
Create a tenant Edge Gateway (Tier-1) in the Service Organization VDC, and dedicate the CSE Tier-0 VRF Gateway to the tenant Edge Gateway, or in the case of VCD 10.4.1 mark it as private	Tier-0 VRF Gateway is dedicated to handling CSE Traffic and need not be shared with other tenants Edge Gateways	
Create a Routed Org Network with a static IP pool under the Service Organization	CSE Server must be placed on a routed org network as it should be able to communicate with the VMware Cloud Director API endpoint. VMware Cloud Director will assign an available IP to the CSE server from the Static IP pool	
Create a SNAT rule for the Routed Organization Network to enable communication with the external network (This is alternative to configuring route re-distribution)	By default, organization networks are not advertised by the Tier-1 created by VCD, SNAT enables CSE Server to successfully communicate with VMware Cloud Director Endpoint	Firewall requirements must be met for the external network to communicate with the VMware Cloud Director API endpoint. Refer Firewall section for more details

Table 7 Network config considerations for CSE Service Organization

NSX Advanced Load Balancer with VCD

NSX ALB is leveraged to provide L4 load-balancing functionalities for the following:

- Services deployed in the Tanzu Kubernetes clusters via CSE.
- Control Plane components of Tanzu Kubernetes Clusters in case of multi-Control plane configuration

As a system administrator:

1. You deploy and configure NSX Advanced Load Balancer to use with your VMware NSX Data Center deployment.
2. You register NSX ALB Controller with VMware Cloud Director: NSX ALB Controller serves as a central control plane for load balancing services. After you register your controllers, you can manage them directly from VMware Cloud Director.
3. You register your VMware NSX Cloud instances with Cloud Director.
4. You import service engine groups into Cloud Director. The load-balancing compute infrastructure provided by NSX ALB is organized into service engine groups. You can assign more than one service engine group to a VMware NSX edge gateway.

After a system administrator assigns a service engine group to an edge gateway, an organization administrator can create and configure virtual services that run in a specific service engine group.

In a non-consolidated SDDC Architecture, NSX ALB service engines are deployed in the compute vCenter, these Service Engines require access to NSX ALB Controllers (typically deployed in Management vCenter) over the port 22(TCP), 8443(TCP), and 123(UDP), refer Firewall Section for more details. To segregate the management traffic from tenant traffic, it's recommended to create a dedicated Edge Cluster under Compute vCenter that handles the Service Engine Management traffic.

Once NSX ALB is configured and Service engine groups are created, you can register NSX ALB Controller with VMware Cloud Director and import the Service Engine groups. You can either register the service engine groups as "Dedicated" or "Shared" based on the tenant requirements.

- **Dedicated Reservation Model**: In this mode for each tenant Organization VDC Edge gateway, NSX ALB will create two Service Engine nodes for each Load Balancing enabled Org VDC Edge GW.
- **Shared Reservation Model**: In this model, SEs part of a SE group is typically shared across all the tenants.

Service Engine Groups Design Options in VCD

Based on the requirement of the tenants, one of the below options can be chosen:

1. **One SE Group per Tier-1/Dedicated Reservation Model**: In this mode for each tenant Organization VDC Edge gateway, NSX ALB will create two Service Engine nodes for each Load Balancing enabled Org VDC Edge GW. This model can be used for tenants who require guaranteed SE resources and a higher level of isolation to host their virtual service. With this approach:
    - The provider maintains the same design regardless of the number of tenants or the number of tier-1 per tenant.
    - Consistent Scalability: For each new tier-1 add a new VIP Data segment and the SE group is created.
    - Guaranteed SE resources
    - Higher Level Isolation: Dedicated data plane, as the SEs are not shared across tenants.
    - Higher cost
    - This could result in less efficient use of SE groups.
2. **One SE Group shared by all tenants/Shared Reservation Model:** In this mode, NSX ALB will create a pool of service engines that are going to be shared across tenants. Capacity allocation is managed in VCD, and NSX ALB deploys and deletes service engines based on usage. With this approach:
    - Optimal usage of SE Group reducing cost
    - Scalability is not consistent. Each SE group can support up to 9 OrgVDC gateways, need to deploy a new SE Groups
    - SE resources are not guaranteed.
3. **One SE Group shared per tenant/Shared Reservation Model:** In this mode, a Service engine group is shared across all Tier-1 under a single tenant. Each tenant will have his own Service Engine group and shared with all the Tier-1 available within the Organization.
    - Better SE Group utilization compared to Option 1
    - Up to 9 VIP Segments per Tenant, basically up to 9 OrgVDC gateways per tenant
    - SE resources are guaranteed at the tenant level.
    - Scalability consistency is achieved at the tenant level.

