Entry-level Clusters on VMware Cloud on AWS

VMware Architecture

Table of contents

Entry-level Clusters on VMware Cloud on AWS

Introduction

VMware Cloud on AWS delivers consistent vSphere-based infrastructure that runs on Amazon EC2 elastic, bare-metal instances that are dedicated to each customer. VMware Cloud on AWS supports several types of clusters that accommodate use cases ranging from evaluation and experimentation to complex applications support. To accommodate the needs for flexible and cost-effective vSphere based cloud infrastructure, we expanded the entry-level cluster offerings in standard and stretched SDDC configuration. Recently added entry-level stretched cluster configurations (two (1-1-1) and four (2-2-1) hosts) enable customers to leverage intelligent AZ placement even with just a handful amount of hosts.

This Feature Brief explains the capabilities and design recommendations for entry-level clusters. We are also covering architecture and resource trade-offs associated with some types of our entry-level cluster offerings.

Entry-level Clusters Definition

For this article and in the context of the technical capabilities of VMware Cloud on AWS, we would count all clusters with two and fewer hosts in a single AZ as **entry-level**. The list includes:

- Two hosts standard clusters;
- Four hosts stretched SDDC (2-2-1);
- Two hosts stretched SDDC (1-1-1).

2-2-1 meaning two hosts in each AB plus a witness node in a different AZ whereas 1-1-1 means one host and a witness node in each AZ. A single host SDDC does not qualify for the production SLA and is not discussed here.

Functional Capabilities

- Qualify for production SLA.
- Support both i3.metal, i3en.metal, and i4i.metal host types.
- Support scale up from a single host SDDC.
- Support scale down from a three host cluster (limitations applies)
- Support as a secondary cluster with custom core count.
- Support stretched cluster SDDC (2-2-1, 1-1-1).

Technical Constraints

IT architects are facing the challenge of combining business requirements, technical capabilities, and reduce risks. It's vital to understand possible trade-offs while planning to deploy and operate a small cluster.

CPU Reservation constraint

Entry-level clusters have different available CPU Reservation capacities based on the host type, cluster configuration, and deployment type (primary or secondary). The most limitations apply to primary clusters with the i3.metal host type. The actual amount of available CPU Reservation capacity per cluster is determined by:

- vSphere HA Admission Control configuration;
- Management VMs CPU Reservation (vCenter, NSX Manager, NSX Edge);
- Host CPU Frequency

vCenter Server or ESXi allows you to power on a virtual machine (VM) only if there are enough unreserved resources to satisfy the VM's reservation. VMs without any configured reservation require at least 32MHz and 100MB Ram to power on successfully.

Note: It's essential to differentiate between how vSphere HA and ESXi CPU scheduler treat reservations. vSphere HA is strictly limited by the available CPU Reservation in a cluster. ESXi (vmkernel) CPU scheduler treats CPU reservation as flexible – if a VM does not need the configured MHz, the CPU scheduler can use them for other VMs. However, in the case of a high CPU load of management VM, they will consume most CPU cycles.

We breakdown the constraint per cluster type:

