Operating Cisco HyperFlex HX Data Platform Stretch Clusters

Version 3.0 rev3

October 2023

Document information

Document summary	Prepared for	Prepared by
V3.0 rev3 Stretch Cluster information for HX 3.5, 4.0, 4.5, 5.0, 5.5	Cisco Field	Aaron Kapacinskas

Changes since 2.0 rev 13
Added information about the Arbitrator for HXDP 5.5
Added information on the operation of Preferred Sites
Added IS Arbitrator link to the More Information section
General document clean up

Intended use and audience

This document contains confidential material that is proprietary to Cisco. The materials, ideas, and concepts contained herein are to be used exclusively to assist in the configuration of Cisco® software solutions.

Legal notices

All information in this document is provided in confidence and shall not be published or disclosed, wholly or in part, to any other party without Cisco's written permission.

Contents

Prerequisites

We recommend reviewing the Cisco HyperFlex™ HX Data Platform release notes, installation guide, and user guide before proceeding with any configuration. The Data Platform should be installed and functioning as described in the installation guide. Please contact Cisco® Support or your Cisco representative if you need assistance.

Introduction

This document is intended to provide operational guidance to supplement the administration guidance for Cisco HyperFlex Stretch clusters. The goal is to help Cisco HyperFlex users understand the characteristics of a Stretch cluster and its day-2 operational features, including its resilience in response to various failure scenarios. Knowledge of the architecture and components of the solution is needed for a full understanding. To this end, the document begins with an overview of general Cisco HyperFlex components that apply to both regular and Stretch clusters.

This document provides recommended configuration settings and deployment architectures for Cisco HyperFlex HX Data Platform solutions specifically related to Stretch cluster deployments. It is intended to be used in conjunction with product documentation. For product documentation, please contact your Cisco representative.

Cisco HyperFlex HX Data Platform general overview: Components and environment

Cisco HyperFlex Stretch clusters and regular clusters are built on common architectural components, but with some slight differences for Stretch clusters related to Cisco Unified Computing System™ (Cisco UCS®) domains, installation processes, and failure modes. This section briefly examines the HX Data Platform components.

Cisco HyperFlex systems are designed with an end-to-end software-defined infrastructure that eliminates the compromises found in first-generation products. Cisco HyperFlex systems combine software-defined computing in the form of Cisco UCS servers, software-defined storage with the powerful Cisco HyperFlex HX Data Platform software, and software-defined networking (SDN) with Cisco unified fabric that integrates smoothly with the Cisco Application Centric Infrastructure (Cisco ACI™) solution. With hybrid or all-flash storage configurations, self-encrypting drive options, and a choice of management tools, Cisco HyperFlex systems deliver a pre-integrated cluster that is up and running in an hour or less. With the capability to integrate Cisco UCS servers as computing-only nodes, you can scale computing and storage resources independently to closely match your application needs.

The following several sections discuss the individual components of the solution, including Cisco UCS, fabric interconnects, and Cisco HyperFlex HX-Series nodes. These components are the same for both Stretch clusters and traditional clusters.

Cisco Unified Computing System

The physical HX-Series node is deployed on a Cisco UCS 220 or 240 platform in either a hybrid or all-flash configuration.

A service profile is a software definition of a server and its LAN and SAN connectivity. A service profile defines a single server and its storage and networking characteristics. Service profiles are stored in supported Cisco UCS 2nd, 3rd, and 4th generation Fabric Interconnects and are managed through specific versions of Cisco UCS Manager (the web interface for the fabric interconnect) or through purpose-written software using the API. When a service profile is deployed to a server, Cisco UCS Manager automatically configures the server, adapters, fabric extenders, and fabric interconnects to match the configuration specified in the service profile. This automation of device configuration reduces the number of manual steps required to configure servers, network interface cards (NICs), host bus adapters (HBAs), and LAN and SAN switches.

The service profile for the HX-Series nodes is created during the cluster build process during installation and is applied to the appropriate devices attached to the fabric interconnects (identified by part number and associated hardware). These profiles should have their own, easily identifiable names and should not be edited after creation. They are preconfigured by the Cisco HyperFlex installer with the settings required for the Cisco HyperFlex system to operate securely and efficiently (VLANs, MAC address pools, management IP addresses, quality-of-service [QoS] profiles, etc.).

Fabric interconnects

A Cisco UCS fabric interconnect is a networking switch or head unit to which the Cisco UCS chassis connects. The fabric interconnect is a core part of Cisco UCS. Cisco UCS is designed to improve scalability and reduce the total cost of ownership (TCO) of data centers by integrating all components into a single platform that acts as a single unit. Access to networks and storage is provided through the Cisco UCS fabric interconnect. Each HX-Series node is dual connected, with one Small Form-Factor Pluggable (SFP) port for each fabric interconnect for high availability. This design helps ensure that all virtual NICs (vNICs) within Cisco UCS are dual connected as well, essentially guaranteeing node availability. The vNIC configuration is automated during Cisco HyperFlex system installation and should not be altered.

Fabric interconnect traffic and architecture

Traffic through the fabric interconnect is of two general types: intracluster traffic (between nodes) and extracluster traffic (traffic related to client machines or replication). All fabric interconnect configurations are managed, accessed, and modified through Cisco UCS Manager.

Cisco UCS Manager requirements

Cisco UCS Manager is the interface used to set up the fabric interconnects for Cisco UCS service profiles and for general hardware management. During installation, the Cisco HyperFlex installer verifies that the appropriate Cisco UCS Manager build is in place for the Cisco HyperFlex system and that the hardware is running a supported firmware version. You are given the option to upgrade these versions during installation if you need to do so.

Cisco recommends disabling the serial-over-LAN (SoL) feature after the deployment is complete because it is no longer needed for VMware ESX configuration. You should also change any default or simple passwords that were used during the installation process.

Virtual network interface cards

For an in-depth discussion of vNICs, see the following:
https://supportforums.cisco.com/document/29931/what-concept-behind-vnic-and-vhba-ucs

The vNICs for each virtual switch (vSwitch) are in a predefined order and should not be altered in Cisco UCS Manager or ESX. Any changes to these (including active or standby status) could affect the functioning of the Cisco HyperFlex system.

East-west traffic

In a regular HX Data Platform cluster, east-west traffic on the fabric interconnect is networking traffic between HX-Series nodes. This traffic is local to the system and does not travel out of the fabric interconnect to the upstream switch. This traffic has the advantage of being extremely fast because of its low latency, low hop count, and high bandwidth. This traffic also is not subject to external inspection because it never leaves the local system unless there is a NIC failure on one of the nodes.

In a Stretch cluster, this traffic will need to traverse the site-to-site link between locations and so will exit a site’s individual fabric interconnects to the Stretch Layer 2 uplink switch and to the complementary site. It still occurs on the dedicated storage VLAN and remains secure. See the following major section “Stretch cluster architecture” to understand why this happens.

North-south traffic

North-South traffic on the fabric interconnect is networking traffic that goes outside the fabric interconnect to an upstream switch or router. North-south traffic occurs during external client machine access to Cisco HyperFlex hosted virtual machines or Cisco HyperFlex system access to external services (Network Time Protocol [NTP], vCenter, Simple Network Management Protocol [SNMP], etc.). This traffic may be subject to VLAN settings upstream. Because site-to-site Stretch cluster traffic needs to traverse the intersite link, a component of north-south traffic is a part of general storage traffic. For the purposes of this discussion, however, north-south generally refers to traffic coming into and out of the cluster (regular or stretch) for interactions between virtual machines and end users.

Upstream switches

Upstream or top-of-rack (ToR) switches are required to manage north-south traffic. You should configure the upstream switches to accommodate nonnative VLANs. The Cisco HyperFlex HX Data Platform installer sets the VLANs as nonnative by default. In a Stretch cluster, these are the switches that manage Layer 2 adjacency for each site.

VLANs

The solution uses several VLANs to separate traffic. It uses management VLANs for VMware ESXi and Cisco HyperFlex control virtual machines. It also uses VLANs for storage data traffic and for hypervisor data traffic (VMware vMotion traffic). You should use a separate subnet and VLANs for each network.

Do not use VLAN 1, the default VLAN, because doing so can cause networking problems, especially if a disjointed Layer 2 configuration is used. Use a different VLAN.

Disjointed Layer 2 networks

If a disjointed Layer 2 network is a requirement for your environment, be sure that you read and understand the following document: https://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/unified-computing/white_paper_c11-692008.html

You can simply add new vNICs for your use case. Cisco supports the manual addition of vNICs and virtual HBAs (vHBAs) to the configuration. Please see the Cisco HyperFlex virtual server infrastructure (VSI) Cisco Validated Design for step-by-step instructions about how to do this safely: https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/UCS_CVDs/HX171_VSI_ESXi6U2.html

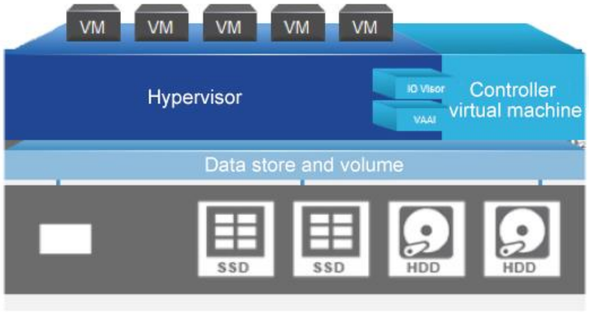
Follow the procedures outlined in the Cisco Validated Design. Do not use pin groups, because they may not properly prune the traffic and can cause connectivity problems because the designated receiver may not be set correctly.

Cisco HyperFlex HX-Series data node

The HX-Series node itself, whether part of a regular cluster or Stretch cluster, is composed of the software components required to create the storage infrastructure for the system's hypervisor. This infrastructure is created when the HX Data Platform is deployed during installation on the node. The HX Data Platform uses PCI pass-through, which removes storage (hardware) operations from the hypervisor, giving the system high performance. The HX-Series nodes use special plug-ins for VMware called VMware installation bundles (VIBs). These are used to redirect Network File System (NFS) data store traffic to the correct distributed resource and to offload to hardware complex operations such as snapshots and cloning.

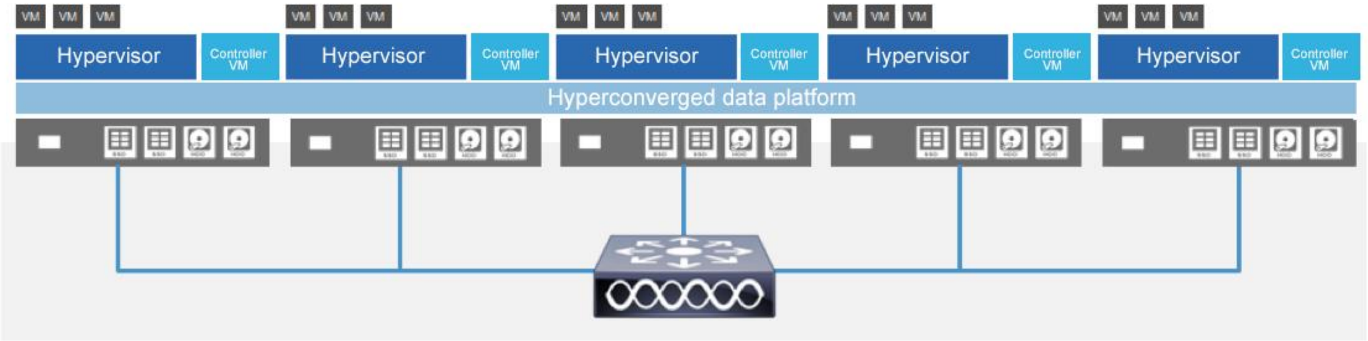
Figure 1 shows the typical HX-Series node architecture.

Figure 1. Typical Cisco Hyperflex HX-Series node



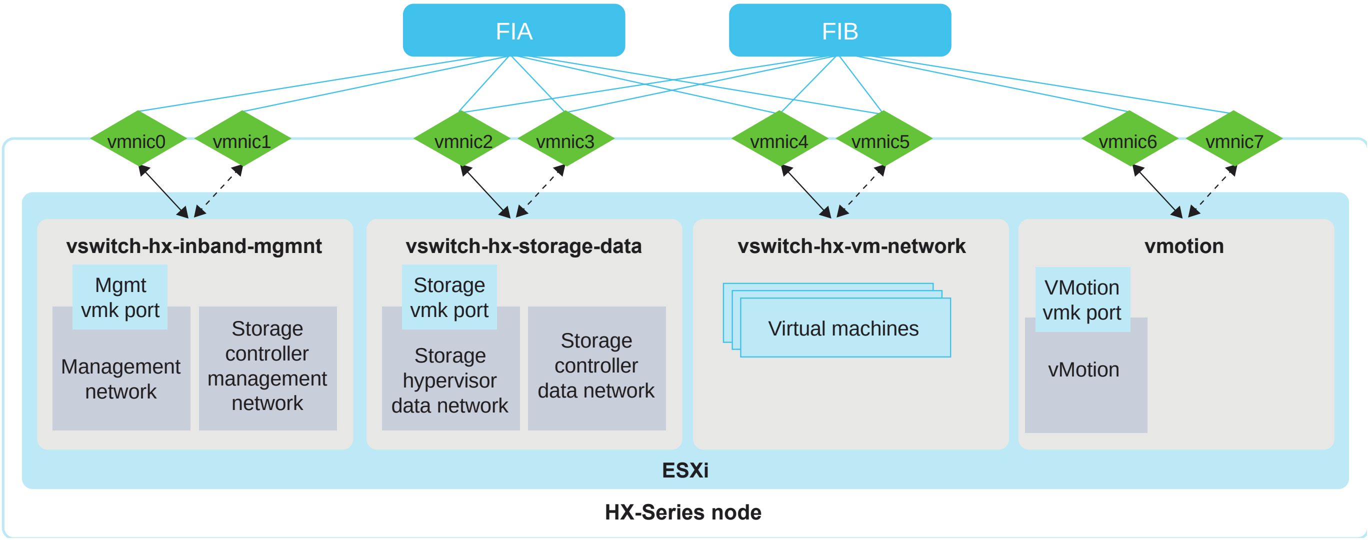
These nodes are incorporated into a distributed cluster using Apache ZooKeeper, as shown in Figure 2.

