Solution Guide for Enabling Highly Resilient Kubernetes Workloads Using vSAN Stretched Clusters

Supporting VMware Tanzu Kubernetes Grid Integrated Edition (TKGI) on vSphere with NSX-T Data Center

VMware MAPBU

June 20, 2022

Legal Notice

© 2022 VMware, Inc. All rights reserved. This product is protected by U.S. and international copyright and intellectual property laws. This product is covered by one or more patents listed at http://www.vmware.com/download/patents.html.

VMware is a registered trademark or trademark of VMware, Inc. in the United States and/or other jurisdictions. All other marks and names mentioned herein may be trademarks of their respective companies.

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

Table of Contents

1. Introduction 5
2. Solution Requirements 6
2.1 Prerequisites and Best Practices 6
2.2 Software Versions 6
2.3 Other Requirements 6
3. Solution Architecture 7
3.1 Topology 1: Dedicated vSphere Clusters 7
3.2 Topology 2: Fully Collapsed vSphere Clusters 8
4. vSphere Clusters and Configurations 9
4.1 Management Cluster 9
4.2 Edge Cluster 9
4.3 Compute Cluster 9
4.4 Cluster Configuration 10
4.4.1 Configuration for all Clusters 10
4.4.2 vSAN Configuration 10
4.4.3 NSX-T Data Center Configuration 10
5. Solution Implementation 11
5.1 Configure Each vSphere Cluster 12
5.1.1 Enable vSphere DRS and Set to Fully Automated Mode 12
5.1.2 Enable vSphere HA 12
5.1.3 Configure vSAN 13
5.2 Configure the vSphere Management Cluster 15
5.2.1 Create vSphere DRS Rule 15
5.2.2 Create VM/Host Groups for DC1 15
5.2.3 Create VM/Host Groups for DC2 16
5.3 Configure the vSphere Edge Cluster 18
5.3.1 Create vSphere DRS Rule 18
5.3.2 Create VM/Host Groups for DC1 18
5.3.3 Create VM/Host Groups for DC2 19
5.4 Configure the vSphere Compute Cluster 21
5.4.1 Create vSphere DRS Rule 21
5.4.2 Create VM/Host Groups for DC1 21
5.4.3 Create VM/Host Groups for DC2 22
5.5 Configure the BOSH Director Tile 23
5.5.1 Configure Availability Zones with Host Groups 23
Configure the BOSH Resurrector 25

© 2022 VMware, Inc. All rights reserved.

5.6 Configure NSX-T Failure Domains 26
5.6.1 Configure NSX-T FDs 26
5.6.2 Verify NSX-T FDs 27
6. Production Considerations 29
7. Addendum: Testing Details 30
7.1 Testbed Topology 30
7.2 Hardware Specifications 31
7.3 Expected Behavior 31
7.4 Testing Configuration 31
7.5 Test Scenario 1: Site 2 down 32
Test Scenario 2: Site 2 up 33
7.6 Test Scenario 3: Site 1 down 34
7.7 Workflow 4: Site 1 up 35

Documentation Revision History

Version	Changes	Time
1.0	The first published version	Mar, 2020
1.1	Updated the product versions for TKGI 1.14 with CSI driver Notes: The UI screenshots in this doc are not updated, but they should be similar in the latest validated product versions.	Jun, 2022

© 2022 VMware, Inc. All rights reserved.

1. Introduction

This solution guide provides technical information for running Tanzu Kubernetes Grid Integrated Edition (TKGI) on vSphere with NSX-T in a vSAN Stretched Cluster environment for the purpose of supporting highly available Kubernetes workloads.

A vSAN Stretched Cluster is useful when you have 2 vSphere Data Center (DC) sites connected to each other over a high speed, low latency network link with at least 10 Gbps bandwidth. A vSphere Management Cluster, Edge Cluster, and Compute Cluster are extended across the 2 DCs with each DC running on top of a shared datastore that is powered by VMware vSAN technology.

The 2 DCs are in active/active mode, which means that vCenter, NSX-T, TKGI Management Plane components, and Kubernetes nodes are deployed across both DCs.

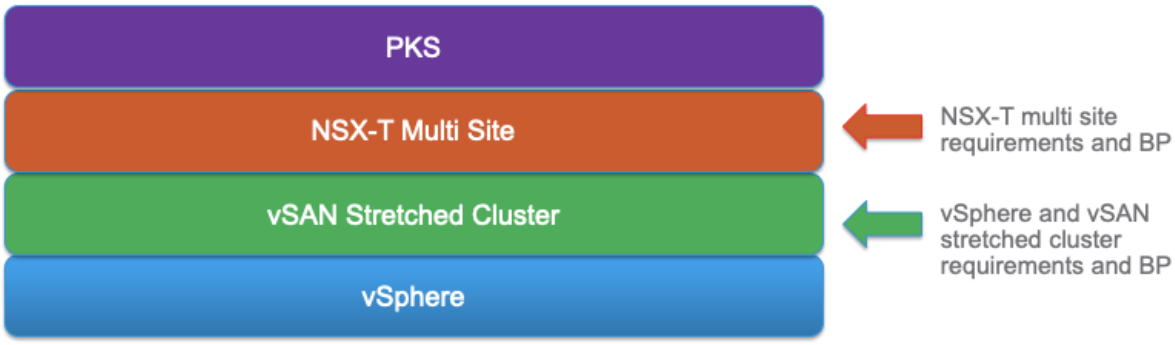
If either DC encounters an outage, workloads on that DC will be restarted on the other DC. Service disruption may appear during this phase, but once the automatic restart and any required manual repair work is completed, all workloads will resume normal operations.

. Before going into production with a vSAN Stretched Cluster supporting TKGI on vSphere with NSX-T with the goal of making Kubernetes workloads highly available, you must carefully read and understand all potential constraints of this solution. See the constraints listed in the “Production Considerations” section below

A vSAN Stretched Cluster topology supporting TKGI on vSphere with NSX-T is made possible by leveraging diverse functionality available with the underlying VMware infrastructure, including:

- vSphere DRS and HA
- vSAN Stretched Cluster
- NSX-T Failure Domain
- TKGI Host Group

The solution set forth in this guide is a multi-layer stack where each layer relies on the technology and services provided by the underlying layer(s). The diagram below illustrates the stack.



To implement the solution, the following requirements must be met, and the corresponding best practices must be implemented:

- vSphere and vSAN stretched cluster requirements and best practices
- NSX-T multi-site requirements and best practices

In addition, you must use only the supported software versions as listed in the “Software Versions” section below.

The steps necessary to configure TKGI in a vSAN Stretched Cluster topology are described in the “Solution Implementation” section below. This document’s recommended hardware specs and failover scenarios are based on observations made while testing and validating TKGI in this topology.

© 2022 VMware, Inc. All rights reserved.

2. Solution Requirements

This section lists the requirements for implementing the solution, including prerequisites and software versions.

2.1 Prerequisites and Best Practices

The following prerequisite requirements must be met, and the corresponding best practices must be implemented, for this solution and topology to be running and supported:

- Requirements for vSAN Stretched Clusters:
  - https://storagehub.vmware.com/t/vsan-stretched-cluster-guide/
- Best practices for implementing vSAN Stretched Clusters:
  - https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/techpaper/vmware-vsphere-metro-storage-cluster-recommended-practices-white-paper.pdf
  - When using automatic deployed vSphere CSI plugin with TKGI, please follow the Design Considerations and Best Practices from CSI documentation.
- Requirements and best practices for running NSX-T in a multi-site environment:
  - NSX-T 2.5: https://docs.vmware.com/en/VMware-NSX-T-Data-Center/2.5/administration/GUID-5D7E3D43-6497-4273-99C1-77613C36AD75.html
  - NSX-T 3.1: https://docs.vmware.com/en/VMware-NSX-T-Data-Center/3.1/administration/GUID-5D7E3D43-6497-4273-99C1-77613C36AD75.html

Any condition that does not comply with these requirements and best practices will result in the non-support of an vSAN Stretched Cluster topology for TKGI on vSphere with NSX-T.

2.2 Software Versions

To implement the solution documented in this guide, you must adhere to the following software versions. No other software versions are supported.

Component	Version (with VCP in-tree driver)	Version (with vSphere CSI Plugin)
vSphere	**6.7 U3+**	**7.0 U3d +**
TKGI	**1.7.0+**	**1.14.0 +**
Ops Manager	**2.8.5+**	**2.10.39 +**
NSX-T Data Center	**2.5.1+**	**3.1.3 +**

2.3 Other Requirements

- TKGI must be deployed with a single node MySQL Database
- TKGI 1.14 supports only the automatically deployed vSphere CSI plugin

© 2022 VMware, Inc. All rights reserved.

3. Solution Architecture

The underlying architecture for the topologies supporting this solution requires 2 Data Centers (DC). Both DC sites are connected by a high speed, low latency link supporting at least 10 Mbps bandwidth. The networking and latency between the 2 locations must meet the vSAN Stretched Cluster requirements.

For this solution, there are two supported topologies:

- Dedicated vSphere Clusters for Management, Edge, and Compute
- Fully Collapsed vSphere Clusters

Each topology is described and depicted below.

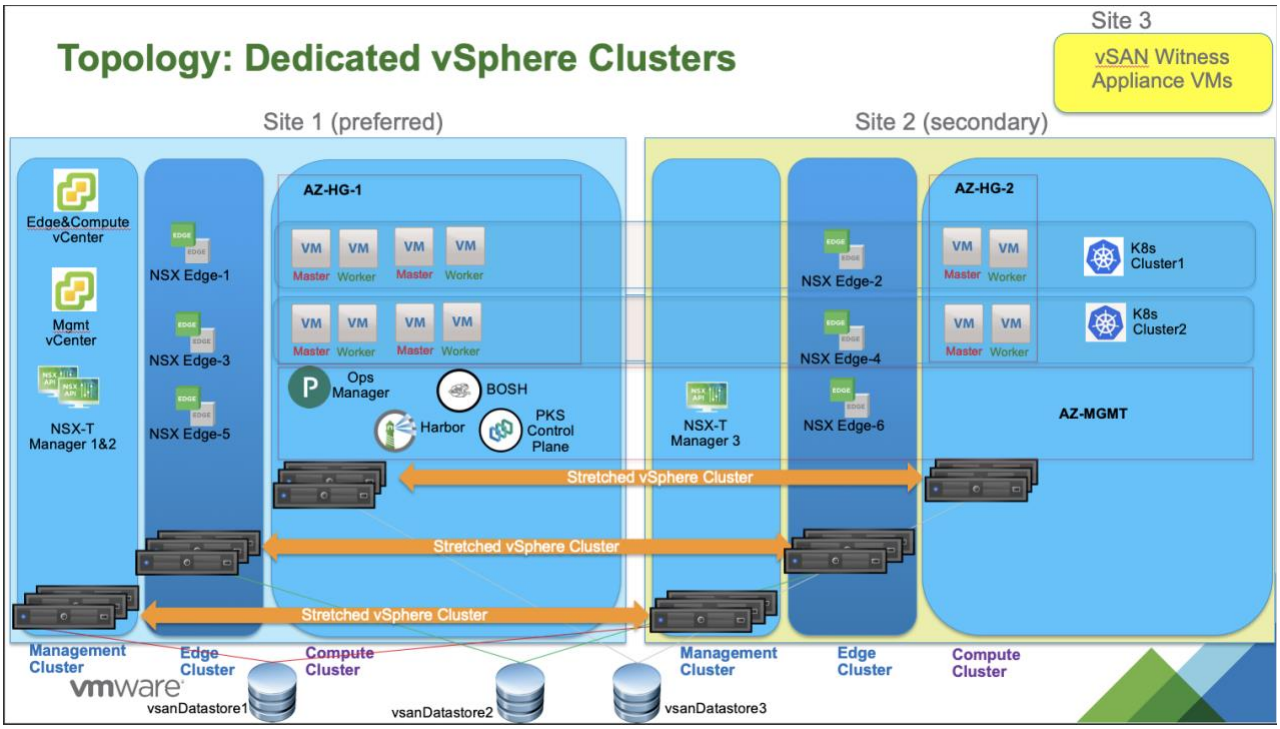
3.1 Topology 1: Dedicated vSphere Clusters

In this configuration, all 3 vSphere clusters (Management, Edge and Compute) are separated. Each cluster has a minimum of 6 ESXi host members dispatched (stretched) across the 2 sites:

- 3 ESXi hosts minimum in Site 1 (or DC1)
- 3 ESXi hosts minimum in Site 2 (or DC2)

The total number of ESXi hosts required for this topology is 18.

