Designlet: VMware Cloud on AWS Management Cluster Planning

VMware Architecture

vmware by Broadcom © VMware LLC.

Table of contents

© VMware LLC.

Designlet: VMware Cloud on AWS Management Cluster Planning

Introduction

This document provides you with recommendations and guidelines to plan the layout of your Software Defined Datacenter (SDDC)'s management cluster. The management cluster in an SDDC is always the first cluster, Cluster-1. It is the cluster where management VMs such as vCenter, NSX, and most importantly, the NSX Edge gateways run in a dedicated Resource Pool called Mgmt-ResourcePool. It is possible to share the management cluster with regular workloads, which run in the Compute-ResourcePool.

This article discusses the reasons and limitations around these different SDDC design considerations to help you select the model most appropriate to your workloads and requirements.

Summary and Considerations

Use Case	Depending on the type and performance characteristics of your workload VMs, there are recommended design guidelines that should be followed for the VMs that run in the management cluster, as well as the SDDC size and scale.
Pre-requisites	Understand network traffic volume (in packets-per-second) and patterns (north-south vs. east-west) for your VMs. Understand storage characteristics of your VMs, such as IOPS, I/O type (read/write), I/O sizes, and total storage requirements. VMware offers vRealize Network Insight and vRealize Operations solutions in both on-prem software and SaaS versions to assist in collecting and analyzing this data.
General Considerations/Recommendations	
Performance Considerations	With i3.metal management cluster hosts, plan for a maximum north-south packet rate of 800k PPS (based on 0.05% packet loss), 2.4M PPS for I3en.metal, and 3.0M PPS for I4i.metal. Packet rates above those values may lead to increased packet loss that can impact application performance. Each Traffic Group in a Multi-Edge deployment can support this rate for north-south traffic. Note that with I4i instances, throughput will be limited to 75Gbps.
Network Considerations/Recommendations	Determine the packet rates to and from VMs that will be migrated when on a stretched network. Although bandwidth (Bytes per second) is the typical measurement used, in cloud environments like VMware Cloud on AWS, limitations are more commonly based on packets-per-second (PPS), and so this is the preferred metric. The relationship between PPS and bandwidth is directly proportional to the average packet size of the traffic. Larger packet sizes typically yield better bandwidth.
Cost implications	Dedicating hosts to a management cluster may require additional hosts in the SDDC. A correctly sized management cluster, however, may allow collapsing multiple SDDCs resulting in fewer total hosts.
Documentation reference	Predefined Clusters and Resource Pools Upsize SDDC Management Appliances Add a Cluster Configuration Maximums for VMware Cloud on AWS Understanding VMware Cloud on AWS Network Performance
Last Updated	March 2023

© VMware LLC.

Note: All rate and capacity numbers provided in this document are based on our testing and experience and are subject to change. They should not be considered as hard limits, but as best-practice guidelines. Exceeding these recommendations may lead to performance issues. The guidance provided is current as of the date published and is subject to change as new features and capabilities are introduced.

© VMware LLC.

Background

In VMware Cloud on AWS, network connectivity to ESXi hosts in the management cluster (Cluster-1) is shared between multiple services. It is important to understand how network traffic generated by or for one service can impact others. In particular, the ESXi host where an active Edge is running determines the capacity available for north-south traffic (traffic that is going between the SDDC and an external location, such as on-prem or to the connected VPC or over Transit Connect). Other services that consume network resources on that same host can reduce the amount of capacity available for the Edge, and therefore limit north-south traffic throughput.

In this document, we also refer to network capacity in packets-per-second (PPS), as opposed to measuring throughput such as Gigabits-per-second (Gbps). This is because the interfaces process individual packets at a certain rate which is not significantly affected by the size of the packet. At a given PPS rate, larger packets will transfer more data, resulting in higher bandwidth.

© VMware LLC.

Planning and Implementation

Refer to the Summary and Considerations table for general information and recommendations. The sections below provide detailed information for various SDDCs such as Large SDDC, Multi-AZ (Availability Zone) SDDCs, and so on.

Workload Impacts

There are two primary considerations for how workloads can consume network capacity from an ESXi host. The first being network traffic sent by or to the VM which, regardless of the destination or source (unless it happens to be on the same ESXi host), will have a direct impact on the host's network capacity. The second is storage I/O, as VMC uses vSAN storage. All hosts in the cluster participate in providing storage capacity over the network to VMs running in that cluster, regardless of which host they run on. In terms of network impact, disk writes are more impactful than disk reads, especially when using Erasure Coding storage policies (e.g. RAID-5/6).

HCX

You can choose to use VMware HCX to migrate workloads into an SDDC, and to extend (L2E) networks between on-prem and your VMC SDDC during the migration period. L2E traffic can have a high PPS value, as there is limited control within the network segment, and there are often many small broadcast packets on a network that must be sent across the L2E. Migration traffic is generally more efficient from a packet-per-second perspective, since it sends large amounts of data using full-sized packets. Still, network traffic between the Edge to the Interconnect appliance (IX), as well as between the IX and the WANOPT appliances need to be taken into consideration. Expect at least 50k PPS north-south from the IX, and 2-4 times that between the IX and WANOPT. Use of the WANOPT is recommended to minimize north-south traffic through its optimizations.