Please refer to Appendix for detailed NSX ALB design decisions. Also, as cloud provider you're encouraged to refer to NSX ALB VMware Validated Solution (VVS) *here*.

GPU Policy for TKG clusters worker nodes

CSP-Cloud Builders can offer vGPU infrastructure to customers starting with release 10.3.2. With CSE 4.0, system admin can publish the vGPU policies to desired tenant organizations and offer GPU as a service along with Tanzu Kubernetes Clusters. Please refer to *Creating and Managing vGPU Policies* in the VMware Cloud Director Service Provider Admin Guide documentation for more information and best practices *here*.

Usage Meter Reporting

VMware Usage Meter for Cloud Service providers allows providers to track their customer's usage of various products such as ESXi, vCenter Server, VSAN, VMware Aria suite, VMware NSX for Service Providers, NSX Advanced Load balancer etc. CSP-Cloud Builders use flex-core bundle to offer Tanzu Kubernetes Grid to the customers on VMware Cloud Director using CSE. With CSE 4.0, Usage meter doesn't support automated usage report to commerce portal for Customers' TKG clusters on VCD tenant portal. There are two ways CSP-Cloud Builders can effectively and easily track TKG usage for each customer:

1. System admin can manually report TKG cluster usage per tenant to Commerce portal for CSP-Cloud Builders. To improve manual reporting experience, system admin can configure a custom placement policy for each customer. This allows to have a specific resource pool for CSE based TKG clusters. System admin can then label this resource pool for TKG usage and easily report usage specific to the created label.
2. Alternatively, CSP-Cloud Builders can use VMware Chargeback 8.10 create custom dashboard to track CSE provisioned TKG clusters and report the usage manually through commerce portal.

VMware Chargeback™ (VMware Chargeback) dashboard for providers

Cloud service providers use VMware chargeback to track and report cost to the customers based on the used by the customers. This tool tracks usage of VMs/vApps resources such as Network, compute, and storage and generates comprehensive report for billing services. VMware Chargeback 8.10 supports customization of providers dashboard to track CSE 4.0 based TKG clusters for each VCD Instance. Customers can generate and view their TKC usage with VMware Chargeback. CSP-Cloud Builders can use the chargeback reported TKG usage to report to CSP-Cloud Builders commerce portal. Figure 6 explains how CSP admin can build a custom usage view for TKG cluster usage for tenant organization with VMware Chargeback 8.10 release. Figure 7 shows how tenant admin can view TKG cluster usage per VCD cell by selecting the VCD cell to view and generate TKG cluster monly usage report from VMware chargeback on VCD tenant portal.