Figure 2. Cisco hyperflex distributed system



Each node consists of the virtual machine NIC (VMNIC) and vSwitch architecture shown in Figure 3.

Figure 3. Cisco Hyperflex HX-Series Node networking architecture



Management interfaces: Cisco HyperFlex Connect and VMware vCenter Plug-in

Cisco HyperFlex Connect is the native HTML 5.0 user interface for the cluster. The vCenter Plug-in for the Cisco HyperFlex system is another management interface available in vCenter after the cluster is deployed. In newer deployments the vCenter plugin is not installed by default. The HTML 5 plugin is installed separately after the cluster is built. HX Connect and the plugin are separate interfaces. Both are accessed through HTTPS in a web browser and are subject to the same user management (including role-based access control [RBAC]) that is available for the command-line interface (CLI) and the API

Apache ZooKeeper

ZooKeeper is essentially a centralized service for distributed systems for a hierarchical key-value store. It is used to provide a distributed configuration service, synchronization service, and naming registry for large distributed systems.

ZooKeeper's architecture supports high availability through redundant services. Clients can thus ask another ZooKeeper leader if the first fails to answer. ZooKeeper nodes store their data in a hierarchical name space, much like a file system or a tree data structure. Clients can read from and write to the nodes and in this way have a shared configuration service. ZooKeeper can be viewed as an atomic broadcast system through which updates are totally ordered.

ZooKeeper offers these main features:

- **Reliable system:** The system is very reliable because it keeps working even if a node fails.
- **Simple architecture:** The architecture of ZooKeeper is quite simple; it uses a shared hierarchical name space, which helps in coordinating processes.
- **Fast processing:** ZooKeeper is especially fast for read-dominant workloads.
- **Scalable:** The performance of ZooKeeper can be improved by adding nodes.

VMware vCenter

The Cisco HyperFlex HX Data Platform requires VMware vCenter to be deployed to manage certain aspects of cluster creation such as VMware ESX clustering for VMware High Availability (HA) and Distributed Resource Scheduler (DRS), virtual machine deployment, user authentication, and various data store operations. The vCenter Plug-in for the Cisco HyperFlex system is a management utility that integrates seamlessly within vCenter and allows comprehensive administration, management, and reporting for the cluster.

VMware ESX

ESX is the hypervisor component in the solution. It abstracts node computing and memory hardware for the guest virtual machines. The HX Data Platform integrates closely with ESX to facilitate network and storage virtualization.

Virtual machines

The Cisco HyperFlex environment provides storage for the guest virtual machines deployed in ESX using VLAN segmented networking. The virtual machines are available for external resources, as is typical of any elastic infrastructure deployment.

Client machines

Client machines are defined here as external hosts that need to access resources deployed in the Cisco HyperFlex system. These resources can be anything from end users to other servers in a distributed application architecture. These clients access the system from external networks and are always isolated from any Cisco HyperFlex internal traffic through network segmentation, firewalling, and whitelisting rules.

Cisco HyperFlex HX Stretch Clusters

This section provides an overview of Cisco HyperFlex Stretch clusters. It details some of the business reasons for deploying such a cluster. It also discusses some of the physical limitations of such a cluster.

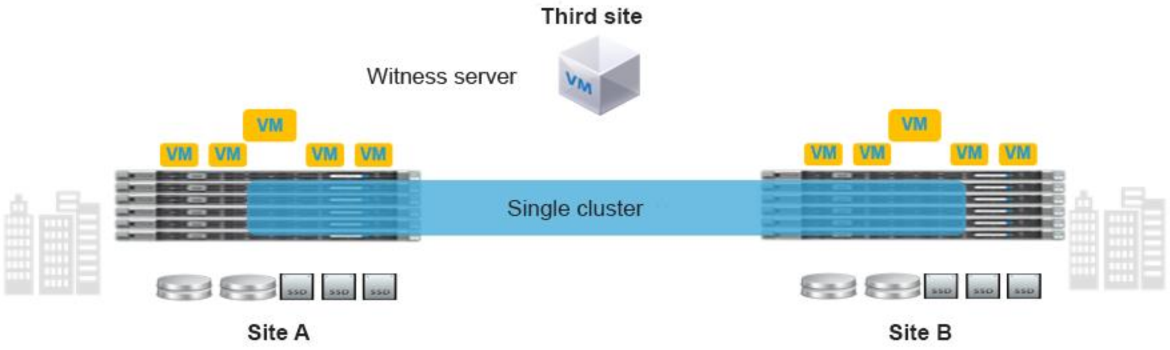
What is a Stretch Cluster

A Stretch cluster is distinct from a non-Stretch, or normal, cluster, in that it is designed to offer business continuance in the event of a significant disaster at a data center location. A Stretch cluster is geographically redundant, meaning that part of the cluster resides in one physical location and another part resides in a second location. The cluster also requires a “tie breaker” or “witness” component, which should reside in a third, separate location. The goal of this design is to help ensure that the virtual infrastructure remains available even in the event of the complete loss of one site. Of course, many lesser types of failures also can occur, and the system is highly available in the event of these as well. All of these scenarios are discussed later in this document.

People often mistakenly think that a Stretch cluster is a set of multiple single clusters. This is not the case. A Stretch cluster is, in fact, a single distributed entity and behaves as such in most circumstances. There are a few differences between a normal cluster and a Stretch cluster, however. These arise solely from the fact that a Stretch cluster must meet some special requirements to provide geographical redundancy for deployments that require it. Georedundancy introduces a few new requirements for the cluster so that certain conditions, such as split brain and node quorum, are handled properly. These are discussed in the following sections.

Figure 4 shows the main features of a Stretch cluster.

Figure 4. Three main components of a Stretch cluster deployment



Note the following characteristics of a Stretch cluster:

- A Stretch cluster is a single cluster with nodes geographically distributed at different locations.
- Storage is mirrored locally and across each site (but not to the tie-breaker witness).
- Sites need to be connected over a low-latency network to meet the write requirements for applications and for a good end-user experience.
- Geographic failover (virtual machine) is like failover in a regular cluster.
- Node failure in a site is like node failure in a regular cluster.
- Split brain is a condition in which nodes at either site cannot see each other. This condition can lead to problems if a node quorum cannot be determined (so that virtual machines know where to run). Split brain is caused by:
    - Network failure
    - Site failure
- Stretch clusters have a witness: an entity hosted on a third site that is responsible for deciding which site becomes primary after a split-brain condition.

Business need for a Stretch cluster

Businesses require planning and preparation to help ensure business continuity after serious incidents or disasters and to resume normal operations within a reasonably short period. Business continuity is the capability of an organization to maintain essential functions during, as well as after, a disaster. It includes three main elements:

- **Resilience:** Critical business functions and the supporting infrastructure must be designed so that that they are materially unaffected by relevant disruptions: for example, through the use of redundancy and spare capacity.
- **Recovery:** Organizations must have in place arrangements to recover or restore critical and less critical business functions that fail for some reason.
- **Contingency:** An organization must establish a generalized capability and readiness to allow it cope effectively with whatever major incidents and disasters may occur, including those that were not, and perhaps could not have been, foreseen. Contingency preparations constitute a last-resort response if resilience and recovery arrangements should prove inadequate in practice.

Stretch cluster physical limitations

Some applications, specifically databases, require a write latency of less than 20 milliseconds (ms). Many other applications require a latency of less than 10 ms to avoid problems with the application. To meet these requirements, the round-trip time (RTT) network latency on the Stretch link between sites in a Stretch cluster should be less than 5 ms. The speed of light (3e8 m/s) at the maximum recommended Stretch cluster site distance of 100 km (approximately 62 miles) introduces about 1 ms of latency by itself. In addition, time is needed for code path and link hops (from node to fabric interconnect to switch), which also plays a role in determining the maximum site-to-site recommended distance.

This section details the components in a typical Cisco HyperFlex deployment. Note that to achieve a secure environment, the various parts must be hardened as needed.

Solution components

A traditional Cisco HyperFlex single-cluster deployment consists of HX-Series nodes in Cisco UCS connected to each other and the upstream switch through a pair of fabric interconnects. A fabric interconnect pair may include one or more clusters. A Stretch cluster requires two independent Cisco UCS domains: one for each site. Therefore, a total of four fabric interconnects (two pairs) are required for a Stretch cluster. Other clusters can share the same fabric interconnects.

Figures 5 and 6 show typical physical layouts for this kind of deployment. Figure 5 shows a single site with its cabling and independent Cisco UCS domain. Figure 6 shows the racks for site A and site B in a Stretch cluster with their respective fabric interconnects and upstream switches. This is an 8-node (4+4) Stretch cluster with Cisco HyperFlex HX220c nodes at each location.

Figure 5. Site a for a Stretch cluster deployment showing a single-site rack: the site contains 4 HX220c M5 nodes and 2 fabric interconnects with a single uplink switch for the Stretch layer 2 network connecting to site b

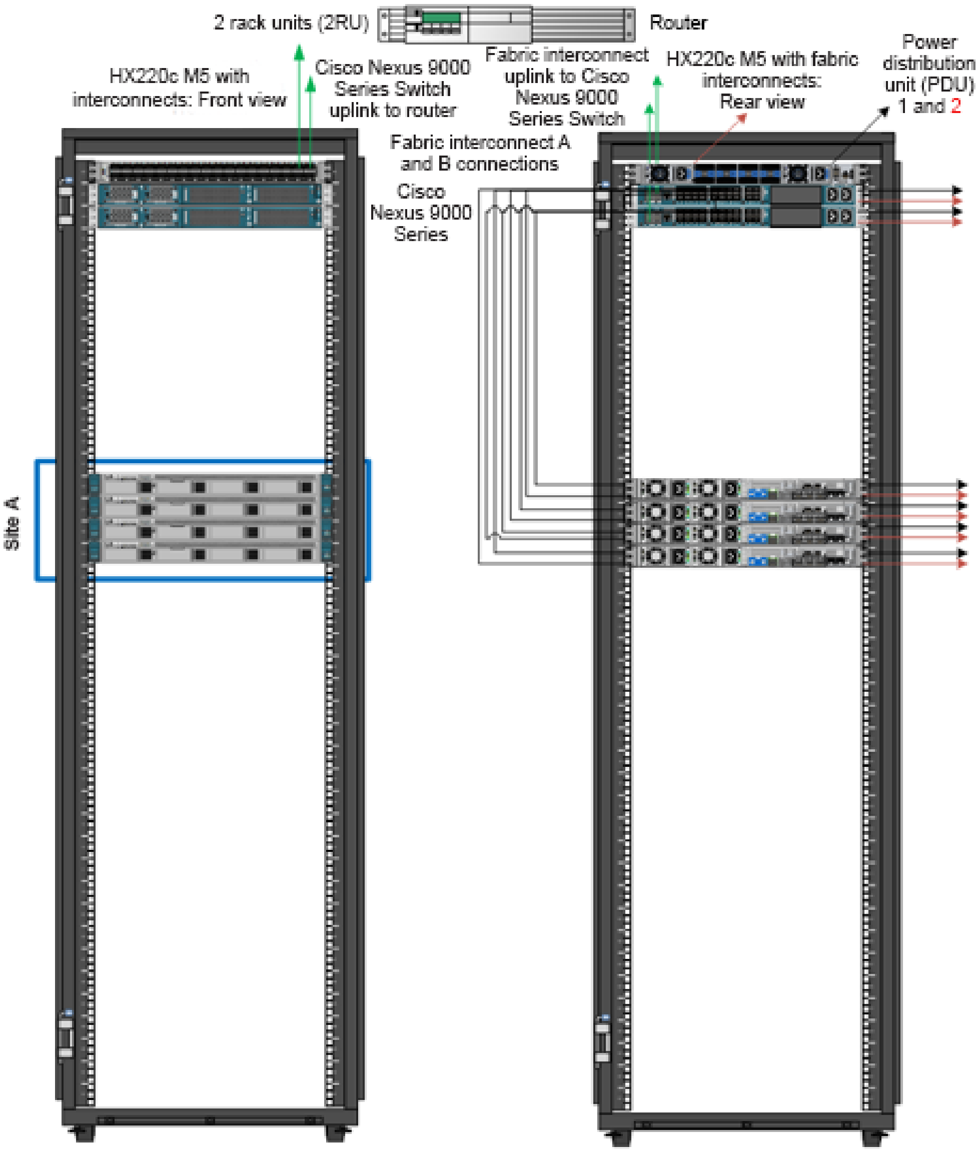
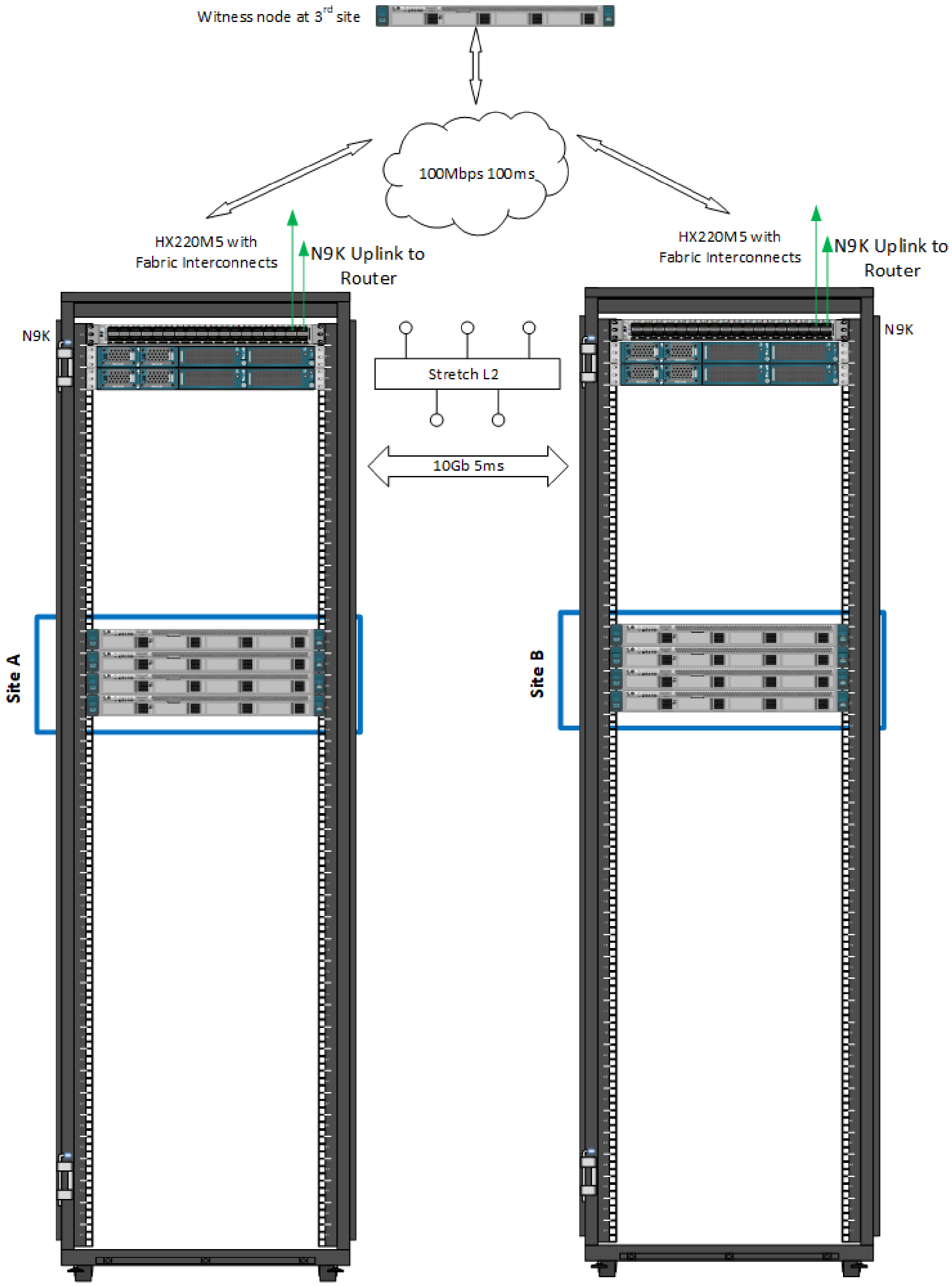