Topology 1: Dedicated vSphere Clusters



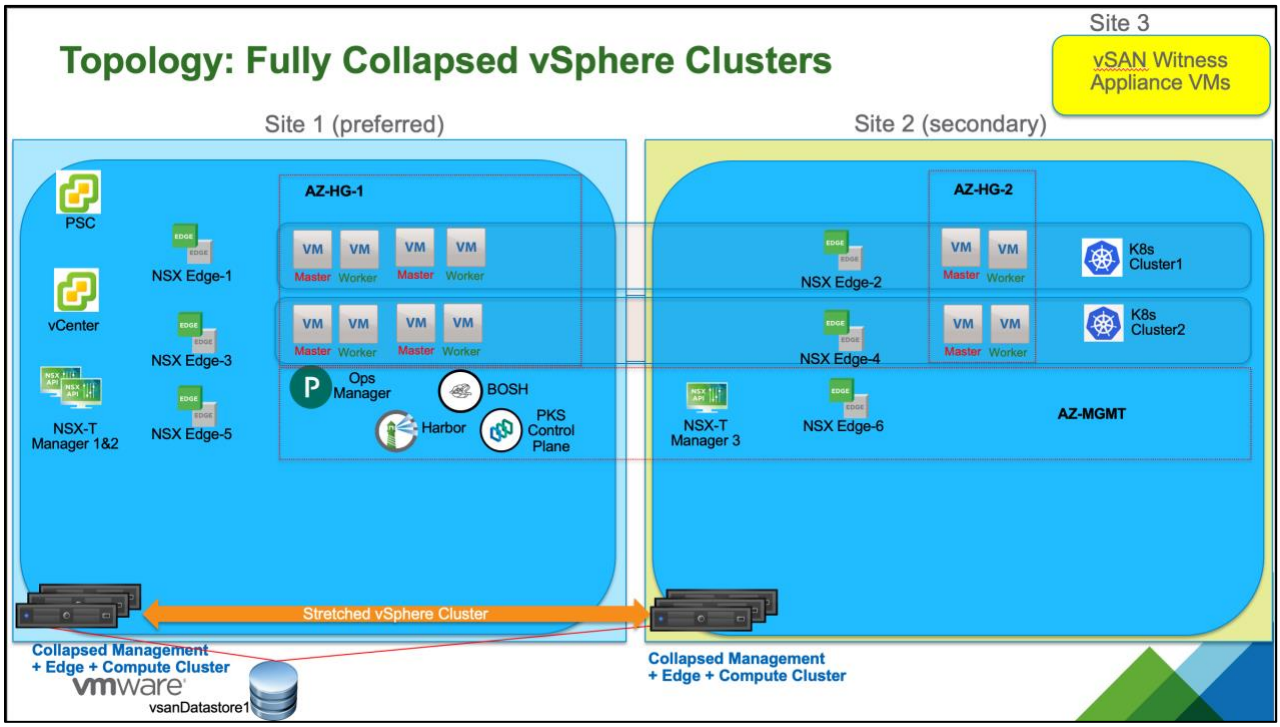
© 2022 VMware, Inc. All rights reserved.

3.2 Topology 2: Fully Collapsed vSphere Clusters

In this topology, the 3 vSphere clusters (Management, Edge and Compute) are collapsed into a single vSphere cluster.

The benefit of this configuration is the total number of ESXi hosts required to run the solution is reduced from 18 to 6. The minimum number of ESXi hosts per DC site remains the same (3).

Topology 2: Fully Collapsed vSphere Clusters



© 2022 VMware, Inc. All rights reserved.

4. vSphere Clusters and Configurations

This section describes the characteristics and details of each vSphere cluster in this solution.

4.1 Management Cluster

The vSphere **Management Cluster** includes 2 vCenter instances and 3 NSX-T manager instances.

Management Cluster characteristics:

- 2 vCenter instances:
  - VC-01: vCenter for MGMT cluster and
  - VC-02: vCenter for Edge and Compute cluster
- 3 NSX Manager instances
  - With a VIP or load balancer fronting the 3 VMs

Management Cluster details:

- Both vCenter instances are pinned to DC1 using a DRS should rule (described in the “Solution Implementation” section below)
- NSX-T Managers 1 and 2 are pinned to DC1 using a DRS should rule
- NSX-T Manager 3 is pinned to DC2 using a DRS should rule
- vCenter access to the Management Cluster is done using the VC-01 instance

4.2 Edge Cluster

The vSphere Edge Cluster contains all NSX-T Edge Node VMs.

Edge Cluster characteristics:

- 6 NSX Edge Node VMs
- All Edge Nodes are part of the same NSX Edge Cluster

Edge Cluster details:

- Edge Node VMs 1, 3 and 5 are pinned to DC1 using a DRS must rule (described in the “Solution Implementation” section below)
- Edge Node VMs 2, 4 and 6 are pinned to DC2 using a DRS must rule
- vCenter access to the Edge Cluster is done using the VC-02 instance

4.3 Compute Cluster

The vSphere Compute Cluster contains TKGI Management Plane VMs and Kubernetes cluster nodes.

Compute Cluster characteristics:

- TKGI Management Plane VMs: TKGI API, TKGI DB, Ops Manager, BOSH Director, Harbor Registry
- Kubernetes cluster control plane and nodes have been provisioned by TKGI

Compute Cluster details:

- All TKGI Management Plane VMs are pinned to DC1 using a DRS should rule (see the “Solution Implementation” section below for details)

© 2022 VMware, Inc. All rights reserved.

- Kubernetes cluster nodes are dispatched across Host Group 1 (AZ-HG-1 in DC1) and Host Group 2 (AZ-HG-2 in DC2)
- vCenter access to the Compute Cluster is done using the VC-02 instance

4.4 Cluster Configuration

This subsection includes configuration details for the vSphere clusters.

4.4.1 Configuration for all Clusters

For both topologies and each vSphere cluster, you must make sure there is enough capacity to support a site down failure situation. This means that all 3 vSphere clusters must be provisioned with enough capacity to support an event where half of the cluster is down. At any time, any workload on any vSphere cluster must be able to run properly on any side of the topology (that is, in DC site 1 or in DC site 2).

4.4.2 vSAN Configuration

This section lists the specific vSAN configurations that must be set for this solution to work and be supported.

- vSAN is enabled on each of the 3 vSphere clusters with the Configuration type set to “Stretched Cluster”. 2 vSAN Fault Domains are created for each cluster to allow for this.
- DC Site 1 is configured as the preferred site and DC Site 2 is configured as the secondary site. A third DC (site 3) hosts the vSAN witness instance.
- vSAN will perform data replication across the 2 sites AND will also perform local data replication within a site as well.

4.4.3 NSX-T Data Center Configuration

This section lists the specific configurations required for NSX-T Data Center.

- The Tier-0 Router is configured in Active/Standby mode.
- There is a total of 6 Edge Nodes (1 per ESXi host) and all Edge Nodes are part of the same Edge Cluster. (Note that the number of Edge Nodes can increase over time.)
- Edge Node 1 is hosting the active T0 instance while Edge Node 2 is hosting the standby instance.
- 2 NSX-T failure domains are created,
- The Edge Nodes are dispatched across the 2 sites as follows:

Site1 (preferred) Failure Domain 1: AZ-HG-1	Site2 (Secondary) Failure Domain 2: AZ-HG-2
Edge Node 1	Edge Node 2
Edge Node 3	Edge Node 4
Edge Node 5	Edge Node 6

- Each ESXi host in the compute cluster has 4 PNICs:
  - 2 allocated for VDS (for vSphere traffic like management, vMotion, vSAN storage).
  - 2 allocated for N-VDS.

However, if you only have 2 PNICs available for each ESXi host, refer to the following documentation for configuration details: Fully Collapsed vSphere Cluster NSX-T Deployment.

© 2022 VMware, Inc. All rights reserved.

5. Solution Implementation

Implementing the overall solution requires multiple steps for each vSphere cluster, vSAN, NSX-T and then TKGI. Listed below is a summary of the required configurations. Detailed instructions for each follow.

- **VMware DRS is enabled and set to fully automated mode for each vSphere Cluster.** This is done by right clicking the cluster object in the vSphere Client and selecting Turn On VMware DRS.
- **VMware HA is enabled for each vSphere Cluster.** This is done by right clicking the cluster object in the vSphere Client and selecting Turn On VMware HA.
- **BOSH resurrector is disabled.** The BOSH resurrector automatically recreates VMs in case of an outage. The BOSH resurrector must be disabled because in the event of a DC site failure, you want vSphere HA to control the restart of all the VMs on the failed site. If BOSH resurrector is enabled, it will conflict with vSphere HA and attempt to resurrect the unresponsive VMs. This could create a race condition a non-deterministic outcome: BOSH deletes a VM before recreating it, which may conflict with vSphere HA trying to restart the VM.
- **NSX-T Failure Domain is configured.** When creating the Tier-1 Routers for TKGI-provisioned Kubernetes clusters, the active instance of the Tier-1 Router is placed on an Edge Node located in Failure Domain 1 (FD1). The standby instance of the Tier-1 Router is placed on a different Edge Node located in Failure Domain 2 (FD2). Note: this is true only if the Tier-1 Router hosts a stateful service (like NAT rules or a load balancer). Otherwise, the Tier-1 Router is instantiated inside the ESXi Transport Nodes. As FD1 and FD2 are in their respective site, this means that if a site is down, this would impact only the instance of the Tier-1 Router located there – therefore enabling the T1 to switchover the active state to the remaining site.
- **vSAN Storage Policy is used to support a dual site topology.** Using a vSAN storage policy makes it possible to instruct vSAN to perform a copy of the data across the 2 sites in addition to a copy that stays local to the site.
- **BOSH DRS rule is set to "should."** The BOSH CPI automatically creates a VMware DRS rule when vSphere Host Group functionality is used for TKGI. By default, this rule is set to "must." However, to support a vSAN Stretched Cluster architecture, you must manually change this rule to "should" so that Kubernetes nodes can be restarted on the other DC site if the DC site where the nodes are currently running suffers an outage.

© 2022 VMware, Inc. All rights reserved.