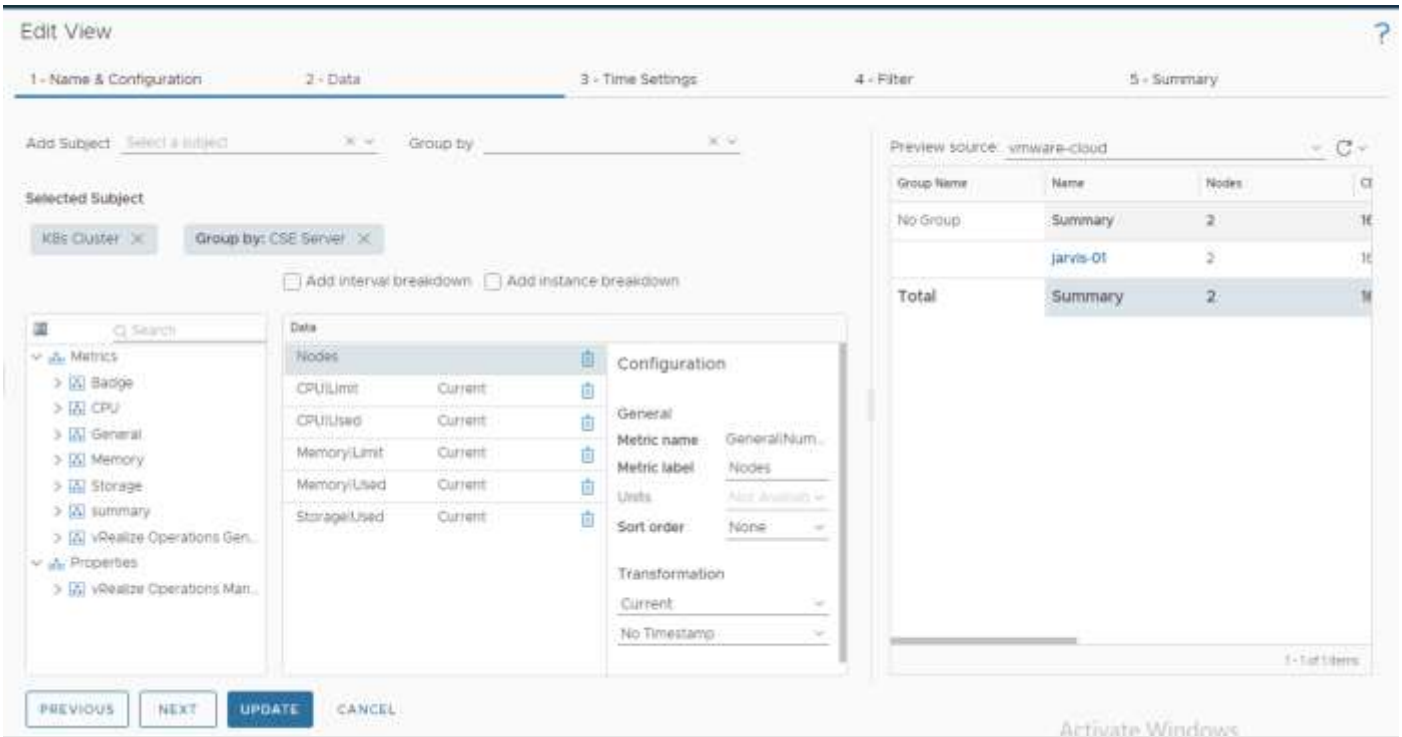


Figure 5 Provider dashboard to track CSE based TKG cluster on VMware Chargeback 8.10

Figure 6 Tenant selecting their VCD cell to generate Bill for TKG Clusters

This summarizes Container Service Extension design and various recommendations for CSP-Cloud Builders, system admins and cse admin user personas. As CSP-Cloud Builders progress their KaaS offering, from few Tanzu Kubernetes clusters to 100s of TKC, CSP should consider introducing 'cse admin' user who can manage significant responsibilities such as customer onboarding, managing CSE server High availability, components such as CAPVCD, CPI, CSI versions, GitHub tokens, syslog monitoring for CSE events, etc.

Tenant Operations

As a Tenant of VCD using CSE 4.0, users with right permissions can perform following operations related to Kubernetes Clusters, including creating TKG cluster with multiple control planes, one or more worker node pools, perform lifecycle operations from UI and VCD APIs, troubleshoot cluster state based on cluster events, and much more. Please refer to CSE Tenant admin guide for following supported operations *here*:

- *TKG Cluster Lifecycle Management using UI*
- *TKG clusters with worker node pools*
- *TKG Clusters with GPU*
- *Deploy Stateful applications on TKG clusters*
- *Ingress Service on TKG clusters*