2 rack units (2RU)
Router
HX220c M5 with interconnects: Front view
Cisco Nexus 9000 Series Switch uplink to router
Fabric interconnect uplink to Cisco Nexus 9000 Series Switch
HX220c M5 with fabric interconnects: Rear view
Power distribution unit (PDU) 1 and 2
Fabric interconnect A and B connections
Cisco Nexus 9000 Series
Site A

Figure 6. Rack diagram showing site a and site b with their respective fabric interconnects and a logical third site at another location for the Stretch cluster witness

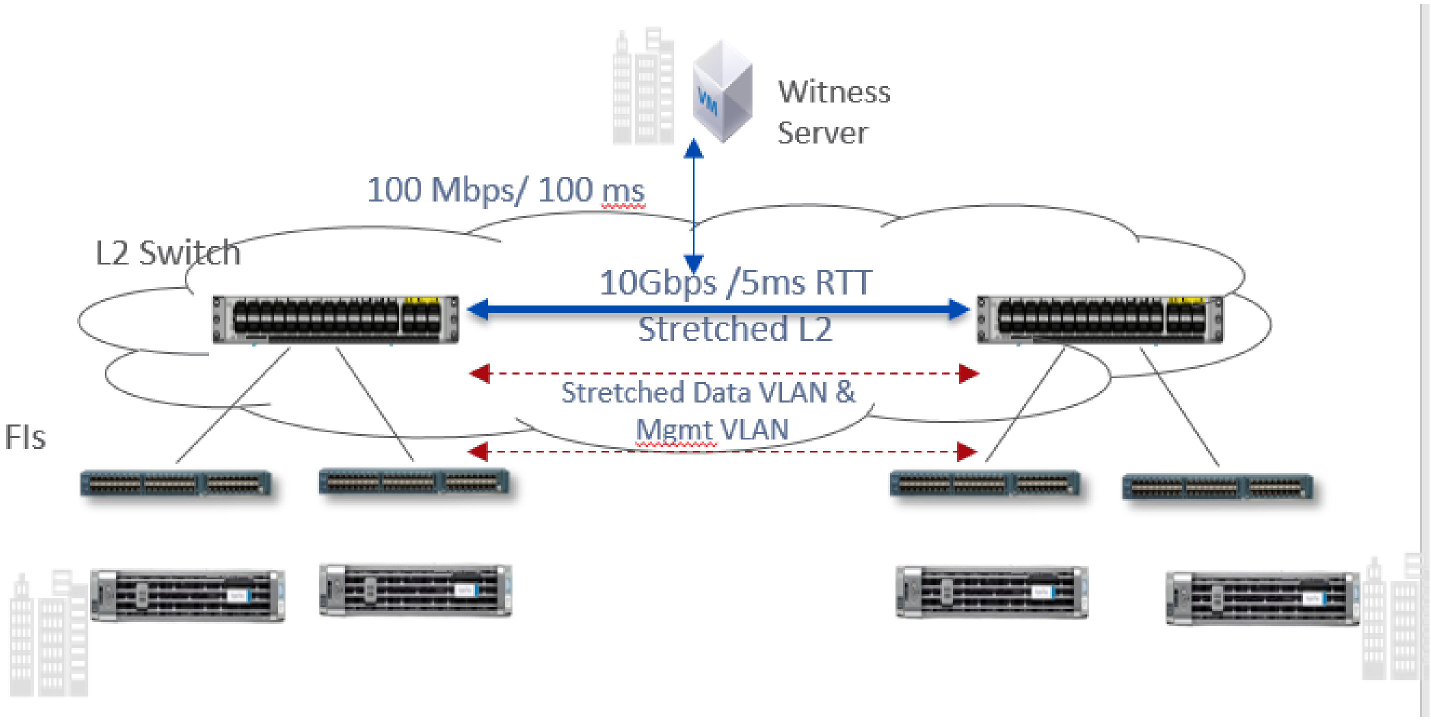


Stretch cluster architecture

This section discusses the specific deployment needs for a Stretch cluster, including hardware, networking configuration, VMware requirements (ESXi and vCenter), failure sizing, and characteristics of the witness (Figure 7). VMware vSphere Enterprise Plus is required because Cisco HyperFlex

Stretch clusters rely on advanced DRS capabilities available only in that premium edition. The requirements are the same across all stacks (even for non-hyperconverged infrastructure [HCI] or traditional storage) that implement Stretch or metropolitan clusters on VMware.

Figure 7. General Stretch cluster network



The first consideration in deploying a Stretch cluster is building the proper site-to-site network. A Stretch cluster requires a minimum of 10 Gigabit Ethernet connectivity and 5-ms RTT latency on the link. The link needs to be Stretch Layer 2 to help ensure network space adjacency for the data storage VLAN network that is used for storage communication. The network between sites requires the following characteristics:

- 10 Gbps (dedicated) for the storage data VLAN
- 5-ms RTT latency between the two active sites
- Data VLAN and management VLAN on a Stretch Layer 2 VLAN
- Stretch Layer 2 VLAN between the two sites
    - Dark fiber and dense wavelength-division multiplexing (DWDM) Layer 2 and 3 technologies are supported.
    - The solution is not currently qualified for Virtual Extensible LAN (VXLAN) unless used with ACI.
    - Stretch Layer 2 characteristics
    - The Stretch data VLAN should use jumbo maximum transmission units (MTUs) for best performance. The installer allows for deployment using an MTU of 1500, however.
    - The Cisco Nexus® 5000 Series Switches are slightly different than the Cisco Nexus 7000 and 9000 Series Switches. The default network-QoS policy does not accept jumbo MTUs, but you can set up jumbo switch policy across the switches.
    - Test the RTT ping using **VMkping –I VMk1 -d -s 8972 x.x.x.x** from any ESXi host in your cluster. This check is also performed by the installer, and if it fails, the installation process will not proceed.
- 100 Mbps and 100-ms RTT latency between the active sites and the witness site

- Different drive types are supported with different nodes limits. See the release notes for your running or target version to determine which drives and nodes you can use. For example, there are LFF drive restrictions and NVME drives began support in 4.0.2x and onward for the HX220 node type.

Support and Limitations

Some deployment limitations exist for Stretch clusters related to the qualified hardware. Most of these limitations are not based on technical factors but simply reflect test bandwidth and the release cycle. After these items have been qualified, they will be removed from the unsupported-features list, and these capabilities will be available for general deployment. For example, ***AMD processors are now supported with Stretch cluster deployments.***

Check the minor version release notes periodically for changes in the support listings.

Minimum and maximum configuration limitations are as follows:

- Minimum
  - Two fabric interconnects per site
  - Two nodes per site
  - One witness
  - One vCenter instance
  - Replication factor: 2+2
- Maximum
  - Two fabric interconnects per site
  - 2:1 maximum ratio for compute to converged nodes
  - Compute nodes can be added asymmetrically with no restriction
  - 16 small-form-factor (SFF) converged nodes per site (32 total, max cluster 64 with compute)
  - 8 large-form-factor (LFF) converged nodes per site (16 total, max cluster 48 with compute)
  - One witness
  - One vCenter or vCenter with HA instance if there is no database update lag
  - Replication factor: 2+2

Stretch cluster support limitations are as follows:

- Self-encrypting drives (SEDs) are not supported.
- Compute-only nodes are supported in HyperFlex 3.5 or higher with a 2:1 ratio to converged nodes. Verify the ratio in the Release Notes for your version.
- ESXi is the only supported hypervisor at this time. Check the release notes for your HX version to see the recommended ESXi version.
- Cisco HyperFlex native replication is supported in HyperFlex 3.5 and greater.
- Cisco HyperFlex N:1 native replication using Edge in HX 4.5.1a with Stretch as the target is not supported
- Expansion of an existing cluster to a Stretch cluster is not supported.
- Stretch clusters are supported only in fresh installations. Upgrade from a standalone cluster to a Stretch cluster configuration is not supported.
- Stretch Clusters must be symmetric (converged nodes). For production environments, this includes Fabric Interconnects.
- Stretch Clusters must be expanded symmetrically (converged nodes). See the admin guide for your version of HX for workflow details.

- Stretch Clusters can be built and/or expanded asymmetrically with compute nodes.
- Online rolling upgrades are supported only for the HX Data Platform. Cisco UCS Manager upgrades must be performed manually one node at a time.
- Stretch clusters are supported on Cisco M5 nodes only. M4 nodes are not supported.
- Logical availability zones are not currently supported in Stretch clusters.
- The witness requires ESXi at the third site (cloud deployment is not currently supported) or use of the Arbitrator
- Disk reshuffling is not supported (e.g., adding empty nodes and "leveling" the disks out)
- Hardware offload (acceleration) cards are supported starting in HXDP version 4.0.2b and greater
- Node removal is not supported
- Single Socket nodes may or may not be supported, depending on your version of HX. Please see the Release Notes.
- **Oracle RAC is not supported**

HyperFlex Stretch Cluster with Oracle

There are some caveats around various Oracle types running on a Stretch Cluster:

- Oracle Single Instance can run on a Stretch Cluster
- Oracle RAC is not supported
- Licensing Oracle for a cluster depends on VM core count. You can reference this link for more information. https://www.flexera.com/blog/it-asset-management/oracle-database-licensing-in-a-vmware-virtual-environment-part-1-of-3-2/

HyperFlex Native Software-Based Encryption (SWE)

Starting with HX 5.0.2(a), Stretch clusters support HX software-based encryption (SWE). This requires the cluster to be claimed in Intersight. The Key Management Server (KMS) for SWE is cloud based and Intersight resident. Please reference the Hardening Guide for details on this encryption mechanism.

About Zones

While logical availability zones are not currently supported in Stretch cluster deployments, you may notice that zone information is available when running the stcli cluster get-zone command as show below:

```
root@SpringpathControllerOHCWUK9X3N:~# stcli cluster get-zone
zones:
    ----------------------------------------
    pNodes:
        ----------------------------------------
        state: ready
        name: 192.168.53.136
        ----------------------------------------
        state: ready
        name: 192.168.53.135
        ----------------------------------------
    zoneId: 51733a6b98df9784:4f8fc27070894bf4
    numNodes: 2
```

```
       ----------------------------------------
       pNodes:
           ----------------------------------------
           state: ready
           name: 192.168.53.138
           ----------------------------------------
           state: ready
           name: 192.168.53.137
           ----------------------------------------
       zoneId: 7b04a6600e3e3ee5:54c7224773a14a9a
       numNodes: 2
       ----------------------------------------
   isClusterZoneCompliant: True
   zoneType: physical
   isZoneEnabled: True
numZones: 2
```

LAZ and Stretch cluster both are implemented using a basic feature called "zones" and that's why you see 'zone' in some of the output. You will not see "logical zones" which is what would appear under LAZ.

note the “zoneType” on the get-zone output.
On Stretch cluster: “zoneType: physical”
On Cluster with LAZ : “zoneType: logical”

Hardware Matching

Stretch Clusters require identical hardware at both sites. This includes node count, type, and drives per node as well. This also applies to expansion. You must expand in converged node pairs.