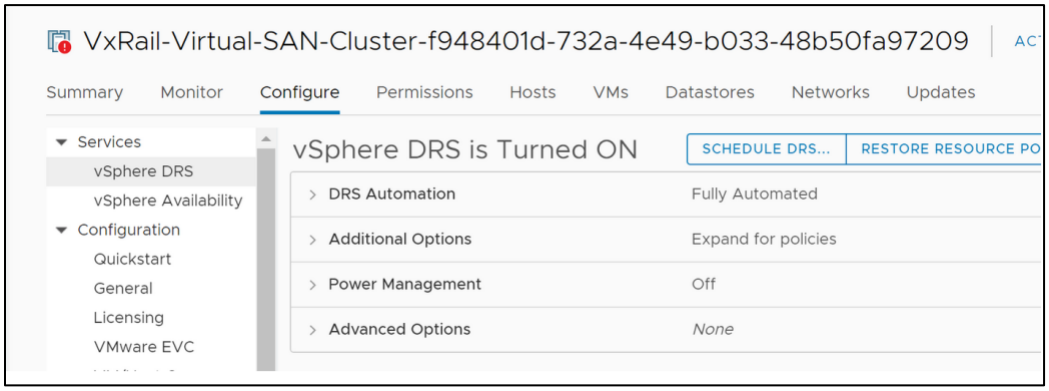
5.1 Configure Each vSphere Cluster

The following configuration settings apply to the vSphere clusters:

- Management Cluster
- Edge Cluster
- Compute Cluster

5.1.1 Enable vSphere DRS and Set to Fully Automated Mode

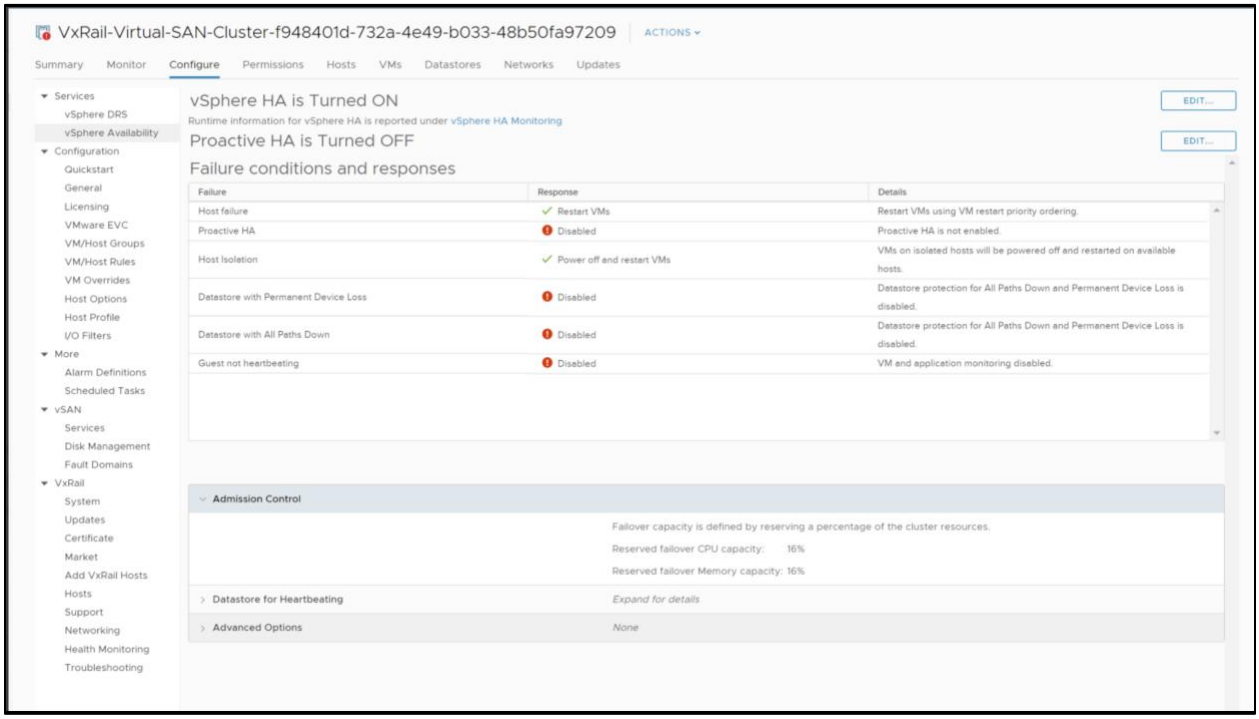
Make sure DRS is turned on and set to fully automated in the vSphere compute cluster as shown below. The default parameters for the other settings are fine.



In addition to its traditional role of distributing workload across available ESXi hosts based on their free resources, DRS allows VMs to move back to their original location upon a failed site recovery case.

5.1.2 Enable vSphere HA

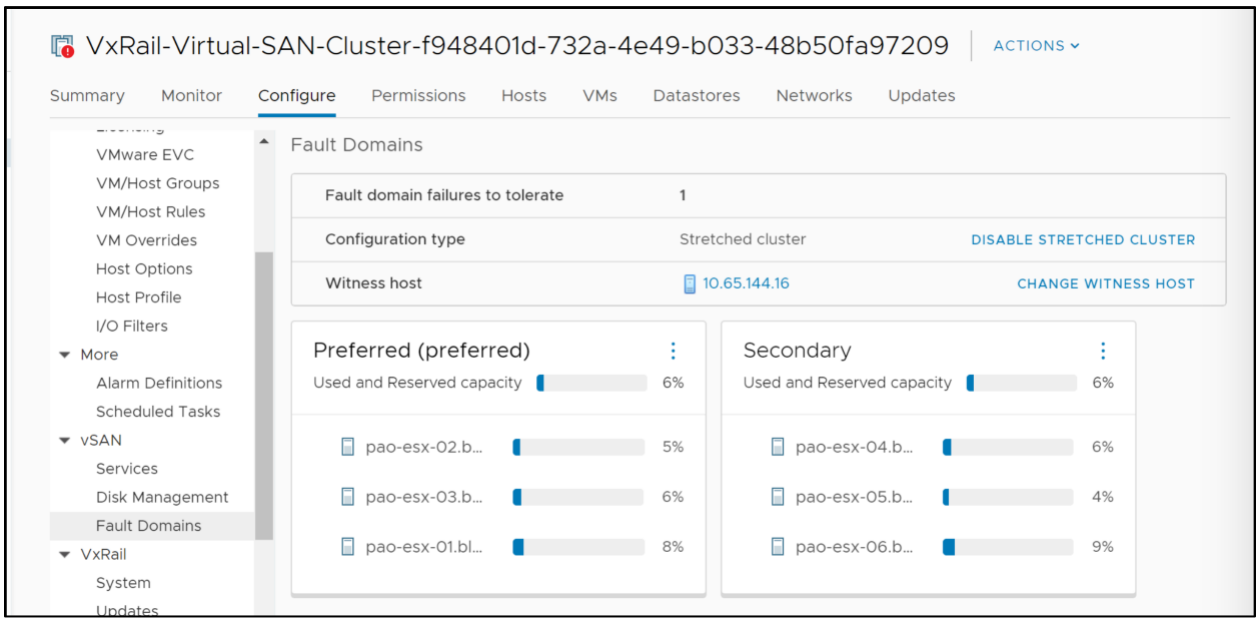
Make sure HA is turned on as shown below. The default parameters for the other settings are adequate.



© 2022 VMware, Inc. All rights reserved.

5.1.3 Configure vSAN

Configure the Management Cluster, Edge Cluster and Compute Clusters with vSAN Fault Domain using DC1 as preferred site and DC2 as secondary site:

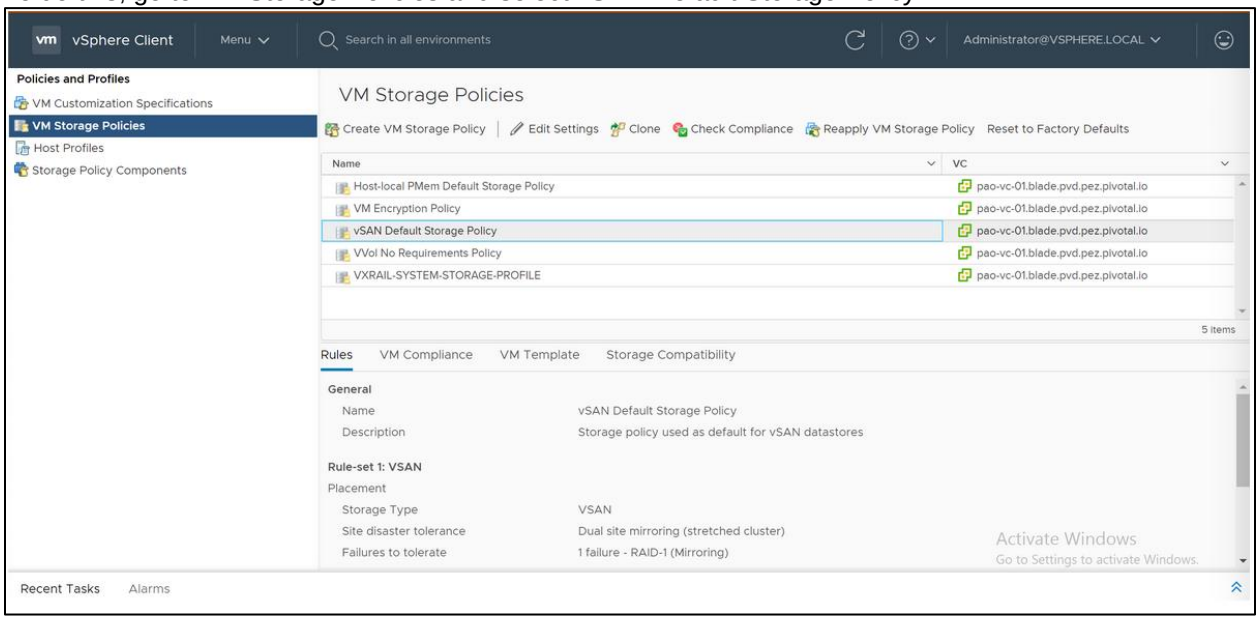


Configure vSAN to handle a dual site topology (replicate data across the 2 sites) and then perform a local data replication as well. Modify default vSAN SPBM with the following parameters:

–PFTT = 1 (Primary FTT)

–SFTT = 1 (Secondary FTT)

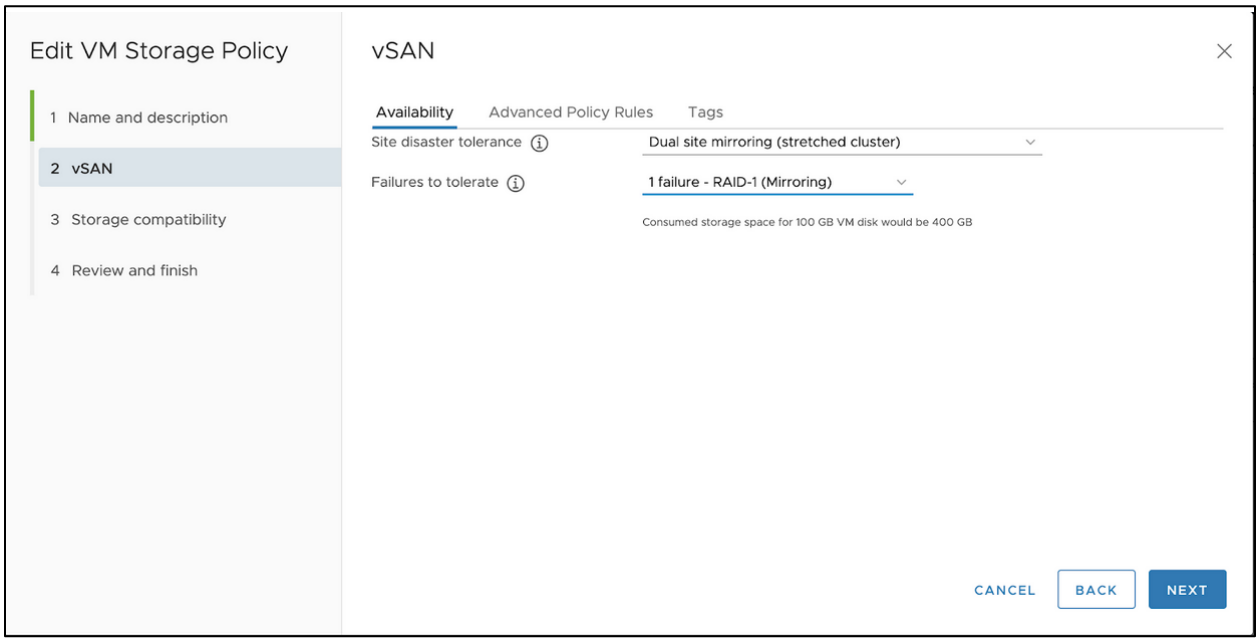
To do this, go to VM Storage Policies and select vSAN Default Storage Policy:



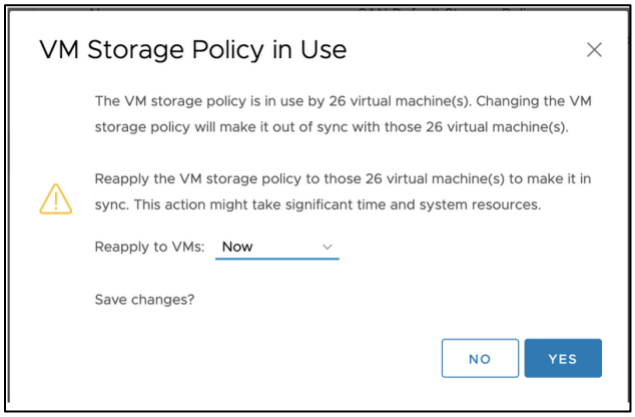
© 2022 VMware, Inc. All rights reserved.

Click Edit and select the parameters as shown below:

- Site disaster tolerance: Dual Site Mirroring (stretched cluster)
- Failures to tolerate: 1 failure - RAID-1 (Mirroring)



Click on Next several times to complete the configuration. You will see the following message at the end of the process. Select Reapply to VMs: Now and click Yes.



Make sure all VMs on the system are using this storage policy and they are all in compliancy with it.

© 2022 VMware, Inc. All rights reserved.

5.2 Configure the vSphere Management Cluster

This section provides configuration instructions for the vSphere Management Cluster.

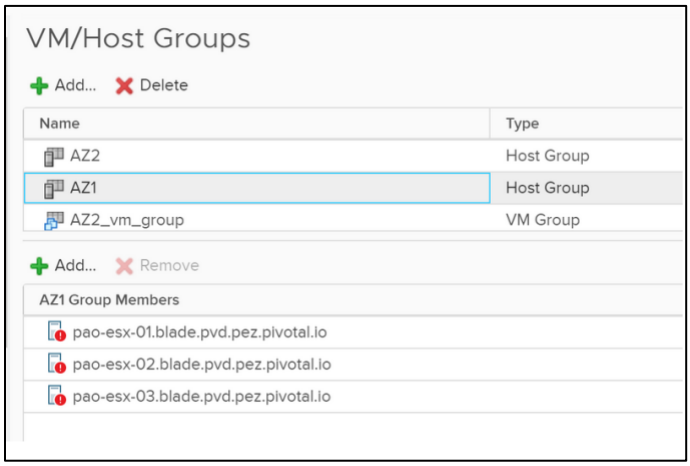
5.2.1 Create vSphere DRS Rule

Create DRS rules (with 'should' policy) to implement the following behavior:

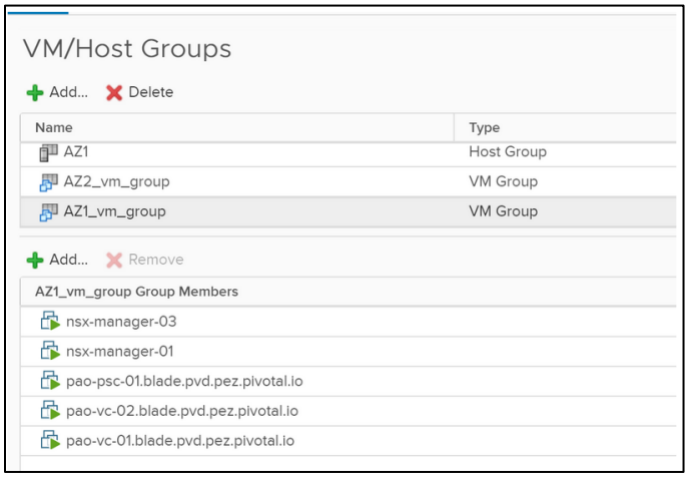
- Both vCenter instances (VC-01 and VC-02) are pinned to DC1
- NSX-T managers 1 & 3 pinned to DC1
- NSX-T manager 2 pinned to DC2

5.2.2 Create VM/Host Groups for DC1

Create Host Group AZ1 which contains all ESXi hosts in DC1:

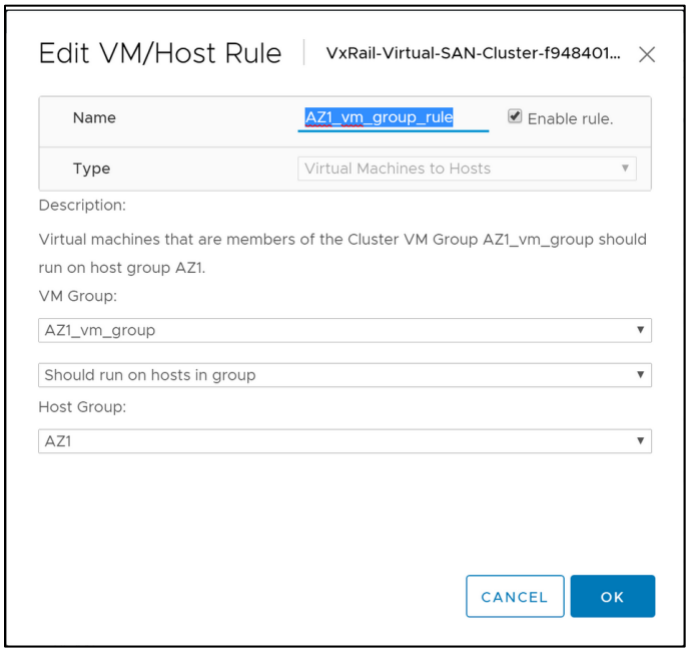


Create VM Group AZ1_vm_group contains all vCenter and NSX-T managers instances 1 and 3:



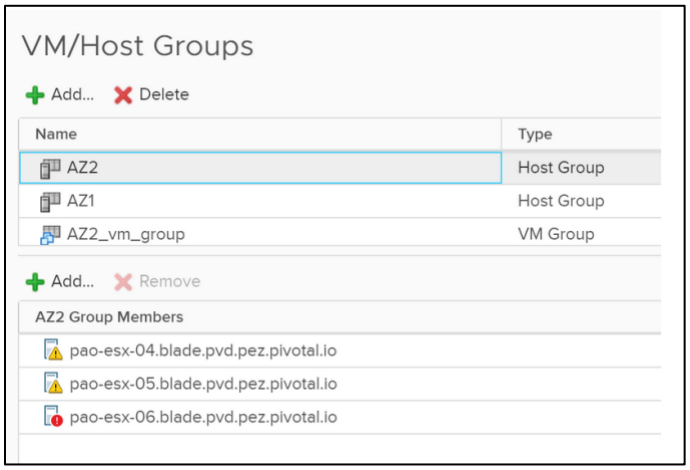
© 2022 VMware, Inc. All rights reserved.

Create the VM/Host Rule for AZ1_vm_group and link the 2 above objects using a ‘should’ policy:



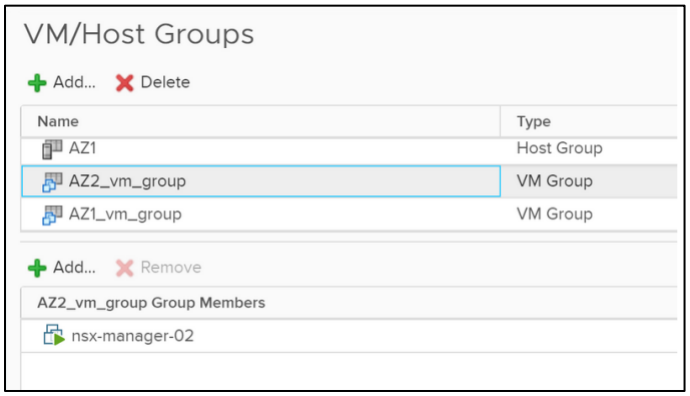
5.2.3 Create VM/Host Groups for DC2

Create Host Group AZ2 which contains all ESXi hosts in DC2:

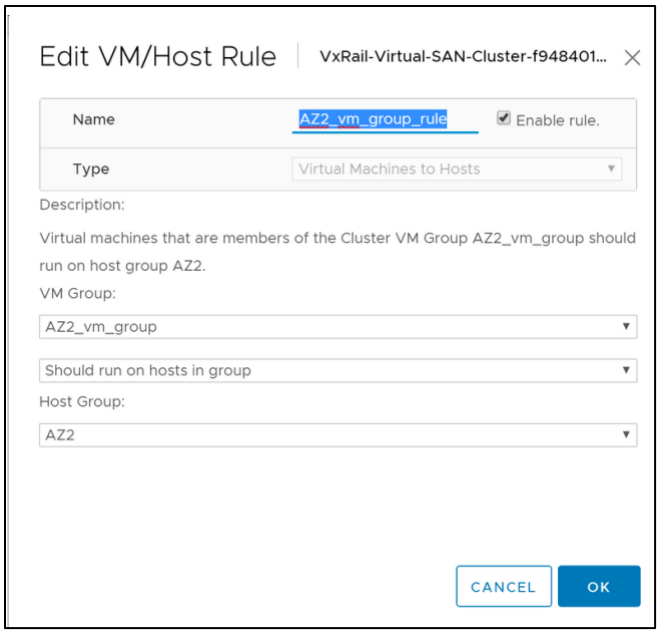


© 2022 VMware, Inc. All rights reserved.

Create VM Group AZ2_vm_group which contains NSX-T manager instance 2:



Create the VM/Host Rule and link the 2 above objects using a ‘should’ policy:



© 2022 VMware, Inc. All rights reserved.

5.3 Configure the vSphere Edge Cluster

This section provides instructions for configuring the vSphere Edge Cluster.