- **Primary two hosts i3.metal standard** cluster supports up to 35 powered on workload VMs or up to 1151 MHz CPU reservation on a single powered-on VM. You decrease the number of simultaneously powered on workload VMs on a scale of 32 MHz if you are adding CPU reservation. For example, if the VM1 has a CPU Reservation of 640 MHz, you can deploy no more than 15 additional powered on VMs without CPU reservations.
- **Primary two hosts (1-1-1) i3.metal stretched** cluster is identical with primary two hosts i3.metal described above.
- **Primary four hosts (2-2-1) i3.metal stretched** cluster features asymmetric CPU reservation availability. vSphere HA Admission control is configured to reserve two hosts (one host in each AZ). A primary AZ (where managements VMs are active) has the same CPU reservation as the primary two hosts' standard cluster. However, the secondary AZ has a capacity of a complete i3.metal host to accommodate VMs with CPU reservations. In case of an AZ failure, vSphere HA can initially power on no more than 35 VMs or a single VM with 1151 CPU Reservation. All other VMs stays powered off. After eDRS completes temporary cluster extension, VMs must be manually powered on.
- **Secondary two hosts i3.metal / primary and secondary i3en.metal/i4i.metal standard and stretched** clusters support the number of powered on VMs up to configuration maximum.
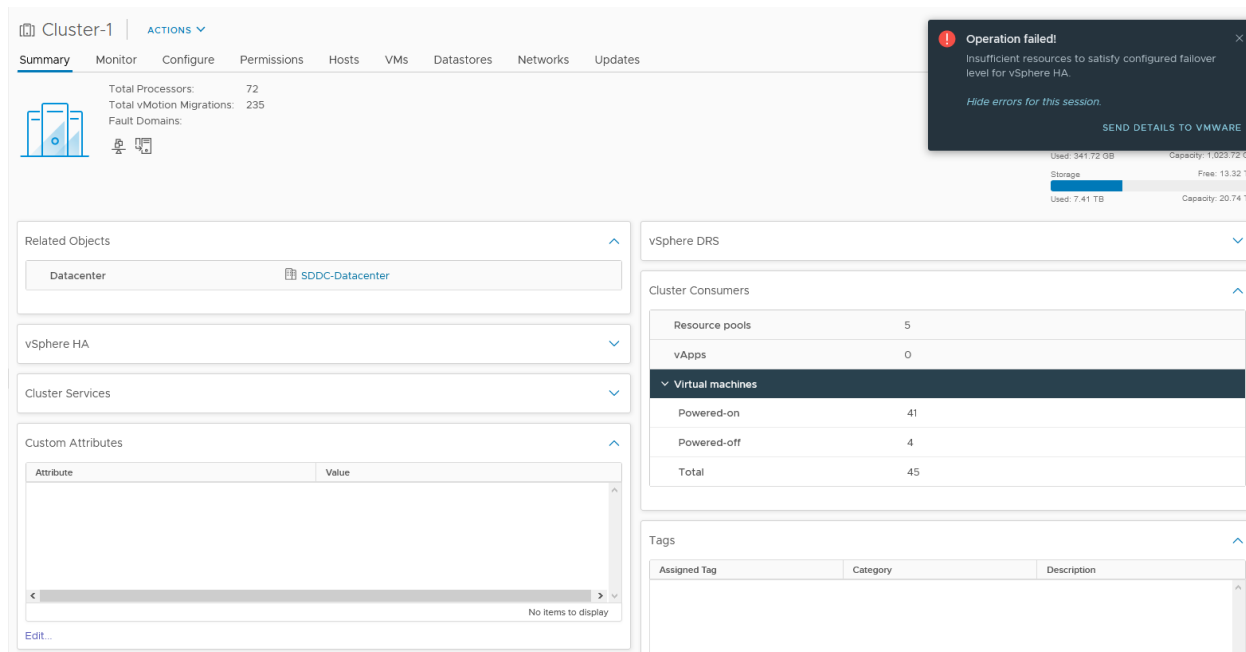
CPU Reservations

Let us take a closer look and understand the reasons for CPU Reservation constraints in the primary two host i3 metal cluster. This type of cluster has a limited amount of available CPU Reservation capacity.

When you create a new or migrate an existing VM and surpass the value of 35 running workloads VMs; or when you set a CPU reservation on a VM above 1151 MHz and power on the VM, your VM experiences a failure and cannot be powered on. You will see a message "*Insufficient resources to satisfy configured failover level for vSphere HA.*"
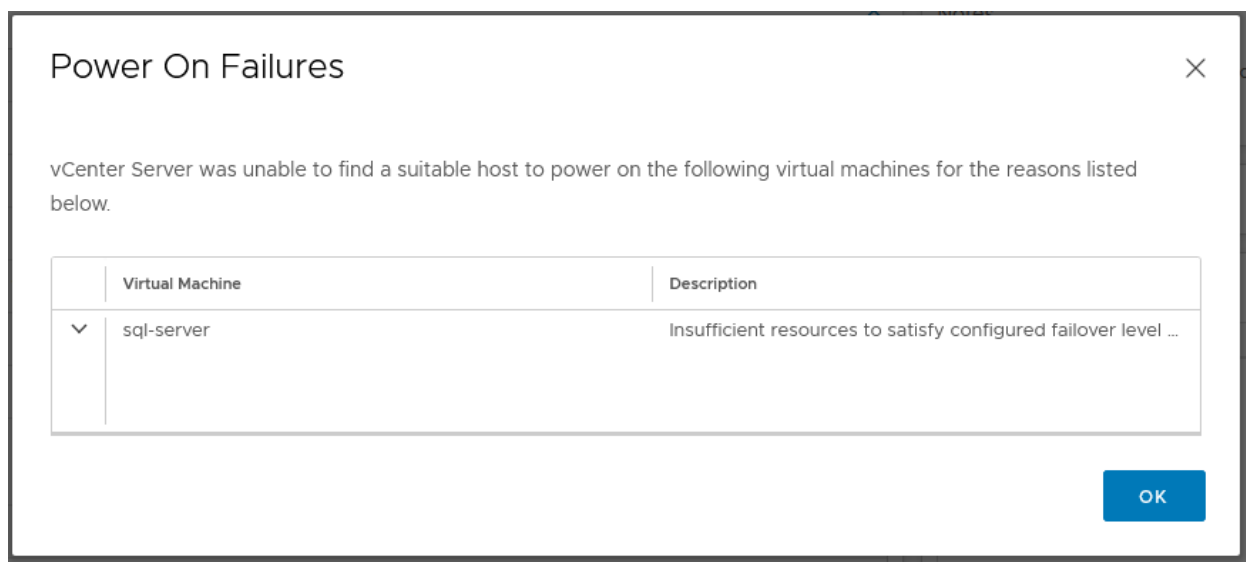
Figure 1: vSphere HA "*Insufficient resources to satisfy configured failover level for vSphere HA*" error message