There are some exceptions to the hardware symmetry requirement. Compute resources are not required to be symmetric between sites. You can have more compute-only nodes on one site than the other. However, care should be taken since a failure scenario from one site with large compute resources to another site with reduced resources may not be sized properly to run the VMs that are started on the surviving site.

Mixing CPU generations is supported within the same family as well. For example, it is ok to mix 8180 Skylake CPUs with 6258R Cascade Lake CPUs. You must, however, size for the less powerful CPU.

A Stretch Cluster will work, i.e., deploy properly and functional as expected, if the FIs are different between sites, but identical within the site. This can be useful for lab and testing environments but is not supported by Cisco for production. FIs must be identical within a site and between sites for production.

Although in general drives need to be identical, there are situations where this is not possible. For example, when existing hardware goes EOS (End of Sale) or EOL (End of Life) a suitable replacement is needed. Here is a list of compatible drives for hardware that has reached EOS or EOL and is no longer available. These drives can be substituted for each other if they are on the compatibility list together. If you are conducting a new install or reinstall with these mixed but compatible drives, the installer will not block the deployment.

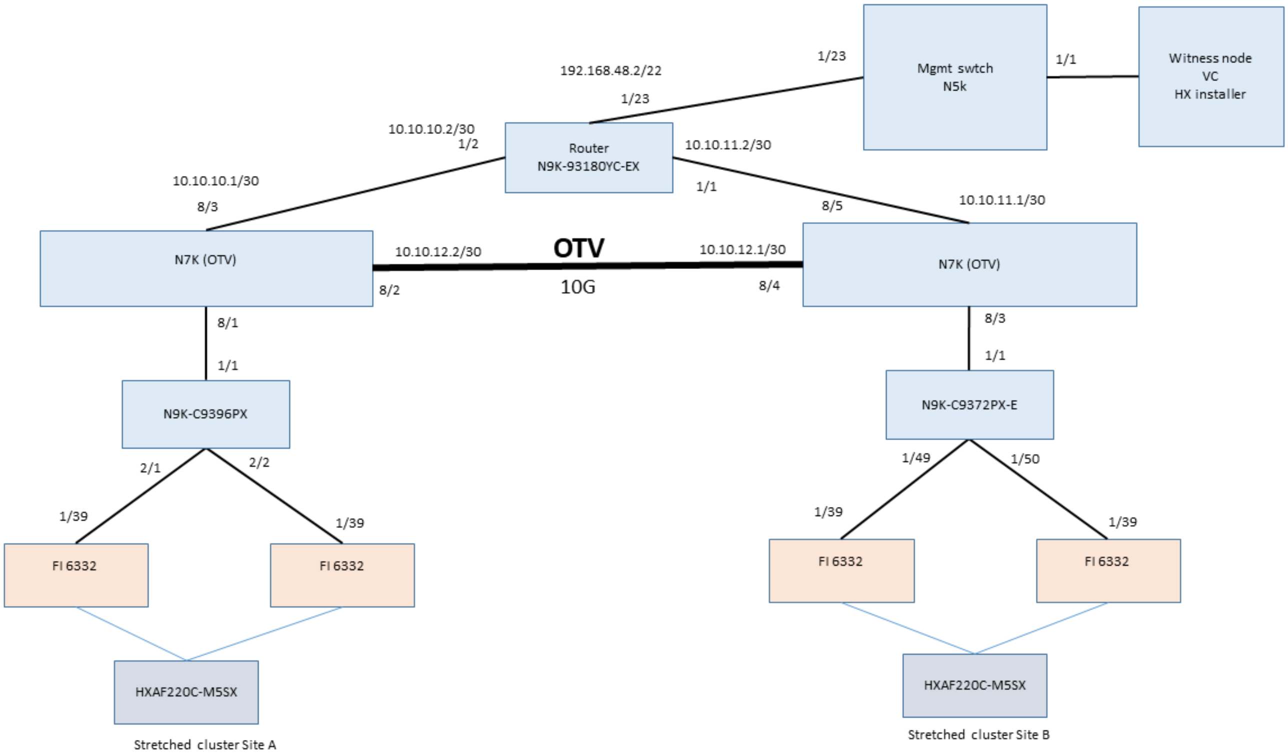
M6 nodes are supported on Stretch Cluster. These nodes can be used exclusively or mixed with M5 nodes via expansion only. In short, if you want to have an M5 and M6 mixed cluster, the cluster needs to be all M5 first and then expanded symmetrically with M6 nodes.

Overlay Networks

Only certain overlay networks are currently qualified for use with Stretch Clusters. OTV is qualified for use with Stretch Cluster. VXLAN is qualified as a standalone protocol or in conjunction with ACI, for use with Stretch Cluster. NSX is now supported as a stand-alone overlay protocol.

See the “More Information” section for the CVDs describing the various VXLAN deployments. VXLAN has some specific requirements (e.g., Nexus 7k/9k only) that are detailed in the design documents presented there.

Cisco Overlay Transport Virtualization (OTV), supported on Nexus Switches, is a networking technology that allows relaying layer 2 (L2) networks over layer 3 (L3) network segments. OTV is important for Stretch Clusters that require Stretch L2 storage and management networks when a dedicated dark fiber type site-to-site connection is not available. The tested and validated OTV design is shown below.



This OTV design was tested for the various failure modes discussed later. It was configured to meet the bandwidth and latency requirements necessary for the proper operation of the Stretch Cluster. It is important to note that layering over L3 can introduce latency since the routed network will necessarily have additional device to device hops. When designing and deploying this type of architecture you must ensure that you are still within the site-to-site communication specification for bandwidth and latency.

The following references are for OTV on Nexus:

https://www.cisco.com/c/en/us/solutions/data-center-virtualization/overlay-transport-virtualization-otv/index.html

https://community.cisco.com/t5/data-center-documents/understanding-overlay-transport-virtualization-otv/ta-p/3151502

Fabric interconnects

Stretch clusters have a specific set of fabric interconnect requirements. Each site is built using its own pair of fabric interconnects in an independent Cisco UCS domain. Therefore, a total of four fabric interconnects are required. The Stretch cluster requires a symmetric deployment, meaning that each site must have the same number and type of fabric interconnects and converged nodes. If site A has 4 hybrid nodes, then site B must also have 4 hybrid nodes. As of Cisco HyperFlex 3.0, the maximum cluster size is 8 nodes per site, for a total of 16 (8 + 8). This has increased in 3.5 and above to 16 converged nodes per site (SFF) with up to a 2:1 compute node ratio for a maximum mixed count of 32 per site. Limits for LFF drives are different. See the release notes for your version of HX to get the latest information on the number and type of supported nodes.

Fabric interconnect and node configuration details are summarized here:

- A total of four fabric interconnects are required, one pair at each site) in unique Cisco UCS domains.
- Do not mix fabric interconnect models within a domain.
- For the fabric interconnects, check the release notes for the recommended UCSM version for your HX version
- Existing fabric interconnects are supported as long as they work with Cisco M5/M6 nodes.
- Node requirements are as follows:
  - You must have the same number and type of nodes per site: All flash or all hybrid.
  - The maximum cluster size and node type depends on your HX version. Check the release notes.
  - These requirements and maximums change frequently, consult the Release Notes for your version.

Note that the UCSM domains present in sites A and B are independent. They do not communicate with each other at the UCSM level. This means that the FI VIP and the FI IP addresses can be in different subnets for the respective domains. It is a good idea to keep the domains either in the same subnet or at least routable to each other in the event this behavior changes in a future release.

Fabric Interconnects Uplink Best Practices

Care should be taken with all deployments of HX when uplinking the Fabric Interconnects to your TOR/edge switches. The best practice surrounding this is designed to make sure that Spanning Tree Protocol (STP) loops are avoided. In a normal cluster these loops will cause FI takeover problems. Due to the multi-domain nature of a Stretch cluster, STP storms can bring the system down. When uplinking the FIs to your redundant swtiches, the virtual port channel (VPC) ports should be set to edge trunk mode so that they do not participate in STP.

This behavior is called out in several locations within Cisco documentation but is reiterated here for convenience. For example, the following document call out using spanning-tree port type edge trunk or the need to disable spanning tree on ports connecting to the FIs from upstream switches:

- https://www.cisco.com/c/en/us/td/docs/hyperconverged_systems/HyperFlex_HX_DataPlatformSoftware/Network_External_Storage_Management_Guide/b_HyperFlex_Systems_Network_and_External_Storage_Management_Guide_3_0/b_HyperFlex_Systems_Network_and_External_Storage_Management_Guide_3_0_chapter_01.html

Cisco FIs appear on the network as a collection of endpoints versus another network switch. Internally, the FIs do not participate in spanning-tree protocol (STP) domains, and the FIs cannot form a network loop, as they are not connected to each other with a layer 2 Ethernet link. The upstream root bridges make all link up/down decisions through STP

.

Uplinks need to be connected and active from both FIs. For redundancy, you can use multiple uplinks on each FI, either as 802.3ad Link Aggregation Control Protocol (LACP) port-channels or using individual links. For the best level of performance and redundancy, make uplinks LACP port-channels to multiple upstream Cisco switches using the virtual port channel (vPC) feature. Using vPC uplinks allows all uplinks to be active passing data. This also protects against any individual link failure and the failure of an upstream switch. Other uplink configurations can be redundant, but spanning-tree protocol loop avoidance may disable links if vPC is unavailable.

When setting the uplinks from the FI as VPC port channels you also need to set the downlink ports on the (e.g.) Nexus 9k to “spanning tree edge” instead of “spanning tree normal”, since the FIs don’t participate in STP. In the absence of this configuration, a spanning tree storm in the N9k will cause a traffic blackhole for HX storage traffic. This in turn will affect all HX traffic in a Stretch cluster. In standard clusters, the problem happens only when there is an FI failover.

In clusters without the ability to use vPC or LACP based link aggregation for redundancy, you should use disjoint layer 2.

VMware vCenter

vCenter is a critical component for normal clusters and is vital for a Stretch cluster. vCenter, with HA and DRS configured automatically manages virtual machine movement in the event of a site failure. The use of virtual machine host groups in the preferred mode, in which virtual machines are pinned to a site for the purpose of local computing and read I/O, is required for optimal performance in a Stretch deployment. Site host groups and the corresponding affinities are created automatically at build time by the Cisco HyperFlex installer.

Data stores also maintain site affinity using host groups as the mechanism to locate the primary copy of virtual machine data. This approach is used to facilitate the asymmetric I/O mechanism that a Stretch cluster uses to increase the cluster response time by localizing read I/O while distributing write I/O (two local-site copies and two remote-site copies). Because both sites in a Stretch cluster are active, virtual machines at one site or the other do not suffer any “second-class citizen” type scenarios, in which one site has preferential performance relative to another.

In a Stretch cluster deployment, a single instance of vCenter is used for both sites. The best approach is to locate this instance at a third location so that it is not affected by site loss. Co-residency with the witness is often the preferred choice because the witness site is required anyway. Nested vCenter (i.e., running the cluster’s vCenter instance on the cluster itself) is not supported. vCenter HA (VCHA) is supported with Stretch Cluster. Be aware the VCHA is a high availability deployment of vCenter itself and does not refer to the enabling HA on vCenter (which is a separate requirement for proper cluster failover behavior).

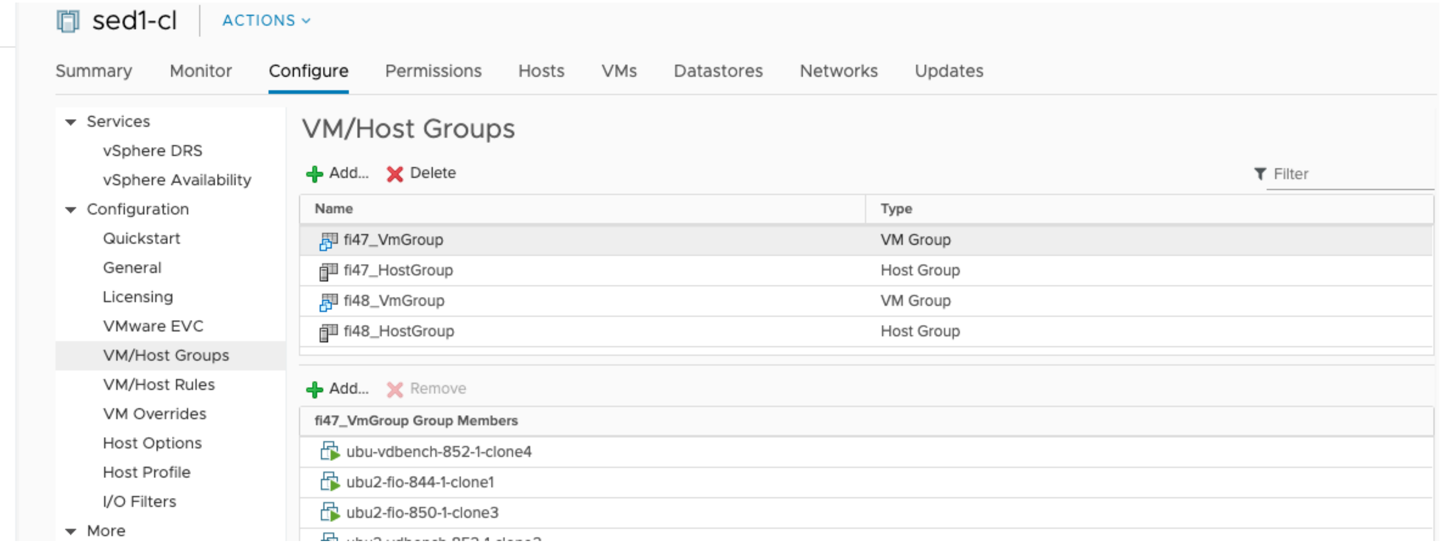
In the vCenter instance, the Stretch cluster corresponds to a single ESXi cluster. Be sure to verify that HA and DRS are set up for the Stretch cluster.