Tanzu Core packages on VCD

TANZU CORE PACKAGES ON VCD			
PACKAGE	PACKAGE REPOSITORY	PACKAGE NAMESPACE	PACKAGE DESCRIPTION
antrea	tanzu-core	Kube-system	Enables pod networking and enforces network policies for Kubernetes clusters. If Antrea is selected as the CNI provider, this package is installed in every cluster
Core-dns	tanzu-core	Kube-system	Provides DNS Service, installed in every cluster
vcd-cpi	tanzu-core	Kube-system	Provides the VMware Cloud provider Cloud Provider Interface. This package is installed in every cluster
vcd-csi	tanzu-core	Kube-system	Provides the VMware Cloud Provider Cloud Storage Interface. This package is installed in every cluster
Kapp-controller	Kapp-controller	tkg-system	Manages packages. kapp-controller is installed in every cluster.
Metrics-server	tanzu-core	Kube-system	Provides Metrics Server. This package is installed in every cluster

Table 8 Tanzu Core Packages on VCD

Tanzu User Managed Packages

As a cluster author or user role, user can install and manage Tanzu User managed packages on TKG clusters provisioned in their tenant organization. Users must be careful to use the right Tanzu CLI version to install and manage the packages. The complete documentation for User managed packages is found *here*.

TANZU CLI MANAGED PACKAGES ON VCD			
PACKAGE	FUNCTION	DEPENDENCY	INSTALL LOCATION
Cert-management	Certificate management	Required by Contour, Harbor, Prometheus, Grafana	Workload Cluster
Contour	Ingress Control	Required by Harbor, Grafana	Workload Cluster
Harbor	Container Registry	-	Provider managed Harbor Registry
Prometheus	Monitoring	-	Workload Cluster
Grafana	Monitoring	-	Workload Cluster
Fluent-bit	Log Forwarding	-	Workload Cluster

Table 9 TKG CLI managed packages for VCD

Summary:

The VMware Container Service Extension 4.0 reference architecture offers a comprehensive guide for implementing and managing container services within a VMware Cloud Director (VCD) environment. It covers key aspects such as the Container Service Extension (CSE) and its components, integration with VCD for network integration, and with NSX Advanced Load Balancer, and resource allocation and management with VCD. The document also provides valuable information on service organization setup, design decisions, and their implications for CSP-Cloud Builders, including topics like network configuration, GPU policy for TKG clusters worker nodes, and usage meter reporting. This reference architecture is designed to help CSP-Cloud Builders deploy and manage container services in a secure, scalable, and efficient manner, offering their customers a robust platform for containerized applications.

Further Reading:

1. VCD Best practices for Cloud Service Providers
2. *VVS for Cloud Providers: Scale and Performance Guidelines*
3. *Container Service Extension 4.0 official documentation*
4. *Troubleshooting TKG clusters on VCD*
5. *Cluster API for VCD (CAPVCD)*

Acknowledgment

Author – Sachi Bhatt Staff Technical Product Manager

Do you want to give credit to the contributors?

Appendix

NSX Advanced Load Balancer Design Decisions

Below are the recommendations for VMware NSX configuration for NSX ALB

VMWARE NSX CONFIGURATION RECOMMENDATION FOR NSX ADVANCED LB		
DESIGN RECOMMENDATION/DECISION	JUSTIFICATION	IMPLICATION
Create a dedicated Edge cluster for NSX ALB Management in workload Domain	NSX ALB Service engines require access to NSX ALB Controllers, dedicated Edge Cluster segregates the management traffic from tenant traffic	Additional Resource requirement
Deploy NSX ALB controller cluster nodes on a dedicated logical segment or VLAN-backed network	Isolate NSX ALB traffic from other infrastructure management traffic	An additional Network (or logical segment) is required

Table 10 VMware NSX Data Center config recommendations for NSX Advanced Load Balancer

NSX ALB Controller

The NSX ALB Controller provides central control and management of the Service Engines. The AVI Controller runs on a VM and can be managed using its web interface, CLI, or REST API but in this case Cloud Director.
It orchestrates policy-driven application services, monitors real-time application performance (leveraging data provided by the Service Engines), and provides for predictive autoscaling of load balancing and other application services. Furthermore, it can deliver per-tenant, or per-application load balancing — increasingly in demand in multi-cloud contexts — and also facilitates troubleshooting with traffic analytics.