5.3.1 Create vSphere DRS Rule

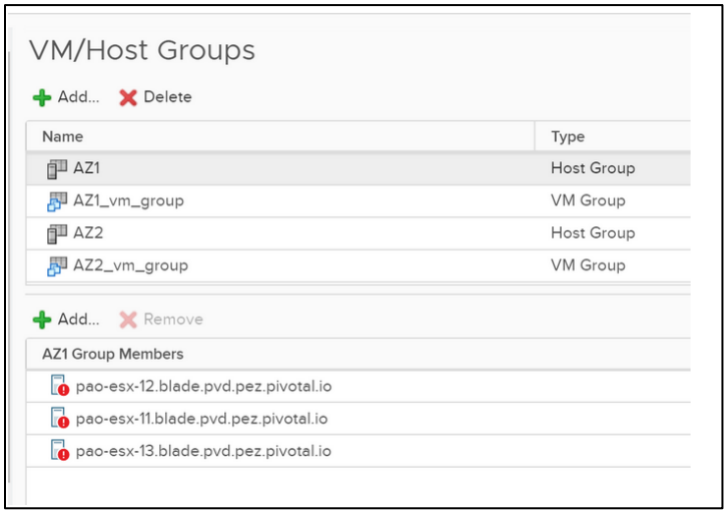
vSphere DRS rule:

Create DRS rules (with 'must' policy) to implement the following behavior:

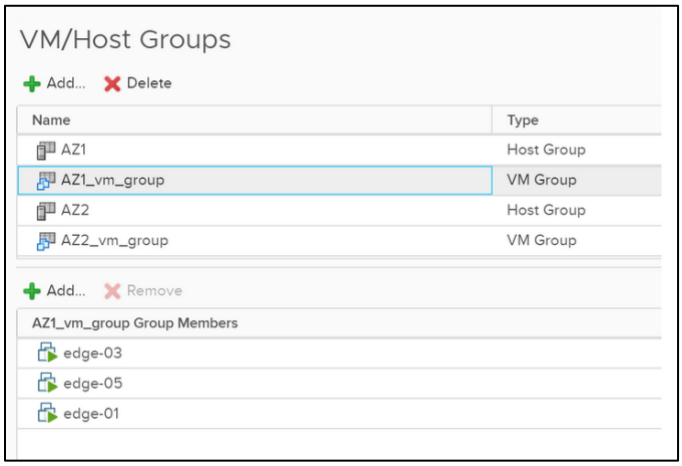
- Edge VM 1, 3 & 5 pinned to DC1
- Edge VM 2, 4 & 6 pinned to DC2

5.3.2 Create VM/Host Groups for DC1

Create Host Group AZ1 and add all ESXi hosts in DC1:

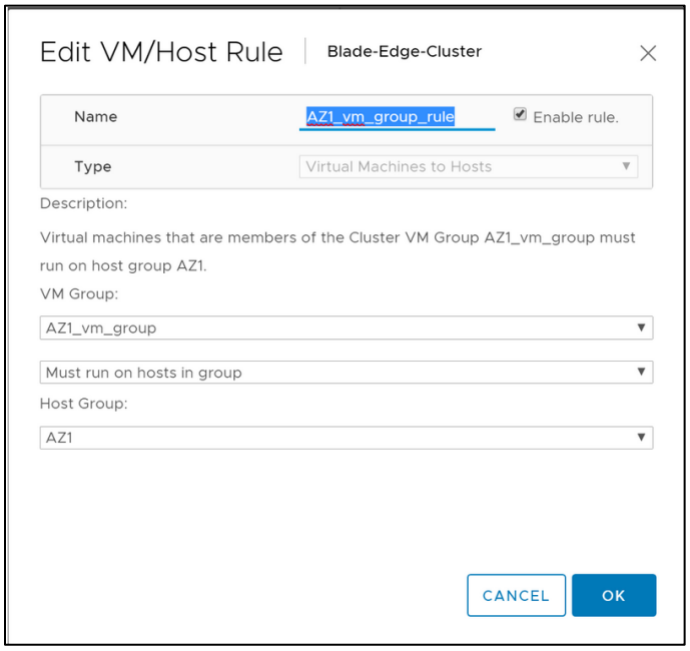


Create VM Group AZ1_vm_group contains all Edge Nodes located in DC1 (Edge Node 1, 3 and 5):



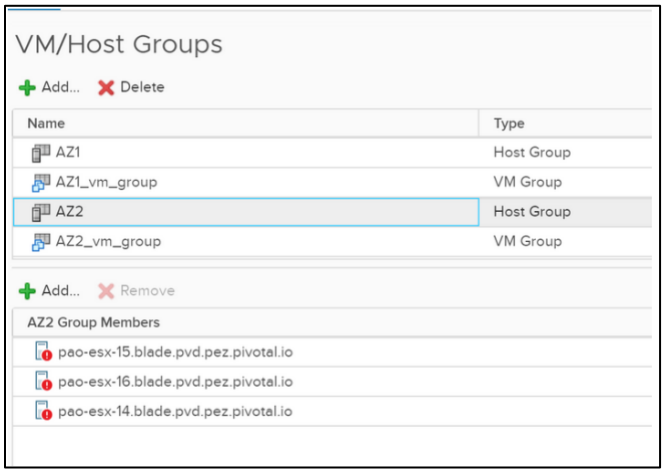
© 2022 VMware, Inc. All rights reserved.

Configure the VM/Host Rule to link the 2 above objects using a ‘’must’ policy:



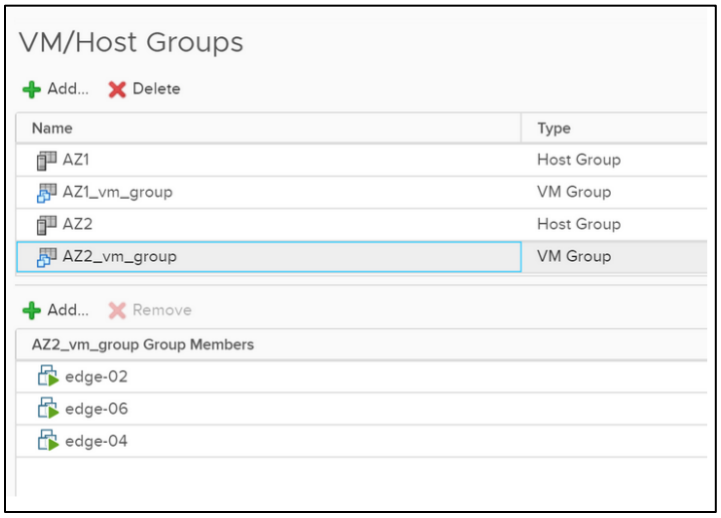
5.3.3 Create VM/Host Groups for DC2

Create Host Group AZ2 which contains all ESXi hosts in DC2:

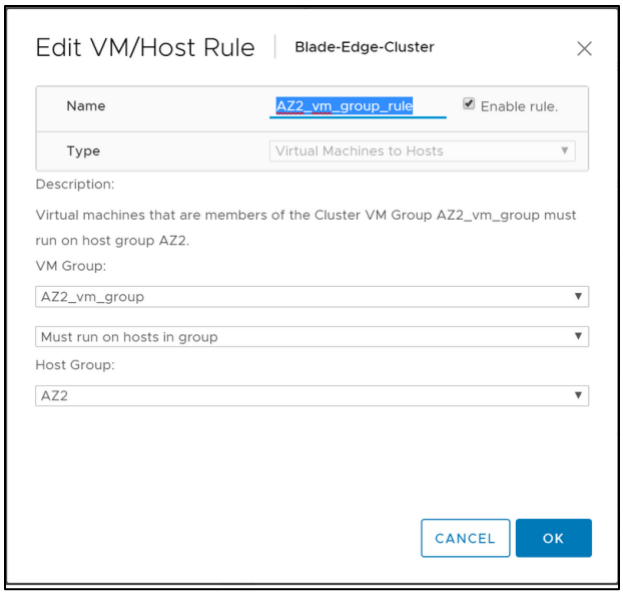


© 2022 VMware, Inc. All rights reserved.

Create the VM Group AZ2_vm_group which contains all Edge Nodes in DC2 (Edge Node 2, 4 and 6):



Configure the VM/Host Rule to link the 2 above objects using a ‘’must’ policy:



© 2022 VMware, Inc. All rights reserved.

5.4 Configure the vSphere Compute Cluster

This section provides instructions for configuring the vSphere Compute Cluster.

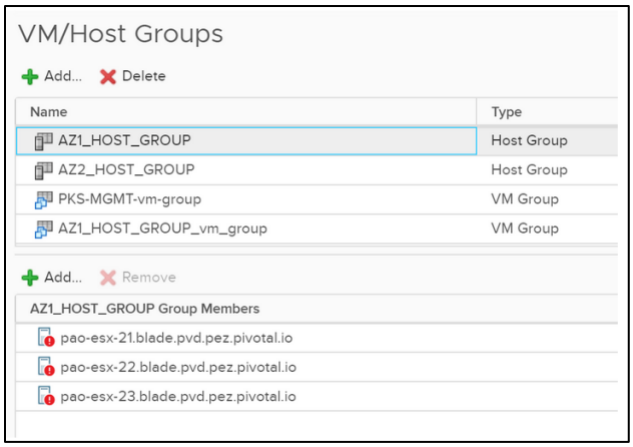
5.4.1 Create vSphere DRS Rule

Create DRS rules (with 'must' policy) to implement the following behavior:

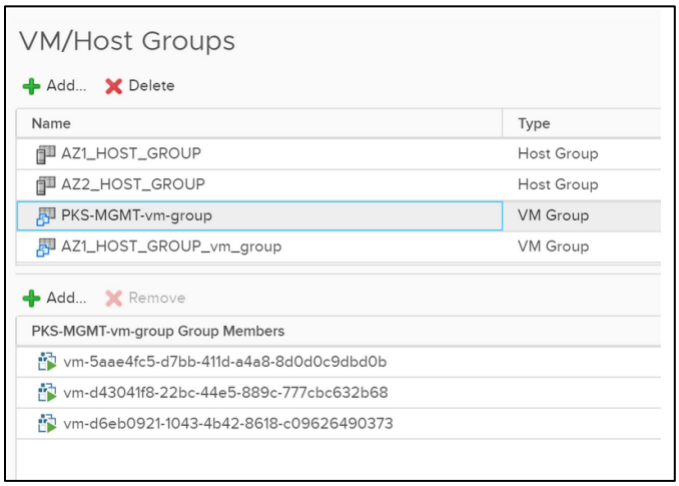
- TKGI Management Plane VMs are pinned to DC1 (Ops Manager, BOSH, TKGI API VM, TKGI DB VM, and Harbor)

5.4.2 Create VM/Host Groups for DC1

Create AZ1_HOST_GROUP: This host group contains all ESXi hosts in DC1.

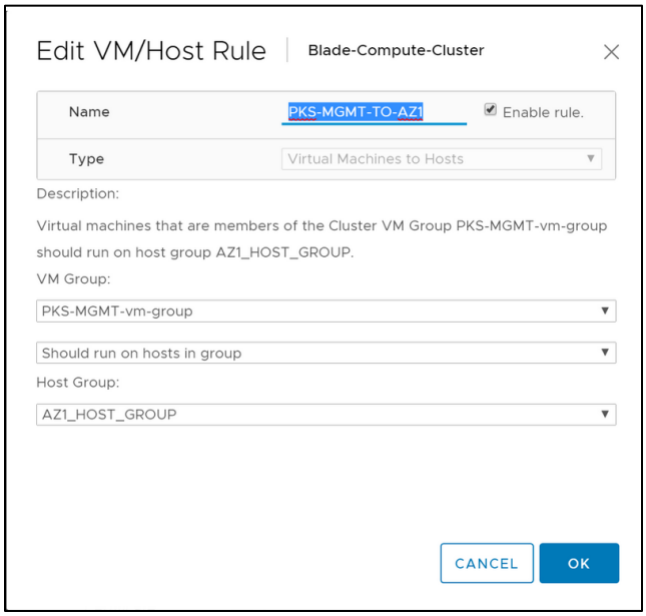


Create VM Group AZ1_vm_group which contains all TKGI management plane VMs:



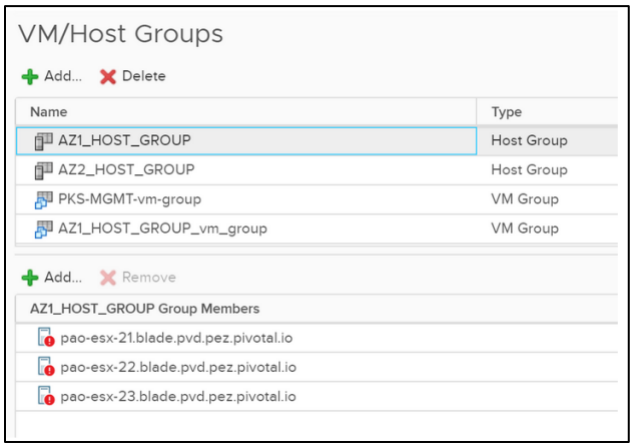
© 2022 VMware, Inc. All rights reserved.