And a pop-up in the upper right corner with the entire message:

Figure 2: VMware Cloud on AWS specific pop-up indicating vSphere HA error messager

The Cluster-1 *Monitor -- Resource Allocation – CPU* section in the vSphere Web Client (Figure 3) provides helpful information about CPU Reservations details to understand this constraint.
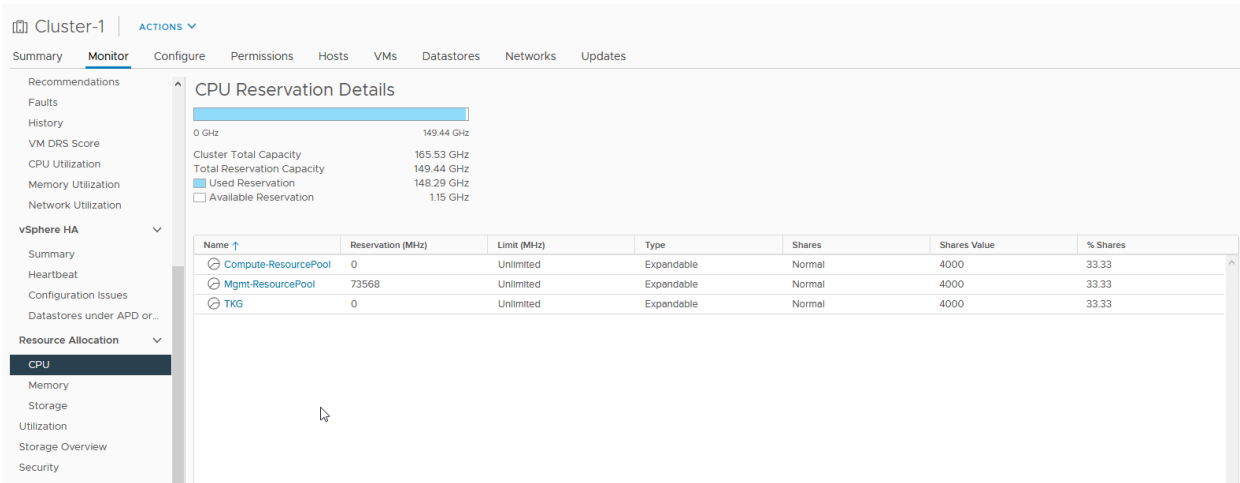
Figure 3: CPU Reservation Detailed information

Let us break down the CPU Reservation details from the screenshot above in a table format, again using the i3.metal as an example as this instance type has the least CPU cycles available compared to i3en and i4i.metal.

Table 1: CPU Reservation in a two hosts standard cluster. The math behind the Numbers.