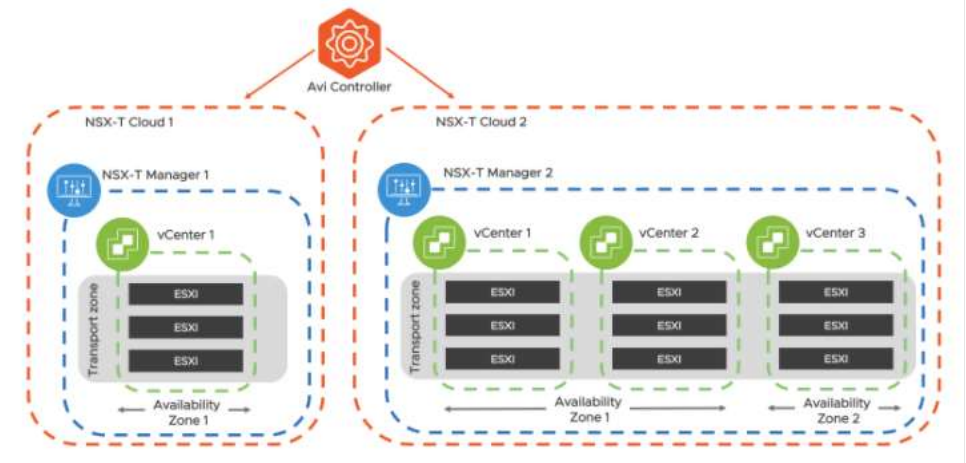
NSX ALB CONTROLLER DESIGN DECISIONS FOR CSE		
DESIGN RECOMMENDATION/DECISION	JUSTIFICATION	IMPLICATION
Deploy NSX ALB in Edge Cluster-2	NSX ALB is a platform solution responsible for providing load-balancing solution similar to other management components such as vSphere and VMware NSX	-
Deploy NSX ALB controller cluster nodes on a dedicated logical segment or VLAN-backed network	Isolate NSX ALB traffic from other infrastructure management traffic	An additional Network (or logical segment) is required
Deploy 3 NSX ALB controller nodes	• To achieve high availability for the NSX ALB platform. In clustered mode, NSX ALB availability is not impacted by an individual controller node failure. The failed node can be removed from the cluster and redeployed if recovery is not possible. • Provides the highest level of uptime for a site	Additional resource requirements
The size of each NSX ALB controller is set to Extra-Large	Provides support to scale up to 5000 Virtual Services and 400 Service Engines	Additional resource requirements
Use Static IPs for the NSX ALB controllers if DHCP cannot guarantee a permanent lease	NSX ALB Controller cluster uses management IPs to form and maintain quorum for the control plane cluster. Any changes would be disruptive	
Reserve an IP in the NSX ALB management subnet to be used as the Cluster IP for the Controller Cluster.	NSX ALB portal is always accessible over Cluster IP regardless of a specific individual controller node failure.	Additional IP is required
Apply vSphere DRS anti-affinity rules for the NSX Advanced Load Balancer Controller cluster nodes.	Ensure that NSX Advanced Load Balancer Controller VMs are distributed across ESXi hosts	Requires additional configuration to set up an anti-affinity rule.
Protect NSX Advanced Load Balancer Controller cluster nodes using vSphere High Availability	Supports the availability objectives for the NSX Advanced Load Balancer Controller cluster without requiring manual intervention during an ESXi host failure event.	
Create a dedicated resource pool with appropriate reservations for NSX ALB controllers	Guarantees the CPU and Memory allocation for NSX ALB Controllers and avoids performance degradation in case of resource contention	
Replace default NSX ALB certificates with Customer CA or Public CA-signed certificates with required SAN entries	To establish a trusted connection with other infra components, and the default certificate doesn't include SAN entries	
Use the internal NSX-ALB backup utility and schedule an automatic backup	Backups are the primary means of a DR, as such is it required that these are taken regularly. These timers can be adjusted based on specific RPO	None. The maximum RPO of the AVI controller configuration is 1 day, adjust the timer as per the RPO requirements.
Store 4 backups locally on the controller.	• Meets the need to be able to restore from a local failure.	Requires additional storage on the controllers