If the need arises to move the cluster from one vCenter to a new vCenter deployment or a different existing vCenter instance, it will be necessary to perform a cluster re-register. Be sure to see the admin guide for detailed notes, but the general workflow is as follows: Create the cluster object in the new vCenter instance and add the cluster ESXi hosts manually. Be sure the HA/DRS is enabled. The re-register is conducted using STCLI from any node or the CIP-M address.

```
admin@ControllerE2L5LYS7JZ:~$ stcli cluster reregister
usage: stcli cluster reregister [-h] --vcenter-datacenter NEWDATACENTER
                                --vcenter-cluster NEWVCENTERCLUSTER
                                --vcenter-url NEWVCENTERURL
                                [--vcenter-sso-url NEWVCENTERSSOURL]
                                --vcenter-user NEWVCENTERUSER
stcli cluster reregister: error: argument --vcenter-datacenter is required
```

In a non-Stretched cluster this is all that is required to remove the cluster from one vCenter instance and move it to a new one. A Stretch cluster, however, requires a few manual steps to complete the process. This is because Host Groups and Affinity Rules are not transferred in the re-registration process. Please note that ICPM needs to be accessible between hosts and vCenter for re-registration to function properly.

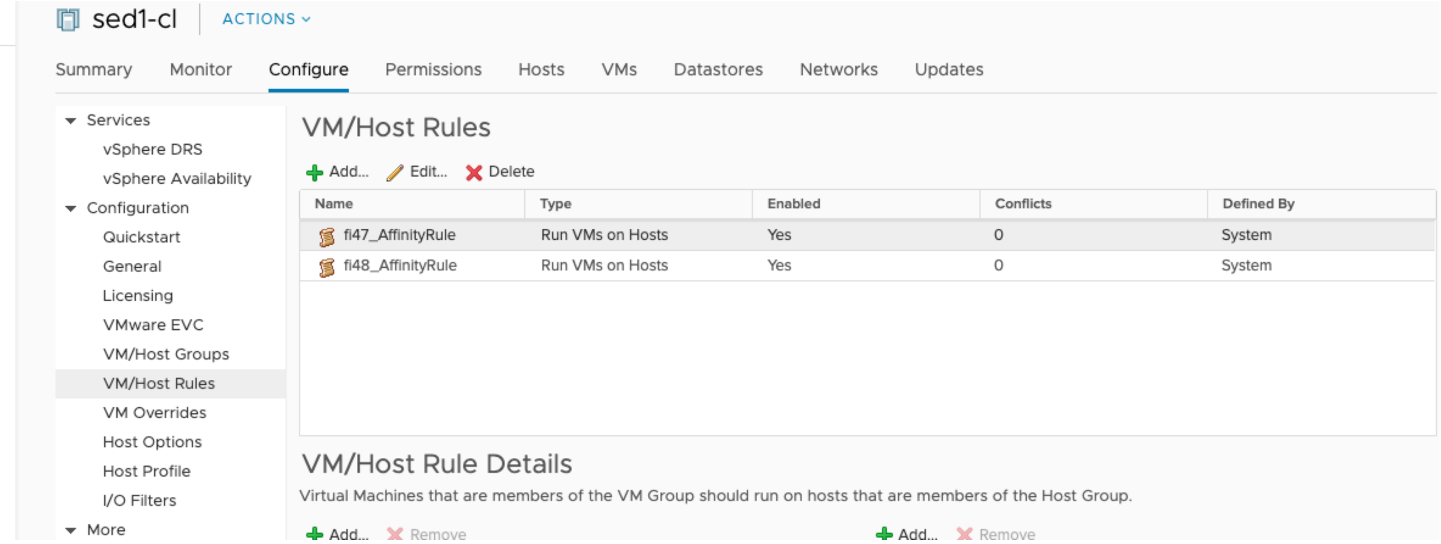
A Stretch cluster relies on a specific naming convention when interfacing with vCenter for implementation of the affinity rules. This is set up automatically, in advance, with the HX Installer when the cluster sites are built. The host group and affinity group naming must follow this convention: <site name>_{HostGroup, VmGroup, SiteAffinityRule} when rebuilding the groups and rules on the new vCenter host. See the screens below for an example. Here, site 1 is called fi47 and site 2 is fi48. Note the naming convention.



VM/Host Groups

Name	Type
fi47_VmGroup	VM Group
fi47_HostGroup	Host Group
fi48_VmGroup	VM Group
fi48_HostGroup	Host Group

fi47_VmGroup Group Members
ubu-vdbench-852-1-clone4
ubu2-fio-844-1-clone1
ubu2-fio-850-1-clone3



VM/Host Rules

Name	Type	Enabled	Conflicts	Defined By
fi47_AffinityRule	Run VMs on Hosts	Yes	0	System
fi48_AffinityRule	Run VMs on Hosts	Yes	0	System

VM/Host Rule Details

Virtual Machines that are members of the VM Group should run on hosts that are members of the Host Group.

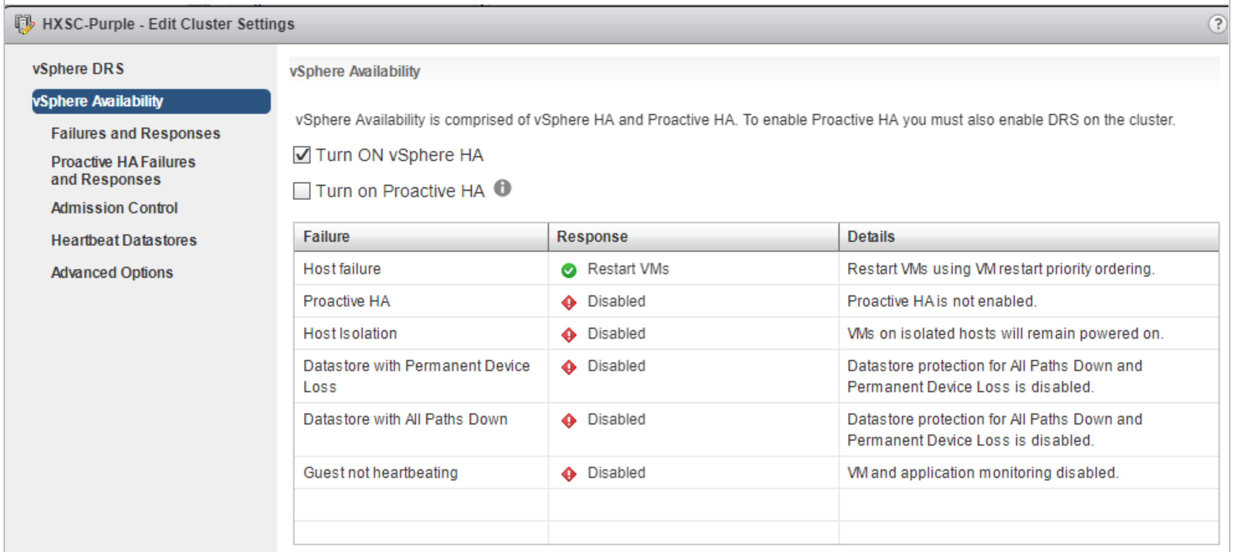
VMware vCenter HA Settings

The settings below are recommended for use in HX Stretch Clusters. This table details the most common settings in vSphere HA that are typically asked about during custom configuration. The screens shots are representative of vCenter 6.5. The cluster will work as designed using the default installation values. If you do not see a value listed below, keep it at the default.

vSphere HA Settings

vSphere HA

Turn on HA. Keep Proactive HA disabled.



HXSC-Purple - Edit Cluster Settings

vSphere Availability

vSphere Availability is comprised of vSphere HA and Proactive HA. To enable Proactive HA you must also enable DRS on the cluster.

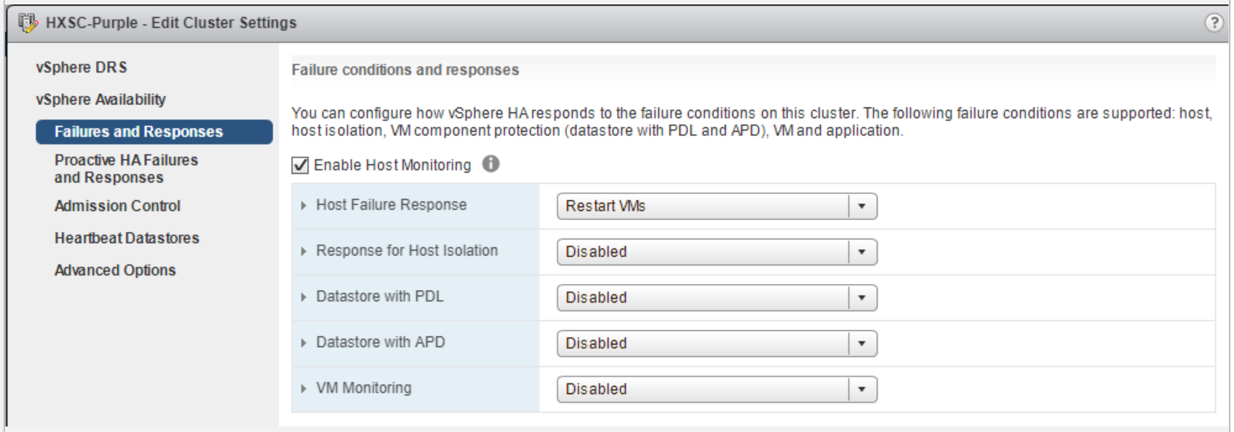
☑ Turn ON vSphere HA

□ Turn on Proactive HA

Failure	Response	Details
Host failure	✓ Restart VMs	Restart VMs using VM restart priority ordering.
Proactive HA	◆ Disabled	Proactive HA is not enabled.
Host Isolation	◆ Disabled	VMs on isolated hosts will remain powered on.
Datastore with Permanent Device Loss	◆ Disabled	Datastore protection for All Paths Down and Permanent Device Loss is disabled.
Datastore with All Paths Down	◆ Disabled	Datastore protection for All Paths Down and Permanent Device Loss is disabled.
Guest not heartbeating	◆ Disabled	VM and application monitoring disabled.

Host Monitoring

Enabled



HXSC-Purple - Edit Cluster Settings

Failure conditions and responses

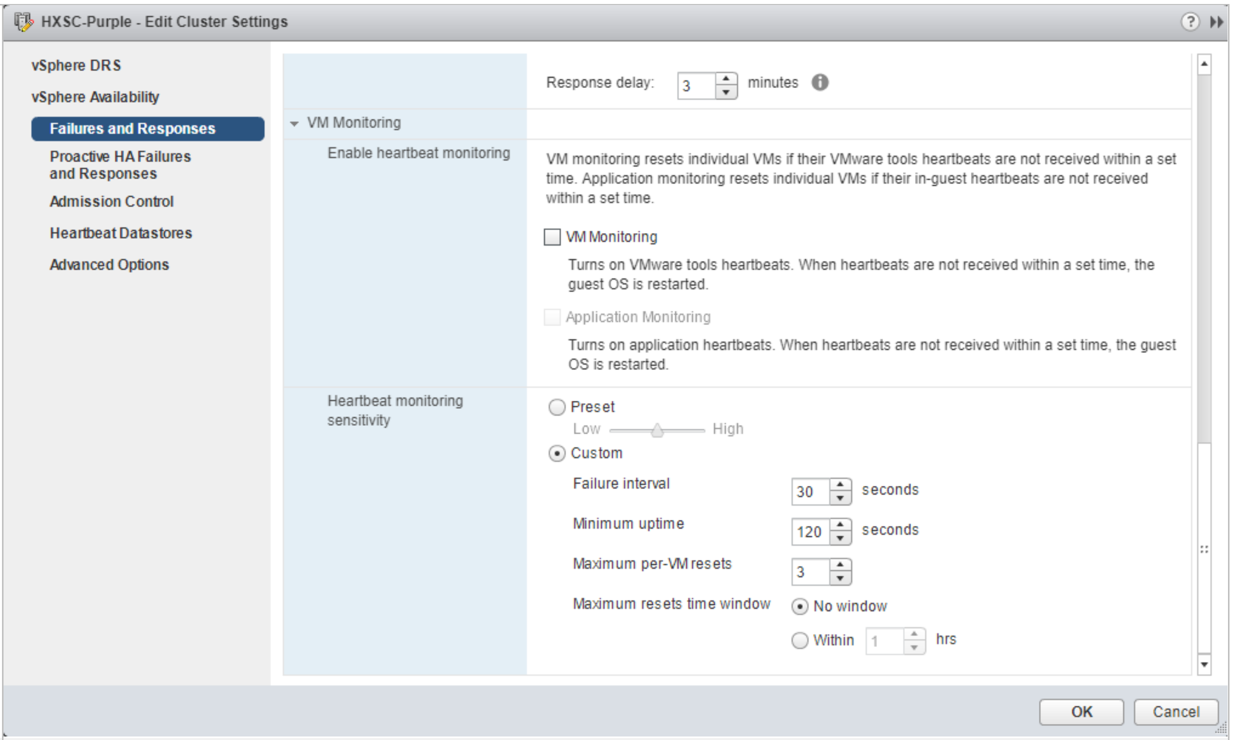
You can configure how vSphere HA responds to the failure conditions on this cluster. The following failure conditions are supported: host, host isolation, VM component protection (datastore with PDL and APD), VM and application.

☑ Enable Host Monitoring

Host Failure Response	Restart VMs
Response for Host Isolation	Disabled
Datastore with PDL	Disabled
Datastore with APD	Disabled
VM Monitoring	Disabled

Virtual Machine Monitoring

Customer Preference – Disabled by default

HXSC-Purple - Edit Cluster Settings

- vSphere DRS
- vSphere Availability
  - Failures and Responses
  - Proactive HA Failures and Responses
  - Admission Control
  - Heartbeat Datastores
  - Advanced Options

Response delay: 3 minutes

VM Monitoring

Enable heartbeat monitoring: VM monitoring resets individual VMs if their VMware tools heartbeats are not received within a set time. Application monitoring resets individual VMs if their in-guest heartbeats are not received within a set time.