Perform a careful analysis that includes determining the packet rates to and from VMs that will be migrated when on a stretched network. L2E and migration traffic is a significant contributor to north-south traffic and must be factored into the overall capacity of the Edge and the host the Edge is on, since virtually all the traffic going over the L2E, as well as migration traffic, will be north-south, there is no secondary impact caused by this traffic. Since HCX appliances must always run in the management cluster, the main consideration required for HCX traffic is its overall contribution to north-south traffic.

Large SDDC

When an SDDC is deployed, there is the advanced option available to deploy the SDDC appliances as medium (default) or large. It is also possible to scale-up an SDDC from medium to large after deployment, however it is not possible to scale down from large to medium. A minimum of 3 hosts are required in the management cluster to scale-up an SDDC. Large appliances should be selected whenever an SDDC is expected to require any of the following:

- Have over 30 hosts
- Have over 3000 VMs
- Deploy multiple Traffic Groups (Multi-Edge)
- Use i3en.metal instances for the management cluster (for the purposes of network throughput)
- Run customer workloads in the management cluster
- Run network latency sensitive applications
- VPN with multiple ECMP tunnels and/or high bandwidth requirements (>2Gbps)
- Contain workloads with bursty-type network traffic
- Have high north-south packet rates (> 600k PPS)
- Have HCX L2 extended networks with high packet rates (>300k PPS)
- Running VMware Horizon / VDI desktops in the SDDC.
- Will deploy DFW at large scale (10k+ rules)

Multi-AZ (Stretched Clusters)

Multi-AZ, or stretched cluster SDDCs, follow all the same guidelines for Single-AZ SDDCs outlined in this document. There are some additional considerations that should be kept in mind, such as VM placement: In most cases, when an SDDC is deployed, the management VMs will all be running in the same AZ, sometimes referred to as the “preferred” fault domain. The SDDC will avoid moving them out of their current AZ, however, in case of failure, there is no affinity to a particular AZ, and especially with multi-Edge deployments, it’s possible that Edge VM HA pairs are either split across AZs or are restarted in a different AZ from where they were previously. While Edge VMs will always be restarted within the SDDC in case of a failure, it may not be in the same AZ

© VMware LLC.

they were originally in. It should not be assumed that the Edges will be in any specific AZ, nor spread over both AZs. This can cause some slight variation in network latency for north-south network traffic, depending on whether the VM communicating is in the same AZ or a different one from the active Edge VM, requiring cross-AZ communication.

Instance Type

The standard i3.metal instance type offers a base level of performance that is sufficient for many workload profiles. However, as it approaches end-of-sale, the newer instances available provide significantly improved network capacity: I3en.metal instances provide approximately 3 times the network performance of i3.metal instances, and I4i.metal improves on I3en.metal by another 25%. However, the I3en.metal is primarily considered a storage-dense instance, and as such may not be an ideal candidate to use in the management cluster where storage requirements are minimal. Using this additional storage for customer VMs will cause a corresponding increase in network utilization, which will offset some of the gains provided by the I3en.metal network capacity. I4i instances, while still providing a significant amount of storage, also provide improved packet-per-second rates. Although throughput on I4i is capped at 75Gbps, only cases with large flows using full-sized jumbo frames are likely to approach that limit. I4i also has a faster processor which provides improved encryption performance resulting in better network throughput when using IPSec VPN or other TLS-encrypted traffic. I3en and I4i instances also provide network level encryption when traffic is destined for another I3en, I4i, or other supported instance type - either within the SDDC, or over Connected VPC. See Encryption between instances for the supported instance type list and other details.

Multi-Edge

Multi-Edge is a solution that deploys additional pairs of Edge appliances and provides a mechanism to direct traffic from specified source IP ranges to those additional Edges. Due to the network capacity demands of an Edge on its host, it is not possible to run multiple Edges on the same host, so the management cluster must have sufficient hosts to accommodate every Edge appliance (2 default + 2 per traffic group). Since the primary purpose of multiple Edges is to provide increased north-south network capacity and running customer workloads in the management cluster consumes network capacity from the hosts, the two strategies should generally be viewed as mutually exclusive.

Multi-Edge with Multi-AZ

Multi-Edge with 1 traffic group can pair well with a Multi-AZ SDDC. Since the stretched cluster operation requires a minimum of 6 hosts (3 in each AZ or Fault Domain (FD)) per cluster, those hosts provide sufficient capacity to support an additional traffic group in addition to the default Edge pair without adding additional hosts, providing a way to leverage the hosts in the second AZ beyond just providing fail-over capacity. Do keep in mind that in case of an AZ failure, there is a small possibility where multiple Edges will have to temporarily co-exist on the same host, given there will be 4 Edge VMs (2 for the traffic group and 2 for the default), and only 3 hosts in one AZ. This can limit north-south network performance since the host’s bandwidth will need to be shared between both Edges until the situation can be remediated.

© VMware LLC.

© VMware LLC.

Copyright © 2005-2024 Broadcom. All Rights Reserved. The term “Broadcom” refers to Broadcom Inc. and/or its subsidiaries.