	• The number of backups can be adjusted based on specific RPO requirements.	
Configure NSX ALB backup with a remote server as a backup location	• Periodic backup of NSX ALB configuration database is recommended. The database defines all clouds, all virtual services, all users, and others. • As best practice, store backups in an external location to provide backup capabilities in case of entire cluster failure	The backup server must be provided by the Capstone. The storage on the backup server is not monitored, and backups are not automatically rotated.
Use the NSX ALB Controller to perform the life cycle management of the NSX ALB platform	NSX ALB Controller performs patching, and upgrade of components as a single process	The operations team must understand and be aware of the impact of a patch, and upgrade operations by using the NSX ALB Controller
Specify external system(s) for the NSX ALB Controller to send events on Syslog	For operations teams to be able to centrally monitor NSX ALB and escalate alerts events must be sent from the NSX ALB Controller	None

VMware NSX Cloud in NSX Advanced Load Balancer

The VMware NSX ALB may be deployed in multiple environments for the same system. Clouds are containers for the environment that NSX ALB is installed or operating within, each environment is called a cloud. Each cloud has its own environment, networking, and NSX ALB Service Engine (SE) settings are maintained separately within each cloud.
During the initial setup of NSX ALB, a default cloud, named "Default-Cloud", is created. Additional clouds must be added, containing SEs and virtual services.

An VMWARE NSX cloud is defined by an VMWARE NSX manager and a transport zone. If an VMWARE NSX manager has multiple transport zones, each will map to a new VMWARE NSX cloud. To manage load balancing for multiple VMWARE NSX environments each VMWARE NSX manager will map to a new VMWARE NSX cloud.



To integrate NSX ALB with VCD, an VMWARE NSX cloud must be configured in NSX ALB and needs to be imported to vCloud Director. In the case of VMWARE NSX cloud setup, the management network of Service engine and data/VIP networks will be VMWARE NSX logical segment.

Before configuring the VMWARE NSX cloud in NSX ALB, the below configuration must be in place on the VMWARE NSX managing workload vCenter

Note: Below steps outline the high-level configuration as per the reference architecture, for more configuration details refer *NSX-T Design Guide for NSX ALB(AVI) VMWARE NSX Design Guide for NSX ALB(AVI)*

1. Create a dedicated Tier-0 Logical Router for AVI Service Engines
2. Create a Tier-1 Logical Router and attach it to the Tier-0 created in the first step.

3. Create a logical Segment for AVI SE Management, and connect it to Tier-1
4. Configure and Enable DHCP Service (provided by VMware NSX) on the NSX ALB SE Management Segment - IPs defined in the DHCP scope will be assigned to Deployed Service Engines

Service Engine Groups and Service Engines

NSX ALB Service Engines (SEs) are created within a group, which contains the definition of how the SEs should be sized, placed, and made highly available. Each cloud will have at least one SE group. The options within an SE group may vary based on the type of cloud within which they exist and its settings, such as no access versus write access mode. SEs may only exist within one group. Each group acts as an isolation domain. SE resources within an SE group may be moved around to accommodate virtual services, but SE resources are never shared between SE groups.

NSX ALB Service Engines (SEs) take the form of distributed software that runs on bare metal servers, virtual machines, and containers. They implement application services across on-premises datacenters, colocation datacenters, and public clouds. They also collect data relating to application performance, security, and clients. As distributed software, Service Engines are capable of horizontal autoscaling within minutes while functioning as service proxies for microservices.

High Availability

NSX ALB Service Engine HA provides SE-level redundancy within an SE group. If an SE within the SE group fails, HA heals the failure and compensates for the reduced site capacity. Typically, this consists of spinning up a new SE to take the place of the one that failed. High availability for NSX ALB SEs is configured at each SE group level. NSX ALB SE groups support the following HA modes.