□ VM Monitoring

Turns on VMware tools heartbeats. When heartbeats are not received within a set time, the guest OS is restarted.

□ Application Monitoring

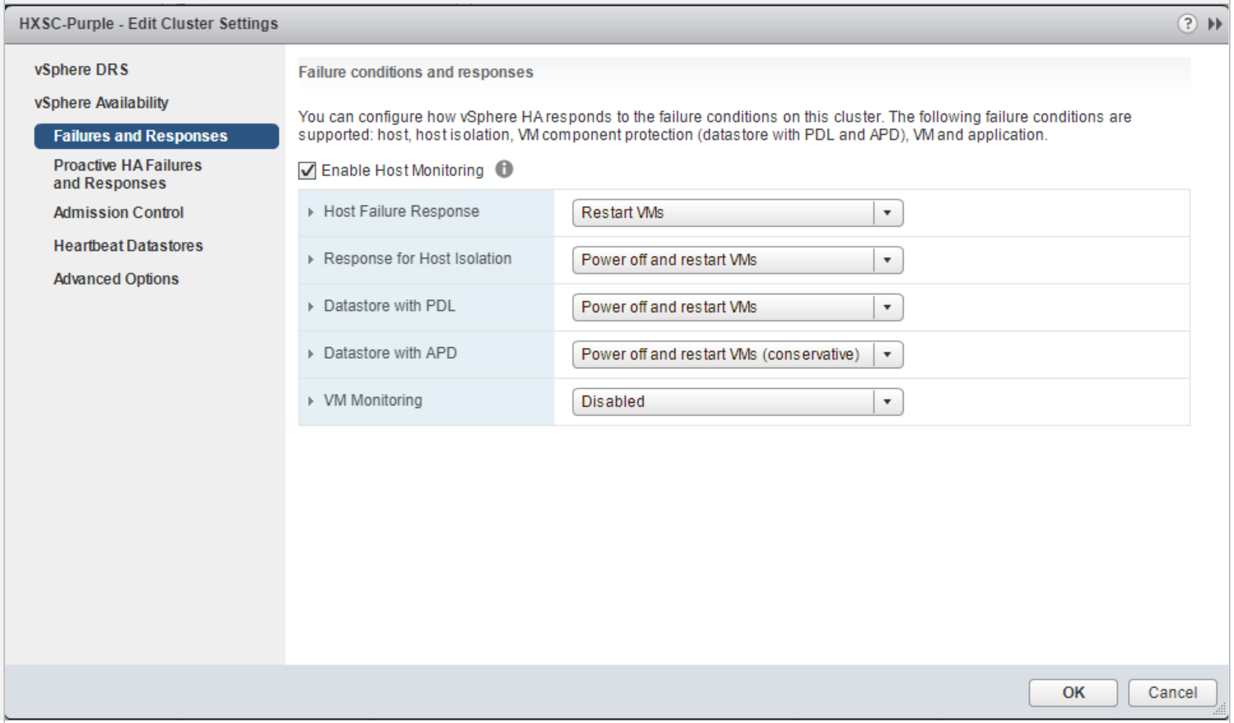
Turns on application heartbeats. When heartbeats are not received within a set time, the guest OS is restarted.

Heartbeat monitoring sensitivity:

○ Preset (Low – High)

● Custom

Failure interval	30 seconds
Minimum uptime	120 seconds
Maximum per-VM resets	3
Maximum resets time window	● No window ○ Within 1 hrs

Failure conditions and VM Response	Host monitoring is enabled, Response for Host Isolation is set to Power off and Restart VMs. For PDL and APD, select Power off and Restart from the drop downs. 

HXSC-Purple - Edit Cluster Settings

- vSphere DRS
- vSphere Availability
  - Failures and Responses
  - Proactive HA Failures and Responses
  - Admission Control
  - Heartbeat Datastores
  - Advanced Options

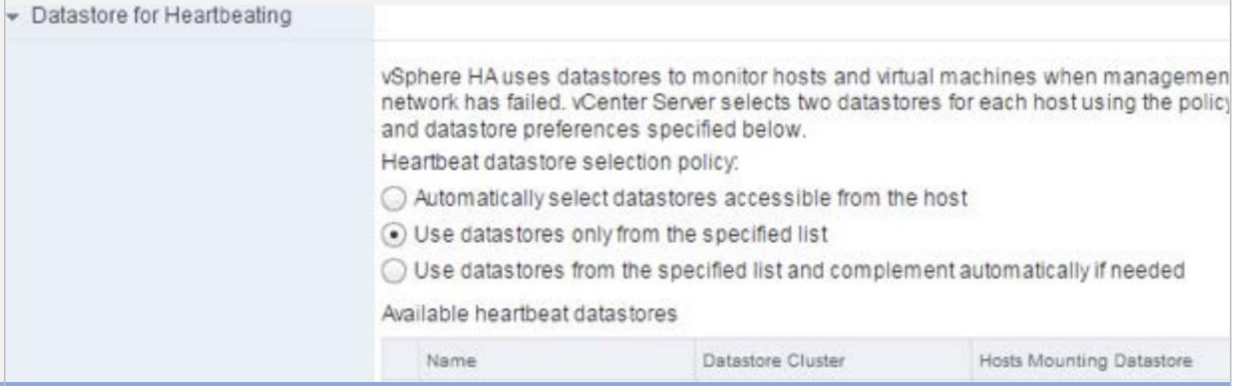
Failure conditions and responses

You can configure how vSphere HA responds to the failure conditions on this cluster. The following failure conditions are supported: host, host isolation, VM component protection (datastore with PDL and APD), VM and application.

☑ Enable Host Monitoring

Host Failure Response	Restart VMs
Response for Host Isolation	Power off and restart VMs
Datastore with PDL	Power off and restart VMs
Datastore with APD	Power off and restart VMs (conservative)
VM Monitoring	Disabled

Admission Control	Set to disable

Datastore Heartbeats	"Use datastores only from the specified list" and select HX datastores. https://kb.vmware.com/s/article/2004739  ▾ Datastore for Heartbeating vSphere HA uses datastores to monitor hosts and virtual machines when managemen network has failed. vCenter Server selects two datastores for each host using the polic and datastore preferences specified below. Heartbeat datastore selection policy: ○ Automatically select datastores accessible from the host ● Use datastores only from the specified list ○ Use datastores from the specified list and complement automatically if needed Available heartbeat datastores Name | Datastore Cluster | Hosts Mounting Datastore
Advanced Settings	
das.usedefaultisolationaddress	False
das.isolationaddress0	IP address for Management Network Gateway
das.isolationaddress1	Existing IP address that is outside cluster. Do not use FI VIPs, Cluster IP (CIP), or cluster host IP

Witness configuration – Witness VM

A quorum is the minimum number of votes that a distributed transaction must obtain to be allowed to perform an operation in a distributed system. A quorum-based technique is implemented to enforce consistent operation in a distributed system. The witness node serves this function. In the event of a split-brain condition, in which both sites are still available but unable to communicate with each other, a virtual machine site leader must be established so that two instances of the same virtual machine are not brought online by HA.

The witness is deployed at a third site and is delivered as an open virtual appliance (OVA) file for use in an infrastructure ESXi deployment at that location. The witness runs an instance of ZooKeeper (see the "Apache ZooKeeper" section earlier in this document for details), becomes a cluster member, and contributes its vote when needed to break a tie.

The witness node must have the following characteristics:

- A third independent site is needed to host the witness virtual machine.
- IP address and connectivity for the witness virtual machine is needed to each Stretch cluster site.
- The witness must be on a routable Layer 3 network.
- The minimum requirements for the witness node are as follows:
  - Virtual CPUs (vCPUs): 4
  - Memory: 8 GB
  - Storage: 40 GB
  - HA: Optional for the witness node
- Latency of at most 100-ms RTT to each site is required.
- Bandwidth of at least 100 Mbps to each site is required.
- For fastest site-to-site failover times, an RTT latency to the witness of less than 10ms is optimal.
- The node must be deployed separately before the Cisco HyperFlex installer Stretch cluster workflow is run.
- **The witness behaves as a quorum node, if you are reinstalling the cluster the witness must be reinstalled as well.**
- **THERE IS ONE WITNESS PER CLUSTER. MULTIPLE CLUSTERS CANNOT USE THE SAME WITNESS.**

While no user data is being sent between the sites and the witness, some storage-cluster metadata traffic is transmitted to the witness site. This traffic is the reason for the 100-Mbps requirement and is in line with competitive products. The witness connection to each site requires 100 Mbps bandwidth with a 100 ms RTT in order to function properly. It is recommended to use a connection with a 100 ms latency for proper system failover behavior. For large clusters and for the best site-to-site failover performance, Cisco recommends witness-to-site latency on the order of 10 ms.

The witness is currently not supported in public cloud deployments because of testing limitations. The OVA file has been tested and is supported for the ESXi platform.

If you need to patch the witness virtual machine for any reason, you can take the witness offline temporarily, implement the update, and bring the witness back online. Cisco recommends that you stage this process and practice it on a test witness to help ensure timely reintroduction of the production system when you implement the actual update. The cluster must be in a healthy condition to conduct this operation. If you need assistance, please contact the Cisco Technical Assistance Center (TAC).

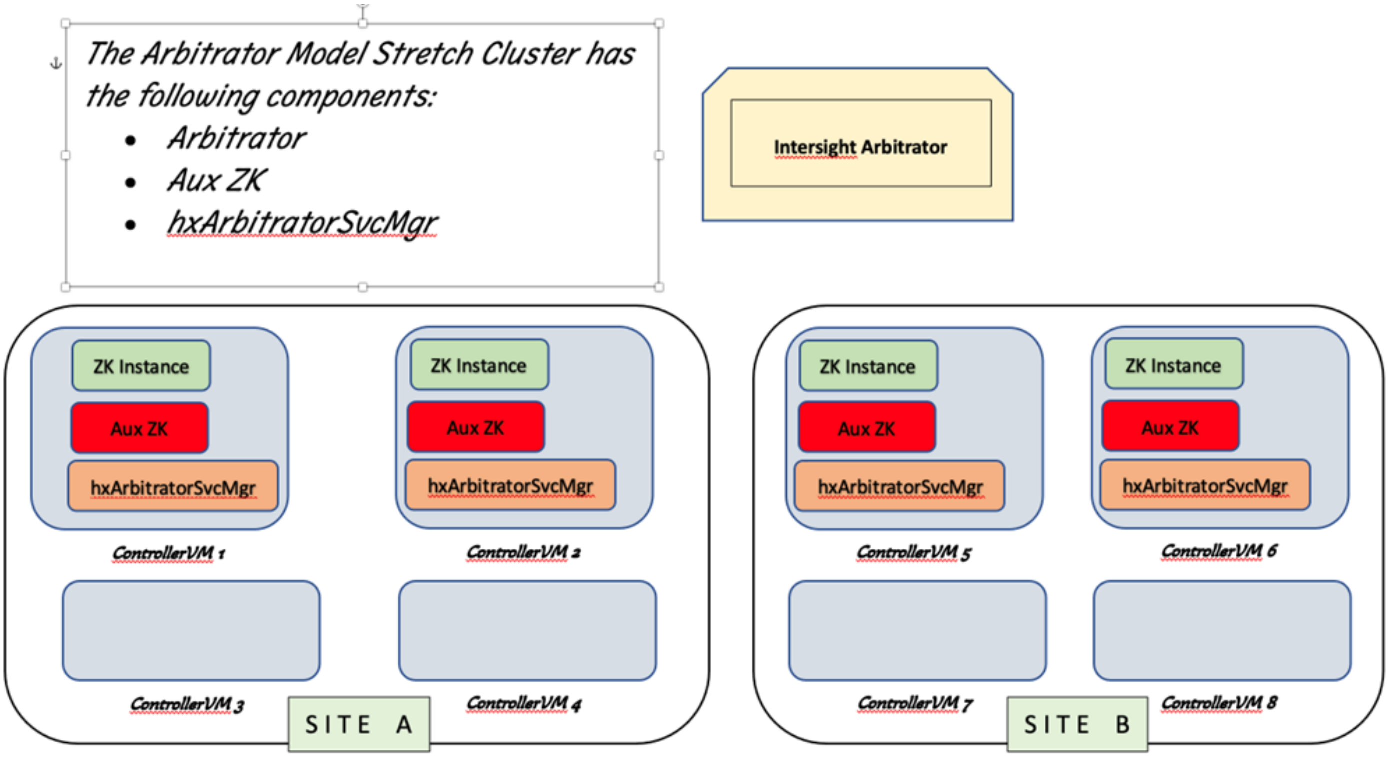
Witness configuration – Intersight Arbitrator

Beginning with HXDP 5.5(1a), stretch clusters can use the Intersight Arbitrator service as a witness. Some new terminology is introduced and defined below to help understand how the Arbitrator functions.