Configure the VM/Host Rule links the 2 above objects using a ‘’should’ policy:



5.4.3 Create VM/Host Groups for DC2

Create AZ2_HOST_GROUP. This host group contains all ESXi hosts in DC2.



© 2022 VMware, Inc. All rights reserved.

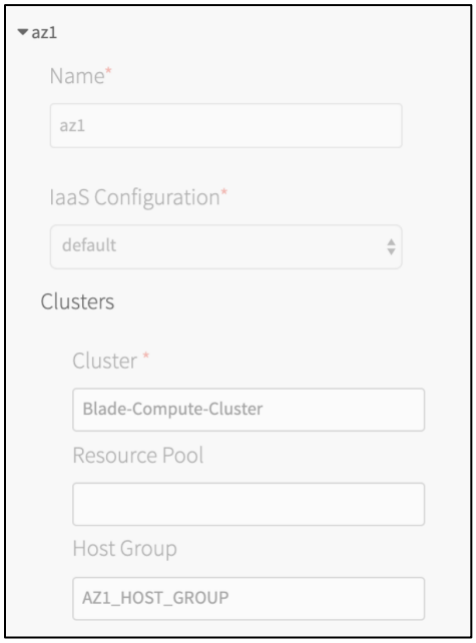
5.5 Configure the BOSH Director Tile

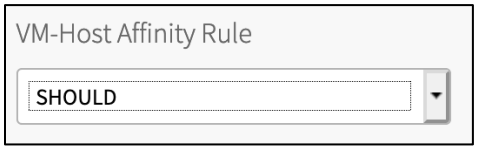
In Bosh tile, configure two AZs (1 per site) as shown below.

5.5.1 Configure Availability Zones with Host Groups

AZ1:

- The cluster field relates to vSphere compute cluster (Blade-Compute-Cluster).
- The Host Group field relates to the vSphere Host Group created in the vSphere compute cluster (AZ1_HOST_GROUP).
- Set the VM-Host Affinity Rule to be SHOULD.

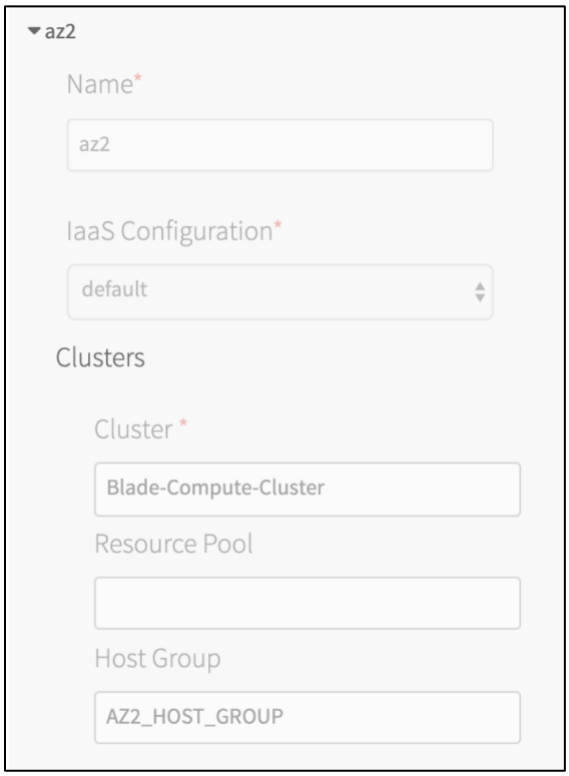


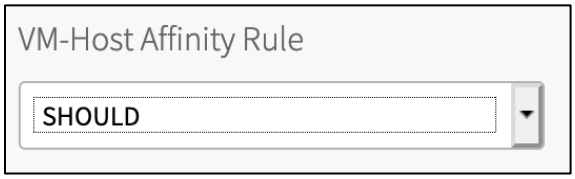


© 2022 VMware, Inc. All rights reserved.

AZ2:

- The cluster field relates to vSphere compute cluster (Blade-Compute-Cluster).
- The Host Group field relates to the vSphere Host Group created in the vSphere compute cluster (AZ2_HOST_GROUP).
- Set the VM-Host Affinity Rule to be SHOULD.





© 2022 VMware, Inc. All rights reserved.

Configure the BOSH Resurrector

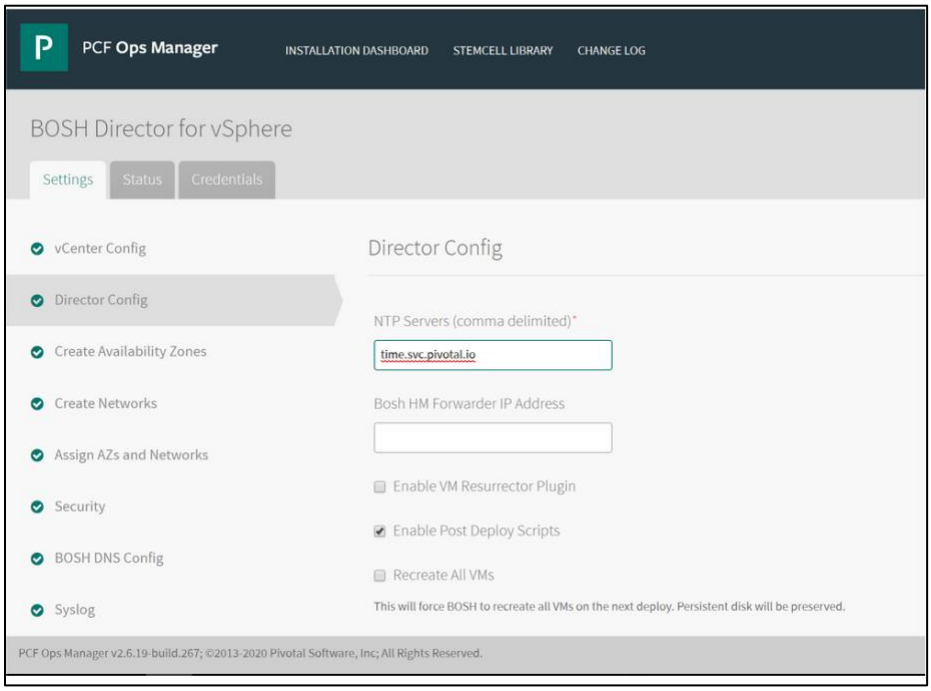
For Bosh to work in a dual site environment, Bosh resurrector MUST be disabled.

During a site down event occurrence:

- vSphere HA must restart all VMs from that site to the other site.
- If BOSH resurrector is enabled, Bosh will try to resurrect VMs by deleting the VMs before recreating them.

A race condition results if both vSphere HA and BOSH resurrector attempt to resolve the site down occurrence. vSphere HA must not conflict with any Bosh tentative resolution actions. To guarantee this, disable Bosh resurrector.

In order to disable Bosh resurrector, use Ops Manager and go to Bosh tile -> Director Config and make sure "Enable VM Resurrector Plugin" is unchecked.



© 2022 VMware, Inc. All rights reserved.

5.6 Configure NSX-T Failure Domains

NSX-T Failure Domains will allow you to correctly handle site failure occurrences for stateful T0 and T1 logical routers. When deploying T1 routers with embedded services, the active instance of the router will land on 1 FD and the passive instance will land on the other FD. For the T0 router, selecting the active and passive Edge Node instance is done at the creation phase (when creating the uplink interfaces). Make sure that those 2 Edge Nodes belong to 2 different FD.

This means that if a site fails down, NSX-T FD will guarantee that a backup instance of the T0 or T1 router live on the other site, and this instance can switch over to the active state if needed.

See the NSX-T documentation for more details.

5.6.1 Configure NSX-T FDs

Create 2 NSX-T Failure Domains and place the Edge Node VMs in their respective NSX-T FD:

- Create the following NSX-T Failure Domains: AZ1 and AZ2
- Place the following Edge Nodes in the AZ1 Failure Domain: Edge Node 1, Edge Node 3, Edge Node 5
- Place the following Edge Nodes in the AZ2 Failure Domain: Edge Node 2, Edge Node 4, Edge Node 6

Site1 (preferred) NSX-T FD AZ1	Site2 (Secondary) NSX-T FD AZ2
Edge Node 1	Edge Node 2
Edge Node 3	Edge Node 4
Edge Node 5	Edge Node 6

© 2022 VMware, Inc. All rights reserved.

5.6.2 Verify NSX-T FDs

You can check the FD configuration and the mapping of an Edge Node into a FD using REST API against NSX-T manager instance:

```
root@pks-client-vm:~/NSX-T/1-NSX-T-FD# curl -k -u admin:xxxxxxxxxxx -X GET 'https://pao-nsxmgr.blade.pvd.pez.pivotal.io/api/v1/failure-domains'
{
  "results" : [ {
    "resource_type" : "FailureDomain",
    "id" : "1db43acb-a708-4ea3-8fb1-01a3bd98a9dc",
    "display_name" : "AZ2",
    "_create_user" : "admin",
            <SNIP>
  }, {
    "resource_type" : "FailureDomain",
    "id" : "4fc1e3b0-1cd4-4339-86c8-f76baddbaafb",
    "display_name" : "system-default-failure-domain",
    "_create_user" : "system",
            <SNIP>
  }, {
    "resource_type" : "FailureDomain",
    "id" : "ce0caa6c-b57f-4766-bdf3-c30ef26b0a27",
    "display_name" : "AZ1",
    "_create_user" : "admin",
    <SNIP>
  } ],
  "result_count" : 3
```

Edge Node 1:

```
root@pks-client-vm:~/NSX-T/1-NSX-T-FD# curl -k -u admin:xxxxxxxxxxx -X GET 'https://pao-nsxmgr.blade.pvd.pez.pivotal.io/api/v1/transport-nodes/bd6fbb27-e7b9-412c-b487-aaf1e1934f04'
<SNIP>
"is_overridden" : false,
  "failure_domain_id" : "ce0caa6c-b57f-4766-bdf3-c30ef26b0a27",
<SNIP>
```

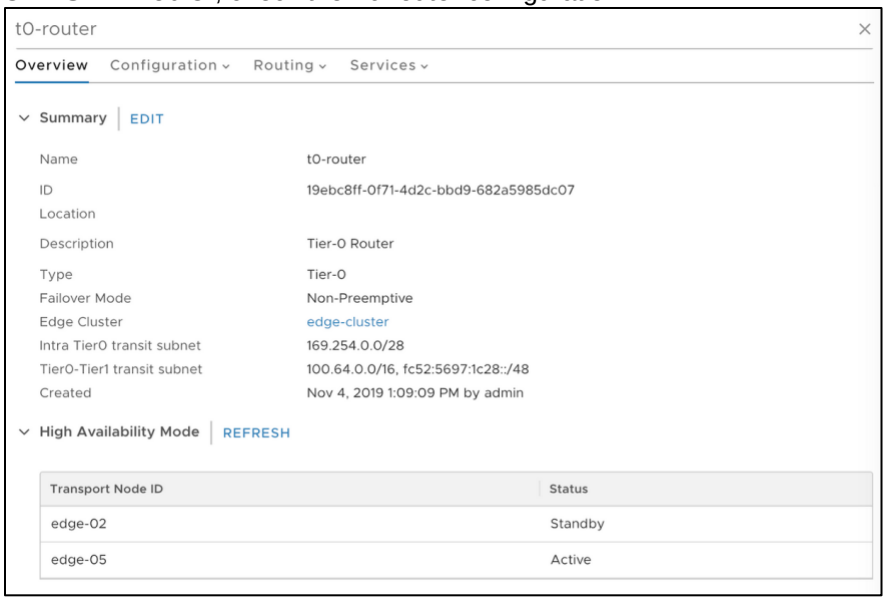
Edge Node 2:

```
root@pks-client-vm:~/NSX-T/1-NSX-T-FD# curl -k -u admin:dNfJ6b+NJ5nfbBM6RF8p -X GET 'https://pao-nsxmgr.blade.pvd.pez.pivotal.io/api/v1/transport-nodes/dd373e67-5b03-4e07-952d-7343d784e95d'
<SNIP>
 "is_overridden" : false,
  "failure_domain_id" : "1db43acb-a708-4ea3-8fb1-01a3bd98a9dc",
<SNIP>
```

© 2022 VMware, Inc. All rights reserved.