Item	Value in MHz	Description
Single host i3.metal total CPU capacity	82,764	36 physical cores * 2299 MHz CPU Frequency
Single host available CPU Reservation capacity	74,719	Total CPU capacity – ESXi CPU Reservation
Two host cluster total CPU capacity	**165,528**	2 hosts * 36 cores * 2299 MHz CPU Frequency
Two host cluster available CPU Reservation capacity	**149,438**	2 hosts * Single host available CPU Reservation capacity
Management VMs CPU Reservation	73,568	Configured on Management VMs in the Mgmt-ResourcePool (Figure 5)
vSphere HA Admission Control CPU Reservation	74,719	Admission control is set to 50%. It reserves one host for failover (Figure 4). Admission Control CPU Reservation is equal to a single host available CPU reservation capacity (See screenshot below)
Used CPU Reservation	**148,287**	Management VMs CPU reservation + vSphere HA Admission Control CPU Reservation
Available CPU Reservation for workload VMs	**1151**	Two host cluster available CPU Reservation capacity – Used CPU reservation: 149,438 – 148,287

Figure 4: vSphere HA Admission Control configuration on a two host standard cluster
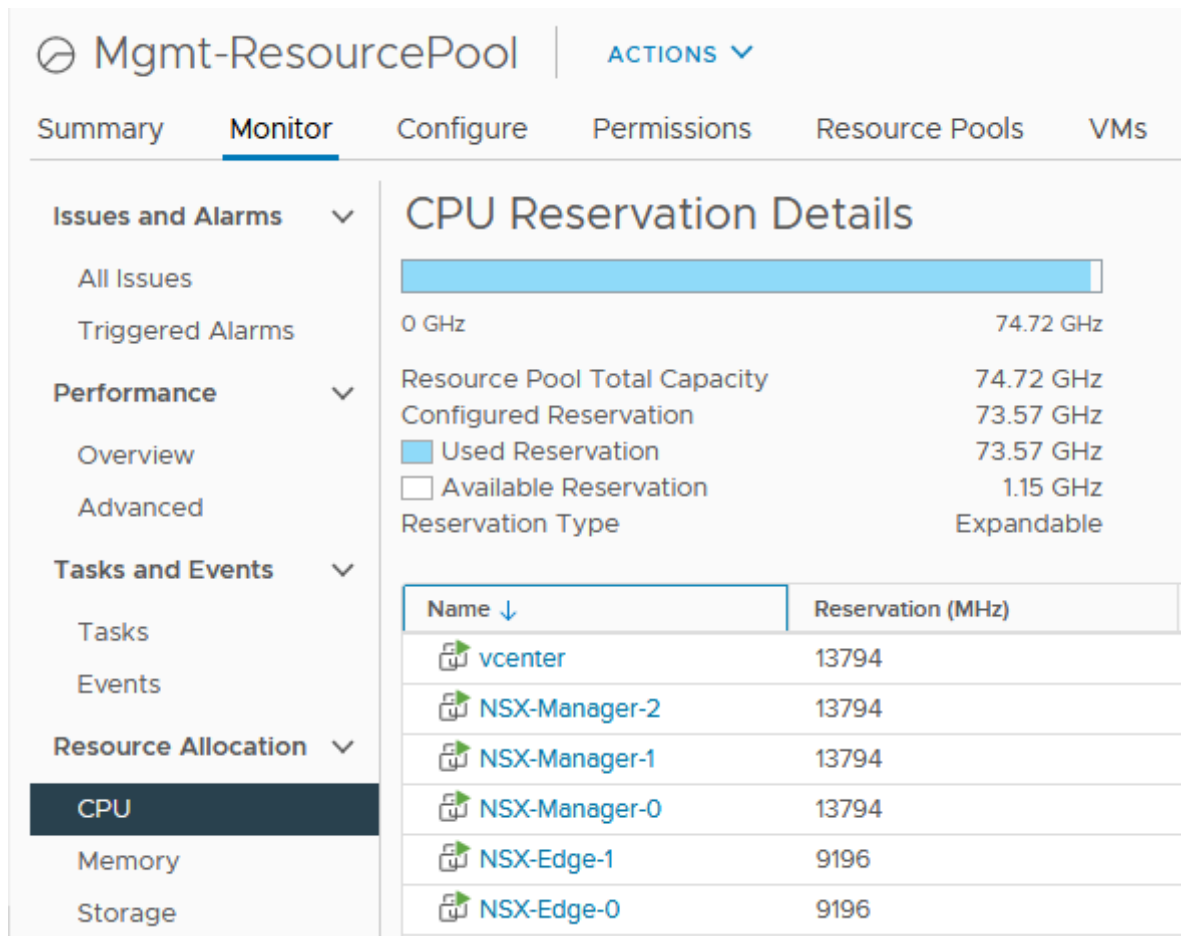
Figure 5: Management VMs CPU Reservation details

Available CPU Reservation determines the upper boundary of a maximum CPU Reservation for workloads VMs. 1151 MHz can be assigned to a single VM or split between multiple VMs for a total of 1151 MHz. In addition, vSphere HA reserves a default value of 32 MHz per powered-on VM. This 32 MHz per VM is acquired from the cluster available CPU Reservation. Knowing that available CPU Reservation is equal to 1151 MHz, we can easily calculate the maximum number of simultaneously powered on VMs by using the following expression: Round down ((Available CPU Reservation)/(default CPU Reservation per VM)).