- **An Arbitrator** is an independent entity which resolves conflicts and settles dispute in a distributed system
- [illegible]

This new architecture is called the Arbitrator Model. Any new or rebuilt stretch clusters created with HXDP 5.5 and above require the use of the Arbitrator Model. The cluster is claimed in Intersight, but not built with Intersight as other HX variants often are. Clusters upgraded to HXDP from a release prior to 5.5 will still be able to retain their physical witness. The HX Installer has been updated to reflect these choices:

- For fresh deployments the installer uses the arbitrator, and a preferred site is designated
- For an upgrade you can continue using the witness VM you have already deployed, or move to the arbitrator service



The Arbitrator Model Stretch Cluster has the following components:

- *Arbitrator*
- *Aux ZK*
- *hxArbitratorSvcMgr*

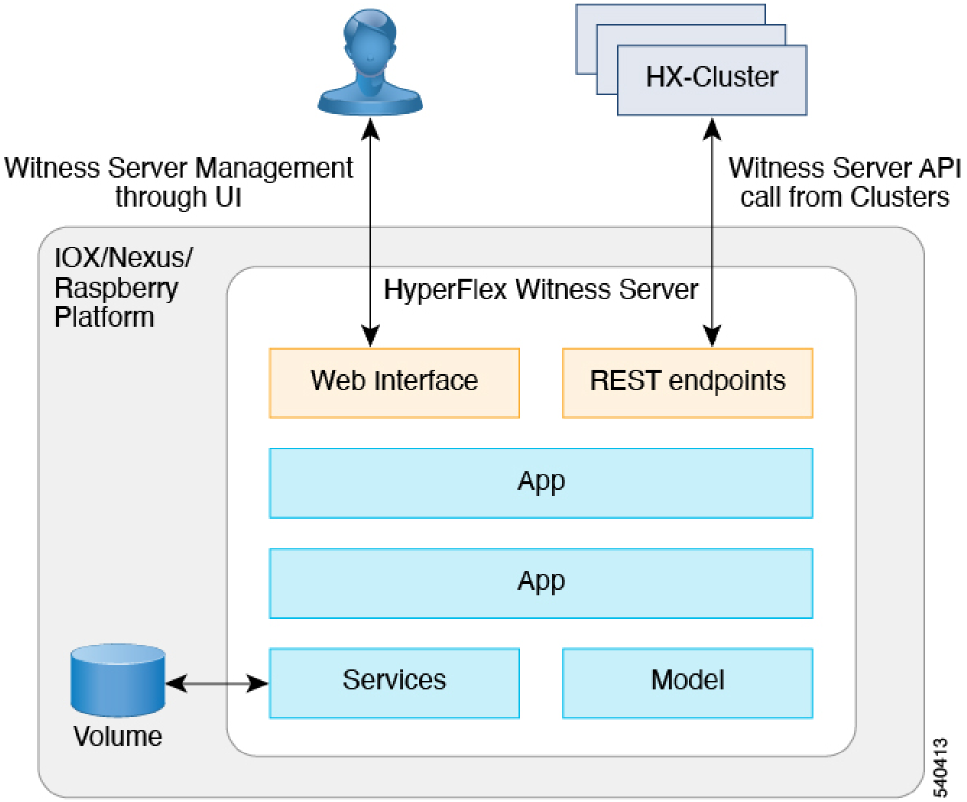
If you have upgraded to HXDP 5.5.x and retained the witness model, you can switch to the arbitrator model by running a set of CLI commands from the secure admin shell. This feature requires SSH to be enabled.

- First, input the auxiliary IP – the auxiliary ip has to be from the data network
- The next input is the preferred site
- Finally, select your arbitrator deployment style: Intersight or custom local third site (PVA or containerized)

The conversion process takes 5 to 6 minutes to complete. There may be some transient errors as stMgr and hxsvc services are restarted. During the switchover the witness is offline. The cluster cannot tolerate a site failure since the minimum zookeeper quorum is unattainable, but a single node failure can still be tolerated. Beyond that, the workflow will fail to complete the conversion.

During the switchover workflow, the witness IP is replaced with the aux IP across all 4 ZK ensemble nodes. Since this IP is floating and can only be brought up or down by hxArbitratorSvcMgr service, it can only happen on the ZK ensemble nodes. There is no way to bring up the Aux IP and Aux ZK in a non-ensemble node. The nodes that are not part of the ZK ensemble (standalone nodes) in the cluster will never get involved in the ZK or Arbitration workflow.

Once the cluster is built with the arbitrator, you cannot revert to a witness VM based deployment. For air-gapped environments, you may use either the Intersight Private Virtual Appliance (PVA) or a containerized arbitrator can be used instead of PVA. This containerized arbitrator is also called a local arbitrator. See the deployment guide for details. The diagram below describes a typical containerized arbitrator deployment.


HX-Cluster
Witness Server Management through UI
Witness Server API call from Clusters
IOX/Nexus/ Raspberry Platform
HyperFlex Witness Server
Web Interface
REST endpoints
App
App
Services
Model
Volume
540413

https://intersight.com/help/saas/resources/deploy_local_container__hyperflex_witness_servers

Arbitrator Operation and Failure Behavior

In a cluster operation steady state, there is no Auxiliary ZooKeeper (ZK) running and there are 2 instances of ZK running on each site. When the number of active ZK nodes goes down to 2, the ZK nodes in the preferred site will immediately contest for the arbitrator lock, but the nodes in non-preferred site will back off for a configured period and then contest for the lock.

- If one site goes down, 2 nodes on the surviving site have the hxArbitratorSvcMgr running on 2 ZK ensemble nodes, out of which one of the hxArbitratorSvcMgr will acquire the arbitrator lock and start the Aux ZK locally on its own node. In this instance there will be two ZK servers running on that node. The local ZK is bound to the eth1 interface and the Aux ZK is bound to Aux Interface. This is similar to how a 2-node HX Edge cluster operates. These two ZK servers and the Aux ZK server that was just instantiated on the site (total 3) will form the quorum.
- In the split-brain case, neither site knows if it's a split-brain or if a site is down. The hxArbitratorSvcMgr service on the non-preferred site waits for a configurable delay (10s default) to allow the preferred site to get an arbitrator lock. After the timeout, it applies for a lock. If it was a true split-brain, the preferred site should have acquired the lock by this point and the non-preferred site is denied the lock. If it was a site down of the preferred site however, the non-preferred site will be able to acquire the lock (after the delay), start Aux ZK, and take over the cluster operations.

A split-brain scenario is the only case where the arbitrator is engaged for establishing takeover by the preferred site. If the preferred site in a split-brain cannot communicate with arbitrator, that is a double failure at the preferred site. If the second site can communicate with arbitrator, then the cluster fails over to the second site. If no site can communicate with the arbitrator, then the behavior is the same as the legacy witness failure scenario. See the Failure Modes section below.

I/O path in a Stretch cluster

A Stretch cluster is in active-active mode at each site: that is, primary copies and read traffic occur for each virtual machine at each site. There is no concept of an active-standby configuration in a Stretch cluster. IO Visor, the Cisco HyperFlex file system proxy manager, dictates which nodes service which read and write requests. In general, a Stretch cluster behaves the same way as a normal cluster with modifications for host affinity and certain failure scenarios (see the "Stretch cluster failure modes" section later in this document). With virtual machine affinity and a replication factor of 2 + 2, the read and write dynamics are as described in the following sections.

Read path

Taking advantage of the host group affinity, all read operations for virtual machine data are served locally, meaning that they come from the nodes at the site to which the data store for the virtual machine is assigned. Read operations are first serviced by the node cache if they are available there. If they are not available, they are read from persistent disk space (in a hybrid node) and served to the end user. The read cache in a Stretch cluster behaves the same way as in a normal hybrid or all-flash cluster with the exception of local service based on host affinity.

Write path

Write operations in a Stretch cluster are a little more complicated than read operations. This is the case because to achieve data integrity, a write operation is not acknowledged as committed to the virtual machine guest operating system until all copies, local and remote, are internally committed to disk. This means that a virtual machine with affinity to site A will write its two local copies to site A while synchronously writing its two remote copies to site B. Again, IO Visor determines which nodes are used to complete each write operation.

The Cisco HyperFlex file system waits indefinitely for write operations to be acknowledged from all active copies. Thus, if certain nodes or disks that host a copy of data for which a write operation is being implemented are removed, write operations will stall until a failure is detected (based on a timeout value of 10 seconds) or the failure heals automatically without detection. There will be no inconsistency in either case.

I/O operations from virtual machines on site A will be intercepted by IO Visor on site A. IO Visor on site B is not involved. The write I/O operations are replicated to site B at the data platform level. In the event of virtual machine migration from one site to another—for example, through VMware Storage vMotion from site A to another data store with affinity to site B—IO Visor will conduct a hand-off. When a virtual machine migrates to site B, IO Visor on site B will intercept the I/O operations. This procedure is also part of the virtual machine failover process internally. After the virtual machines have migrated from site A to site B, virtual machine I/O operations will not be intercepted by the site A IO Visor, but rather by the site B IO Visor.

Sizing

Typically, you start sizing exercises by profiling the workload or already knowing the requirements for the virtual machines that you need to run. However, you come by this information, the next step is to use a sizing tool (unless you want to do the math yourself). Cisco provides a sizing tool that can run workload estimates for a Stretch cluster with a typical VSI profile:

Cisco HyperFlex sizer tool: https://HyperFlexsizer.cloudapps.cisco.com/ui/index.html#/scenario

Sizing a Stretch cluster requires an understanding of the replication factor used for data protection. Each site runs a replication factor of 2: that is, each site has a primary copy and a replica. Each site also runs a replication factor of 2 for the complementary site, so that for each virtual machine, across both sites, there is a primary copy and three replicas: equivalent to a replication factor of 4. This configuration is required so that any individual site can tolerate the loss of its complementary site and still be able to run. Note that the loss of a site does not guarantee the ability of the surviving site to tolerate a disk or node loss because the affected node might be a zookeeper node. When the cluster is created, a zookeeper leader is elected at a given site. The leader is used to make updates to the ensemble. In the event of a site or zookeeper leader failure, a new leader is elected. This is not configurable.

Survivability while maintaining online status requires a majority zookeeper quorum and more than 50% of nodes (the witness counts as both an active zookeeper node). It is possible that the surviving site could tolerate a node or disk loss (in a cluster greater than 2+2) if that node is not a zookeeper node, but it is not guaranteed.

The data protection and workload profile (I/O requirement) considerations allow you to determine the number and type of disks required to meet your capacity needs. You then need to determine the node count needed to meet your vCPU and virtual machine memory needs.

Here are some sizing guidelines:

- For VSI an option is available in the sizer for selecting the Stretch cluster. Use this option for your sizing exercises.
- In general, a Stretch cluster uses a replication factor of 4: that is, replication factor 2 + replication factor 2 (a replication factor of 2 at each site with full replication to the complementary site, also at a replication factor of 2). This configuration effectively results in a replication factor of 4.
- You can use a replication factor of 2 for one site and then apply the same factor to the second site. If you want to be able to run all workloads from either site, then you must be sure that you have enough capacity at each site by accounting for the overall workloads and thresholds. The sizer automatically performs this verification for you.
- Consider the virtual machine and vCPU capacity: everything must be able to run comfortably at one site.
- The total virtual machine vCPU capacity is required.
- The total virtual machine memory capacity is required.

Failure sizing

It is not enough to size your deployment for normal operations. Ideally, you should size your deployment for a scenario in which you have lost a site, and the surviving site has lost a non-zookeeper node. This is the worst-case continuous-operation scenario for resource distribution to your overall virtual machine workload. Everything must be able to run comfortably on one site for a Stretch cluster deployment to offer true business continuance in the event of a disaster.

If it is sufficient to run only certain virtual machines at the surviving site, you may be able to undersize the system, but you need to be aware of this and take it into consideration when planning disaster-recovery runbooks. Keep in mind that the automated recovery mechanism of the Stretch cluster will launch virtual machines from failed sites without user intervention. You may find yourself in a situation in which you need to turn off failover virtual machines if they exceed the capacity of the surviving site.

Bandwidth Considerations for the Inter-Site Link Based on Workloads

Read bandwidth is normally local only, so there is no dependence or impact on the site-to-site link. Non-local VMs, i.e., VMs running on nodes that do not have the assigned datastore affinity, will incur link read traffic. This is not the typical situation but should be considered in corner-case scenarios.

Write bandwidth is necessarily relevant to the link: Replicas traverse the link (2 copies). There is also meta data overhead for the filesystem that traverses the link making the write bandwidth some multiplier greater than 2. A typical good estimate is to add 20% to the write IOPS to account for this overhead (i.e., a factor of 1.2).

Workloads are almost never 100% read or 100% write. Typical benchmarks use a 70% Read/ 30% Write workload distribution. This means that for a 100,000 IOPS workload, 70,000 would be reads and 30,000 would be writes with a typical block size of 4k in the application. While the cluster writes do not map one-to-one with application writes (they are concatenated and written in chunks), the overall size of the write(s) match.

Link Bandwidth = WIOPS(2 replicas)(1.2 metadata overhead)(4kB) + RIOPS(4kB) + ResynchIOPS(2 replicas)(4kB) + vMotionBW

Where WIOPS are Write IOPS, RIOPS are Read IOPS, ResynchIOPS are resynchronization operations from any potential failure recoveries, and vMotionBW is the bandwidth taken up by a VM move (both compute and storage to account for datastore affinity when moved between sites). Resynchronizations only happen on failure recovery and are transitory operations so we will ignore them here. Storage vMotion is also typically not undertaken, but we will consider it in the example below.

Example: 20,000 IOPS total cluster workload, one affinity-displaced VM contributing 1000 IOPS in 4kB Reads, no resynchronization, and 1 full SVMotion running at 500Mb/s. Assume 70/30 breakdown for the read and writes in the main workload. Note that the 20k workload RIOPS are not factored in the link traffic since this traffic is locally served by the site(s) and does not traverse the link.

Link BW = 0.3(20000)(2)(1.2)(4kB)+1000(4kB) + 0 + 500Mb/s

Link BW = 14,400 kB/s + 4000 kB/s + 500Mb/s = 18400 kB/s + 500 Mb/s = (18400)*8/1024 Mb/s + 500 Mb/s

Link BW = 143.75 Mb/s + 500 Mb/s

Link BW ≈ 643.75 Mbps

Since you will not often do resync or vMotion, this can be considered a peak link value for the 20000 IOPS workload examined. There are times, for example, during large, frequent deletes, where the file system cleaner can incur larger metadata traffic on the link. To estimate those, you can use a multiplier of 1.3 to 1.5 instead of 1.2 for the (temporary peak) metadata value.