Make sure T0 is configured with 1 Edge Nodes in different FD:

On NSX-T web UI, check the T0 router configuration:



Active edge node is edge-05 which is in FD=AZ1
Standby edge node is edge-02 which is in FD=AZ2

© 2022 VMware, Inc. All rights reserved.

6. Production Considerations

You must take into consideration the following caveats when implementing a vSAN Stretched Cluster solution before going into production.

- **When Site 1 and Site 2 are Up and Running**:

  The BOSH resurrector function is disabled and will not restore failed or dysfunctional Kubernetes nodes:

  - Losing Kubernetes master nodes quorum will result in losing access to the Kubernetes cluster.
  - If a Kubernetes worker node becomes dysfunctional for any reason (file system full, issue with the vNIC, etc.), BOSH will not be able to repair them.

- **When Site 1 is Down:**
  - TKGI create cluster and TKGI resize cluster actions are not allowed because an AZ is down, and Bosh will try to create Kubernetes nodes in this AZ which will ultimately fail.
  - Kubernetes nodes in DC1 will be restarted in DC2 (thanks to vSphere HA):
    - However, Zone Labels on each node won't change (will still refer to original AZ where they were created).
    - 'bosh vm' command will display the original AZ where the node was created.
    - Note: when DC1 recovers, Kubernetes nodes will move back to their original location (thanks to vSphere DRS) and Zone Labels and 'bosh vm' command display will then be correct.
- **When Site 2 is down:**
  - TKGI create cluster and TKGI resize cluster actions are not allowed because an AZ is down, and Bosh will try to create Kubernetes nodes in this AZ which will ultimately fail.
  - Kubernetes nodes in DC2 will be restarted in DC1 (thanks to vSphere HA):
    - However, Zone Labels on each node won't change (will still refer to original AZ where they were created).
    - 'bosh vm' command will display the original AZ where the node was created.
    - Note: when DC2 recovers, Kubernetes nodes will move back to their original location and Zone Labels and 'bosh vm' command display will then be correct.

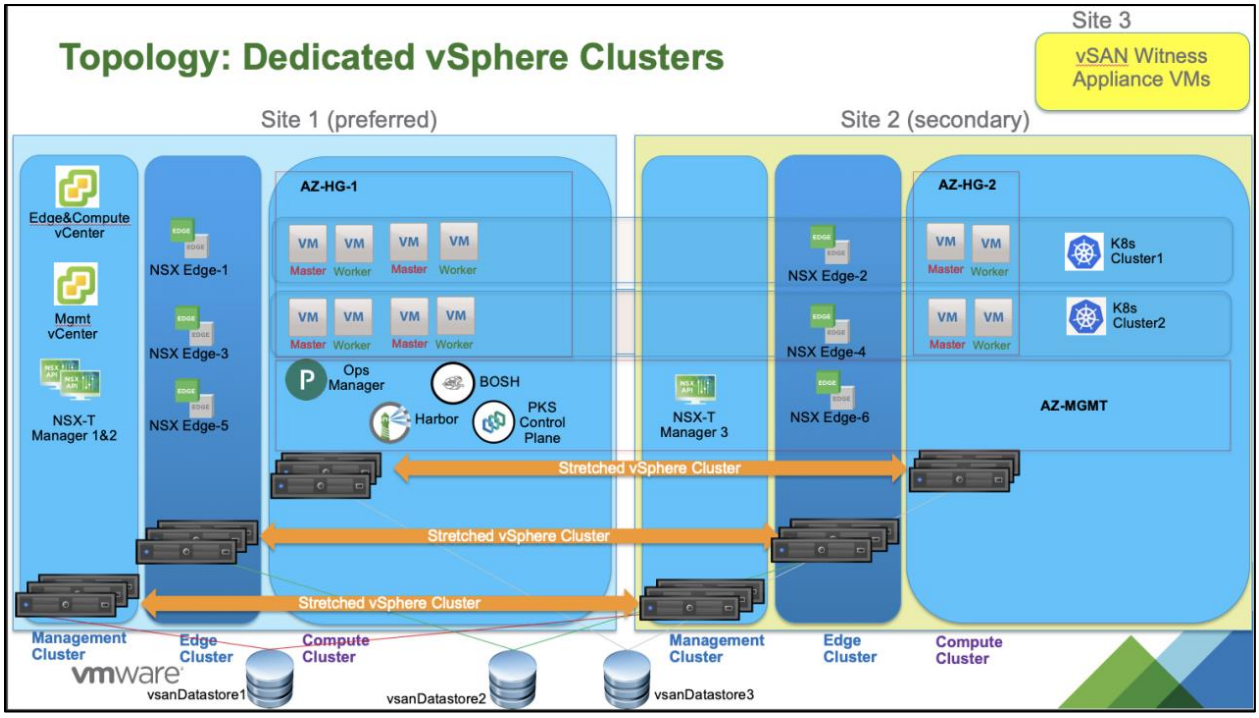
© 2022 VMware, Inc. All rights reserved.

7. Addendum: Testing Details

This addendum provides details on the testing done to support and document the solution.

7.1 Testbed Topology

The topology shown below depicts the environment and configuration that was used during the test and validation phase:



As a reference, this is the repartition of the ESXi hosts across the 2 sites for the 3 vSphere clusters:

vSphere cluster	Site1 (preferred) AZ-HG-1	Site2 (Secondary) AZ-HG-2
Management cluster	pao-esx-01.blade.pvd.pez.pivotal.io pao-esx-02.blade.pvd.pez.pivotal.io pao-esx-03.blade.pvd.pez.pivotal.io	pao-esx-04.blade.pvd.pez.pivotal.io pao-esx-05.blade.pvd.pez.pivotal.io pao-esx-06.blade.pvd.pez.pivotal.io
Edge cluster	pao-esx-11.blade.pvd.pez.pivotal.io pao-esx-12.blade.pvd.pez.pivotal.io pao-esx-13.blade.pvd.pez.pivotal.io	pao-esx-14.blade.pvd.pez.pivotal.io pao-esx-15.blade.pvd.pez.pivotal.io pao-esx-16.blade.pvd.pez.pivotal.io
Compute cluster	pao-esx-21.blade.pvd.pez.pivotal.io pao-esx-22.blade.pvd.pez.pivotal.io pao-esx-23.blade.pvd.pez.pivotal.io	pao-esx-24.blade.pvd.pez.pivotal.io pao-esx-25.blade.pvd.pez.pivotal.io pao-esx-26.blade.pvd.pez.pivotal.io

© 2022 VMware, Inc. All rights reserved.

7.2 Hardware Specifications

As a reference, the hardware specification of the ESXi hosts used in the lab is listed below:

VxRail Node Type E560
VxRail HCI System Software 4.7.300
CPU (x2) Intel(R) Xeon(R) Gold 5120 CPU @ 2.20GHz
VxRail NIC Intel(R) 10GbE 4P X710 rNDC 4 port
RAM (x12) per 32GB (x12) DDR-4 2400Mhz
Flash Storage (Cache) (x2) per Toshiba PX05SMB080Y 800GB
HDD (Capacity)(x4) per Seagate ST2000NX0423 2TB

This is just for information. Any other HW specifications certified by VMware will work for this topology.

7.3 Expected Behavior

We tested the overall system behavior under the following workflows to test the failure and recovery scenarios:

- Workflow 1: Site 2 down
- Workflow 2: Site 2 up
- Workflow3 : Site 1 down
- Workflow 4: Site 1 up

The goal was to determine the behavior of TKGI and the components (vCenter, NSX-T, vSAN, TKGI and Kubernetes clusters) in different site failure scenarios.

7.4 Testing Configuration

6 Kubernetes clusters were deployed.

Each Kubernetes cluster hosts stateless apps and stateful apps:

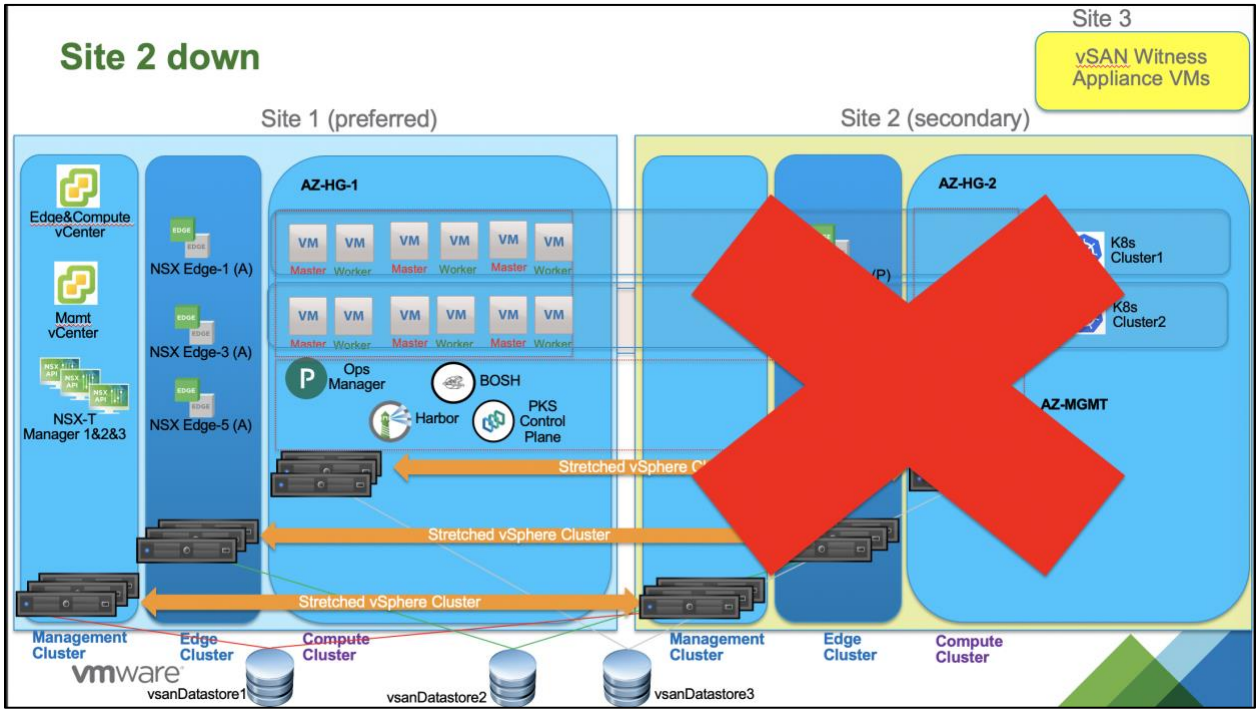
- Stateless apps: Guestbook and Tea-coffee
    - Guestbook leverages Kubernetes service of type LB
    - Tea-coffee leverages Kubernetes Ingress
- Stateful apps: Guestbook-pv and Cassandra DB
    - Guestbook-pv leverages PV using PVC
    - Cassandra DB leverages statefulset with 3 replicas

© 2022 VMware, Inc. All rights reserved.