Round (1151/32) = 35

If a powered-on VM has CPU Reservation, it will be deducted from the Available CPU Reservation. It will decrease the maximum number of powered-on VM by an increment 32 MHz. You can use the following expression to calculate the number of powered-on VMs if CPU Reservation is configured.

Round (((Available CPU Reservation – Total VM CPU Reservation)/(default CPU Reservation per VM)))

Integrated Services Compatibility

In general, you can deploy additional VMware solutions or VMware Cloud on AWS Integrated Services if the required CPU Reservation for the deployed VMs does not exceed 1151 MHz on selected cluster types. Setting up CPU Reservation to 1151 MHz would prohibit the ability to power on other VMs even without reservations. It's recommended to check the Add On specific documentation for more details.

VMware HCX

VMware HCX supports two-host and 1-1-1 stretched cluster SDDCs on a best-effort basis.
An HCX deployment in a two-node cluster is only supported for limited production or proof-of-concept purposes.

For SDDCs with i3.metal hosts, HCX installations have the following limitations:

- Single Site Pairing with one Service Mesh.
- One IX and one NE appliance.
- WAN Optimization is NOT supported.
- One stretched network.
- No CPU reservations in the HCX Cloud Manager Compute Profile.
- One concurrent migration of any type.
- Bulk or Cold migrations are recommended.
- vMotion, Replication-Assisted vMotion (RAV), and OSAM migrations require more resources and are not recommended.

With the above limitations, all HCX operations work in SDDCs with i3.metal hosts, assuming the resource consumption of the workloads is not too high. If an SDDC runs out of compute resources and the HCX infrastructure fails due to CPU contention, it may be necessary to scale up the SDDC to add more i3 hosts or scale down the workload allocation.

The above restrictions also apply to SDDCs with i3en hosts, except for the following:

- WAN Optimization can be used but it is NOT recommended.
- Bulk migrations are limited to six concurrent workflows.

Site Recovery

Deploying of the Site Recovery Add On on entry-level clusters is possible. Before activating the Site Recovery Add On, make sure that at least 64 MHz CPU Reservation is available in the primary cluster. The vSphere Replication ad Site Recovery Manager appliances does not have CPU Reservation set by default; hence, it would slightly reduce the Available CPU Reservation in the primary cluster for the value of 1087 MHz.

Note: The Deployment of a vRNI collector VM with required CPU Reservation is not possible on entry-level clusters limited to 1151 MHz CPU Reservation.

Recommended Use Cases

Cloud architects may found the following use cases helpful for their cloud design ad implementation.

- Use of primary two hosts' i3.metal cluster as a pure management cluster in a multicluster SDDC. No workload VMs are running in this cluster under normal operation. You isolate the management domain and significantly reduce the cost of the management cluster.

- Use of secondary i3/i3en/i4i.metal two hosts clusters to reduce costs for hosting business-critical applications with per-host licensing or Microsoft SPLA. This option can be combined with the custom core count feature to reduce further licensing costs (performance limitations apply).

- Initial cost-effective deployment for further scalability. You can start with just two hosts cluster and leverage production SLA, and grow this cluster as needed up to a supported maximum of 16 hosts.

- A cost-effective platform for Disaster Recovery. You can use entry-level clusters under normal operation and leverage eDRS in case of a DR event.

Note: if you want to leverage the Optimize for Rapid Scale-Out eDRS Policy, you must use at least three hosts cluster.

Summary

Entry-level clusters in VMware Cloud on AWS provide many design benefits and cost savings. Care should be taken when operating entry-level clusters based on i3.metal hosts not to surpass the available CPU reservation. This is of less concern when using i3en and i4i.metal clusters.

VMware Cloud on AWS is a rapidly evolving service. Please check this article and VMware Cloud on AWS Release Notes frequently for updates.