1. **Legacy HA**: Emulates the operation of 2-device hardware **active/standby** HA configuration. The active SE carries all the traffic for a virtual service (VS) placed on it. The other SE in the pair is the standby for the VS, carrying no traffic for it while the active SE is healthy. In this mode VS failover is quick, however, service engines cannot be scaled.
2. **Elastic HA**: Provides fast recovery for individual virtual services following the failure of an SE. Depending on the mode, the virtual service is already running on multiple SEs or is quickly placed on another SE. The following modes of cluster HA are supported:
    a. **Active/Active**: In this mode, NSX ALB places each virtual service on more than one SE, as specified by the "**Minimum Scale per Virtual Service**" parameter under the Service Engine group. If an SE in the group fails, then.
        - Virtual services that had been running are not interrupted. They continue to run on other SEs with degraded capacity until they can be placed once again.
        - If NSX ALB orchestrator mode is set to write access, a new SE is automatically deployed to bring the SE group back to its previous capacity. After waiting for the new SE to spin up, the NSX Adv Load Balancer Controller places on it the virtual services that had been running on the failed SE.

    b. **N + M**: In this mode, each virtual service is typically placed on just one SE, this default behavior can result in slow failover (due to the controller need to re-assign services), but efficient SE utilization.
        - The "N" in "N+M" is the minimum number of SEs required to place virtual services in the SE group — this calculation is done by the NSX ALB Controller based on **Virtual Services per Service Engine** parameter defined under SE group configuration. N will vary over time, as virtual services are placed on or removed from the group. The maximum number of Service Engines is determined based on the **Max Number of Service Engines** parameter defined under SE group configuration.
        - The "M" of "N+M" is the number of additional SEs the Avi Controller spins up to handle up to M SE failures without reducing the capacity of the SE group. M appears in **Buffer Service Engines** defined under SE group configuration
          Note: The buffer SE in N+M mode is the number of SE failures the system can tolerate for the VS to be operationally up (placed on at least 1 SE), but not to be at the same capacity. If there is a specific minimum scale per VS set in the SE Group and if additional SE is required above that, then you should increase the buffer SE according to the calculations.

Please refer NSX ALB official documentation for performance and Sizing recommendations *here*.

SERVICE ENGINE GROUP AND SERVICE ENGINE DESIGN DECISION		
DESIGN DECISION	JUSTIFICATION	IMPLICATION
NSX ALB Service Engine High Availability set to Active/Active	Provides higher resiliency compared to N+M and Active/Standby	Requires enterprise licensing
Enable ALB Service Engine Self Elections	Enable SEs to elect a primary amongst themselves in the absence of connectivity to the NSX ALB controller	An additional Network (or logical segment) is required
Set the SE size to 2vCPU and 4GB of Memory	This configuration should meet the most generic use case	For services that require higher throughput, these configuration needs to be looked into and modified accordingly
Reserve Memory for Service Engines	Guarantees the memory allocation for SE VM and avoids performance degradation in case of memory contention	If the reservation cannot be met, the SE VM will not power on
DESIGN DECISION	JUSTIFICATION	IMPLICATION
The minimum number of active Service Engines for the Virtual Service is set to 2	Ensures that any Virtual Service is active on at least 2 Service Engines, which improves SLA in case of individual Service Engine failure	
NSX ALB Service engines placed in Workload domain/clusters	NSX ALB Service engines provide Load Balancing services for tenant workloads and applications	NSX ALB SE components must be considered while sizing the workload clusters
Only include datastore shared across all hosts in the clusters for NSX ALB Service Engines	Ensures that the NSX ALB SEs are deployed only on the datastore specified which is shared across all SEs	

Glossary

VM	Virtual Machine
vApp	Virtual Application – comprises of multiple Virtual Machines which provides a solution
CSE	Container Service Extension
VCD	VMware Cloud Director
TKG	Tanzu Kubernetes Grid
TKC	Tanzu Kubernetes Cluster
API	Application programming interface
GW	Gateway
FW	Firewall
CAPVCD	Cluster API for VMware Cloud Director
CSI	Container Storage Interface
CPI	Cloud Provider Interface
CNI	Container Networking Interface
PV	Persistent Volume

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 vmware.com Item No: vmw-wp-tech-temp-a4-word-101-proof 6/20