7.5 Test Scenario 1: Site 2 down

Within 5 minutes, all Kubernetes nodes located in Site 2 are properly restarted by vSphere HA in site 1. Note that time will vary based on the load of the system in production.

Same behavior for NSX-T Manager 3 instance.



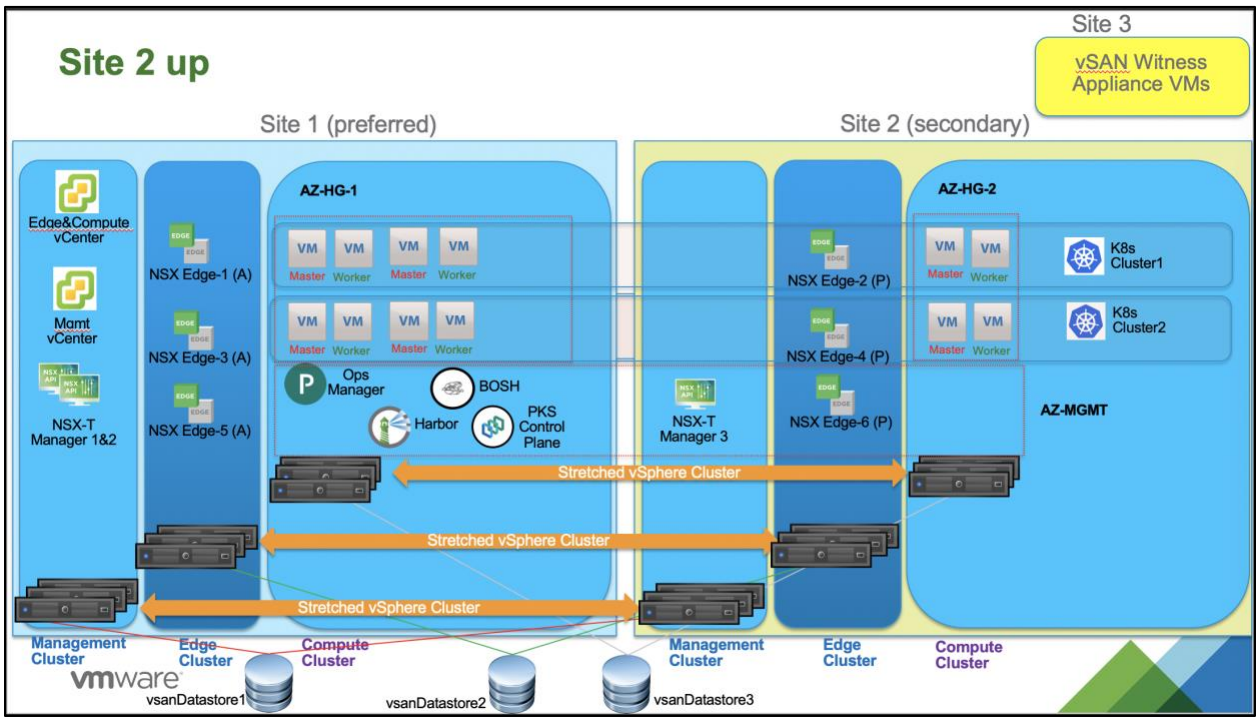
VC-01: vCenter-management	OK
VC-02: vCenter-compute	OK
NSX-T managers	OK
vSAN	OK
TKGI mgmt plane (Ops Mgr, Bosh, Harbor)	OK
Kubernetes clusters	OK
Stateless Apps	OK
Stateful Apps	OK

© 2022 VMware, Inc. All rights reserved.

Test Scenario 2: Site 2 up

Kubernetes nodes which were originally in Site 2 are placed back to that site with the help of vSphere DRS.

NSX-T manager instance 3 moves back to site 2 as well (because of the DRS rule that pins it to site 2).



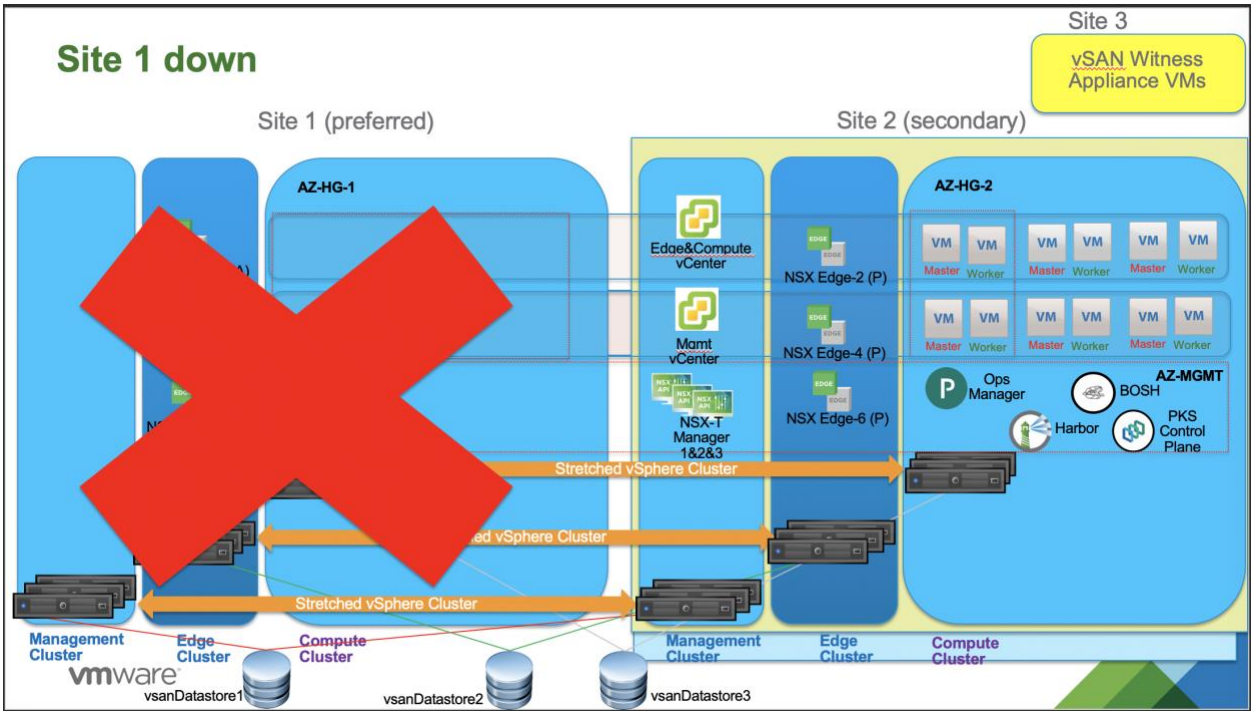
VC-01: vCenter-management	OK
VC-02: vCenter-compute	OK
NSX-T managers	OK
vSAN	OK
TKGI mgmt plane (Ops Mgr, Bosh, Harbor)	OK
Kubernetes clusters	OK
Stateless Apps	OK
Stateful Apps	OK

© 2022 VMware, Inc. All rights reserved.

7.6 Test Scenario 3: Site 1 down

Within 5-10 minutes, all Kubernetes nodes located in Site 1 are properly restarted by vSphere HA in site 2. Note that time will vary based on the load of the system in production.

Same behavior for both vCenter instances, NSX-T Manager 1 & 2 instances and TKGI management plane components.



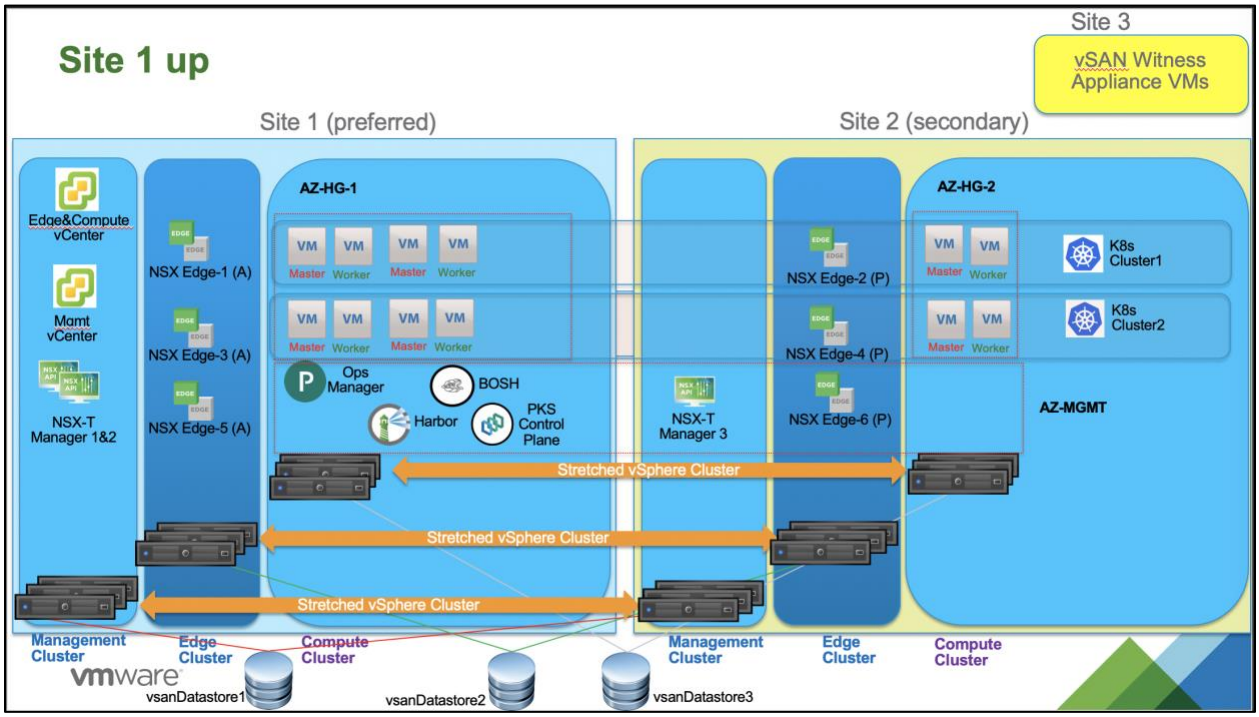
VC-01: vCenter-management	OK
VC-02: vCenter-compute	OK
NSX-T managers	OK
vSAN	OK
TKGI mgmt plane (Ops Mgr, Bosh, Harbor)	OK
Kubernetes clusters	OK
Stateless Apps	OK
Stateful Apps	OK

© 2022 VMware, Inc. All rights reserved.

7.7 Workflow 4: Site 1 up

Kubernetes nodes which were originally in Site 1 are placed back to that site with the help of vSphere DRS.

Both vCenter instances, NSX-T manager instance 1 & 2 and TKGI management plane VMs moves back to site 1 as well (because of the DRS rules that pin them to site 1).



VC-01: vCenter-management	OK
VC-02: vCenter-compute	OK
NSX-T managers	OK
vSAN	OK
TKGI mgmt plane (Ops Mgr, Bosh, Harbor)	OK
Kubernetes clusters	OK
Stateless Apps	OK
Stateful Apps	OK

© 2022 VMware, Inc. All rights reserved.