Stretch cluster installation

Before conducting any installation, please refer to and complete the preinstallation checklist maintained here:

http://www.cisco.com/c/dam/en/us/td/docs/hyperconverged_systems/HyperFlex_HX_DataPlatformSoftware/HyperFlex_preinstall_checklist/Cisco_HX_Data_Platform_Preinstallation_Checklist_form.pdf

This checklist is essential to a smooth and timely installation process. You must also read the release notes for the Cisco HyperFlex build that you will be installing:

https://www.cisco.com/c/en/us/support/hyperconverged-systems/HyperFlex-hx-data-platform-software/products-release-notes-list.html

The installation process has several prerequisites:

- Locations must be appropriate. Distances cannot exceed 100 km.
- Two fabric interconnects with uplinks to the site-to-site switch must be in place at each location.
- A Stretch Layer 2 network must be in place between sites for the storage data VLAN and the management VLAN, and the storage data VLAN should use jumbo frames for best performance but 1500 MTU is supported.
- The site-to-site bandwidth should be at least 10 Gbps with a RTT latency of at most 5 ms.
- Test the RTT ping using **VMkping –I VMk1 -d -s 8972 x.x.x.x** from any ESXi host in your cluster. This check is also performed by the installer, and if it fails, the installation will not proceed.
- Use symmetric converged nodes and a new installation or clean repurposing of the supported model.
- The witness should meet these criteria:
  - Use ESXi at the third site to deploy the OVA file.
  - The connection, at worst, should be 100 Mbps with a 100-ms RTT latency when tested with ping.
  - Site failover times are directly related to witness latency, so for the fastest possible failover times, the RTT to the witness should be under 10ms.
- vCenter must be installed at the third routable site in advance.
- Deploy the installer OVA in your network so that it can reach both sites and the witness (you can deploy it at the witness location on that infrastructure if needed).
- Complete the preinstallation checklist.
- Read the release notes.
- Be sure to run the post-installation script from the installer CLI once the deployment is complete.
  - You will be prompted for both UCSM domain credentials.
- **The witness behaves as a quorum node, if you are reinstalling the cluster the witness must be reinstalled as well.**
- script from the installer CLI once the deployment is complete.

Note that when you are installing the cluster and selecting available nodes, select the nodes you are going to add to the cluster and not the “existing nodes”. See the Stretch Cluster administration documentation for additional information.

Enabling jumbo frames for the storage network during installation is a simple checkbox. There are some things going on behind the scenes, however.

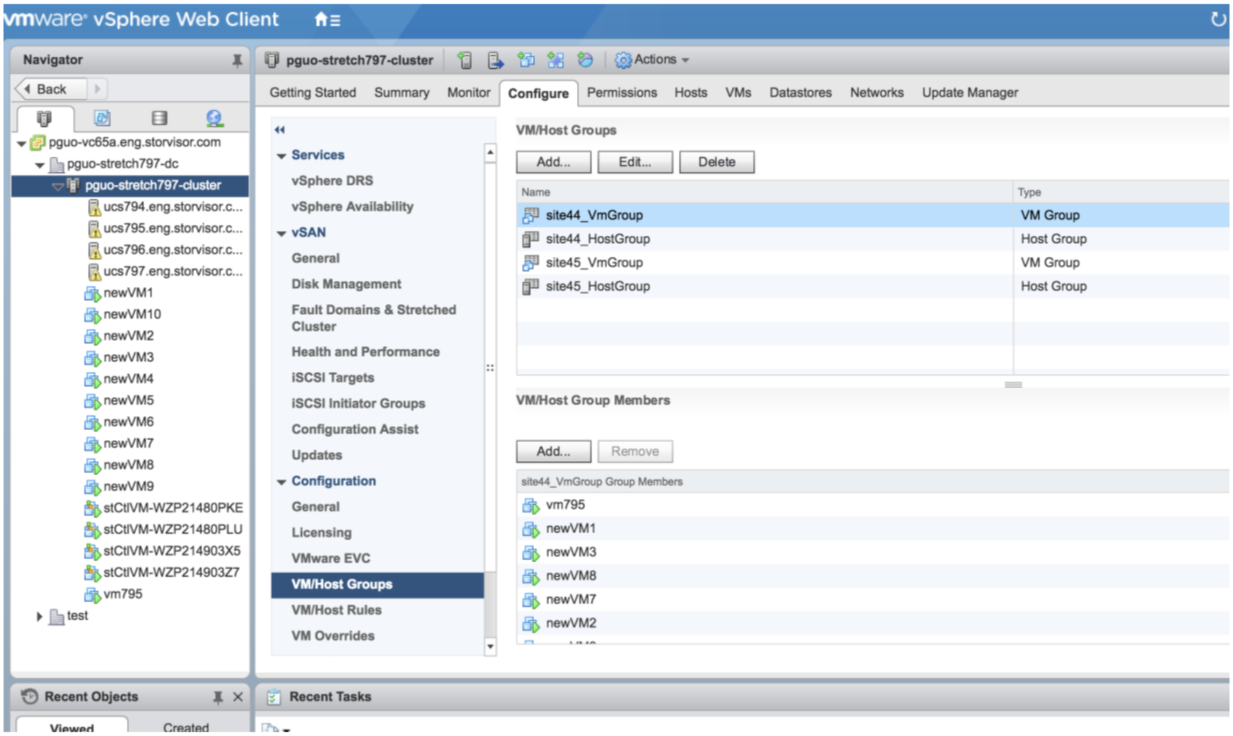
- When users deploy HX with jumbo frames enabled (default setting), the MTU is configured as 9000 bytes in the controller VM interface (for storage traffic).
- The actual maximum size of an ethernet frame on the wire would therefore be:
    - 9000 bytes of IP payload (includes 20 byte IP header)
    - 14 byte Ethernet header
    - 4 byte Ethernet CRC
    - 4 byte VLAN header
    - 9022 bytes in total
- This 9022 bytes does not include other overheads/encapsulations that may be in use across the inter-site link. You must add these if anything else is added to the ethernet frame above
- Be careful with the definition of MTU – sometimes it is used to mean IP MTU, sometimes it is the full packet size including all ethernet headers and CRC.
- As long as the 9000 byte IP packet can traverse the network without being fragmented, you should be safe to proceed. Please do some exhaustive ping tests with fragmentation bit set to DF to ensure that the end-to-end path can support the storage traffic. These tests can be run from the admin shell of the controller VMs.
    - E.g. *ping -M do -s 8972 -I eth1 <eth1 IP of other controller VM on other site>*
    - Repeat across multiple controller VMs and between controller VMs and vmk1 on ESXi across both sites

VMware vCenter has some installation preinstallation and post-installation requirements as well, but the detailed site configuration is handled automatically and is therefore simple to implement.

- VMware DRS and HA should be enabled. DRS is enabled automatically. If DRS is not enabled, virtual machines will run on any site. The VM-Host Affinity rules added to the vCenter during the deployment will not be enforced properly, hence the VMs which are intended to run on site-1 can move to site-2, this will un-necessarily increase the utilization of the WAN link and increase read latency.
- Set site affinity to the preferred host groups per site. Virtual machine affinity rules and host groups are created automatically.
    - Verify the virtual machine and host affinity groups: one for each site.
    - The affinity group consists of virtual machines and hosts from each site.
    - Verify that VM/Host Rules is set to the “should” clause.

Figure 8 shows the vCenter screen on which you can verify host affinity groups.

Figure 8. Virtual machine and host groups affinity verification in VMware vCenter



VM/Host Groups

Name	Type
site44_VmGroup	VM Group
site44_HostGroup	Host Group
site45_VmGroup	VM Group
site45_HostGroup	Host Group

VM/Host Group Members

site44_VmGroup Group Members
vm795
newVM1
newVM3
newVM8
newVM7
newVM2

Moving VMs Around After Deployment

In general, compute vMotion of VMs between sites should not be conducted. This will break the local read operations established by site affinity which will cause read latency. If there is a need to move a VM from Site 1 to Site 2 permanently, perform a Compute and a Storage vMotion to a datastore which has affinity set to Site2.

Site-to-Site Link Security

To begin the discussion regarding the inter-site link, we need to define and accept some common terms:

Dark Fiber

The term ‘Dark Fiber’ has evolved to encompass the practice of leasing ‘dark’ fiber optic cables from network providers and operators. A client will lease or purchase unused strands of ‘dark’ fiber optic cable to create their own privately-operated optical fiber network rather than just leasing bandwidth. The Dark Fiber network is separate from the main network and is controlled by the client rather than the network provider.

Dark Fiber networks can be set up in a variety of ways, including dark fiber rings, point to point or point-to-multipoint configurations. With Dark Fiber, a client can expect to get high levels of performance, a highly secure network and superfast speeds.

Layer 2 (L2)

Layer 2, also known as the Data Link Layer, is the second level in the seven-layer OSI reference model for network protocol design. Layer 2 is equivalent to the link layer (the lowest layer) in the TCP/IP network model. Layer2 is the network layer used to transfer data between adjacent network nodes in a wide area network or between nodes on the same local area network.

Layer 3 (L3)

Layer 3 refers to the third layer of the Open Systems Interconnection (OSI) Model, which is the network layer.

Layer 3 is responsible for all packet forwarding between intermediate routers, as opposed to Layer 2 (the data link layer), which is responsible for media access control and flow control, as well as error checking of Layer 1 processes.

Traditional switching operates at layer 2 of the OSI model, where packets are sent to a specific switch port based on destination MAC addresses. Routing operates at layer 3, where packets are sent to a specific next-hop IP address, based on the destination IP address. Devices in the same layer

2 segment do not need routing to reach local peers. What is needed however is the destination MAC address which can be resolved through the Address Resolution Protocol (ARP)

VLAN

A VLAN (virtual LAN) abstracts the idea of the local area network (LAN) by providing data link connectivity for a subnet. One or more network switches may support multiple, independent VLANs, creating Layer 2 (data link) implementations of subnets. A VLAN is associated with a broadcast domain. It is usually composed of one or more Ethernet switches.

VLANs make it easy for network administrators to partition a single switched network to match the functional and security requirements of their systems without having to run new cables or make major changes in their current network infrastructure. Ports on switches can be assigned to one or more VLANs, enabling systems to be divided into logical groups -- based on which department they are associated with -- and establish rules about how systems in the separate groups are allowed to communicate with each other. These groups can range from the simple and practical (computers in one VLAN can see the printer on that VLAN, but computers outside that VLAN cannot), to the complex and legal (for example, computers in the retail banking departments cannot interact with computers in the trading departments).

Each VLAN provides data link access to all hosts connected to switch ports configured with the same VLAN ID. The VLAN tag is a 12-bit field in the Ethernet header that provides support for up to 4,096 VLANs per switching domain.

Trunk port

A trunk port is a port that is assigned to carry traffic for all the VLANs that are accessible by a specific switch, a process known as trunking. Trunk ports mark frames with unique identifying tags – either 802.1Q tags or Inter-Switch Link (ISL) tags – as they move between switches.

Since VLANs exist in their own layer 3 subnet, routing will need to occur for traffic to flow in between VLANs. In a Stretch Cluster the management VLAN is configured to be routed from the switch to the gateway. The data network VLAN is not routed and can only access layer 2 adjacent nodes (same VLAN same subnet connected by the switch).

Switch configurations are performed by the network administrator. Dark fiber connections are linked via switches when the switch uplink ports between site switches are configured by the network administrator. Dark fiber connections are leased or purchased from a local provider.

The Witness is not accessed over the Stretched link.

Configuring a Trunk Port on a Cisco Nexus 5k.
Configuring VLANs on a Cisco Nexus 5k.

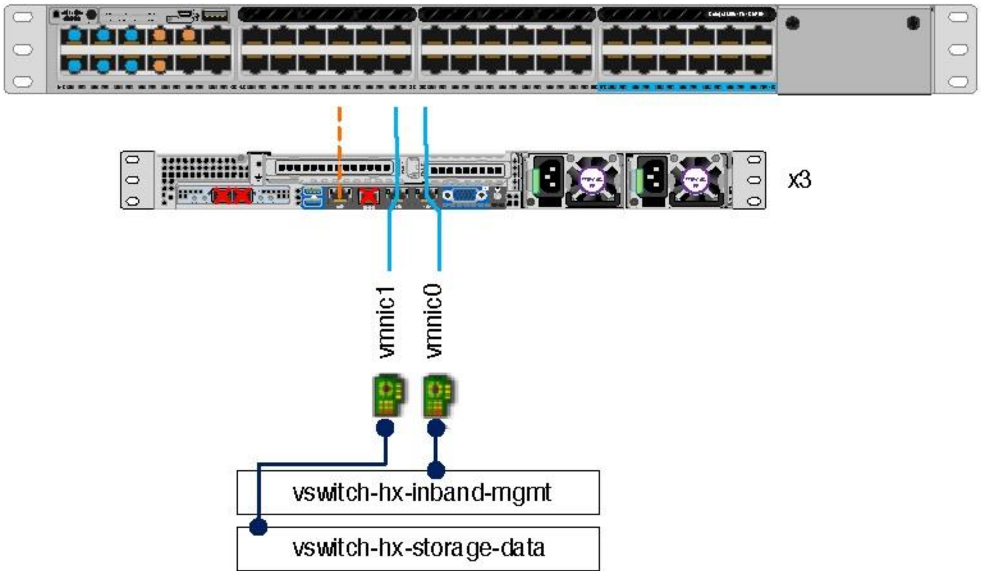
The layer 2 (L2) connection between the Stretch cluster sites is an extension of the layer 2 connectivity present in the Storage Controller Data Network and the Management Network vSwitches. These networks are segmented out by VLAN on the vSwitches.

The Storage Controller Data Network is private and un-routed. All communication on the Storage Controller Data Network is between cluster components that are L2 adjacent on the same network segment. There are no other components on this network and there is no mechanism for outside devices to access this private network.

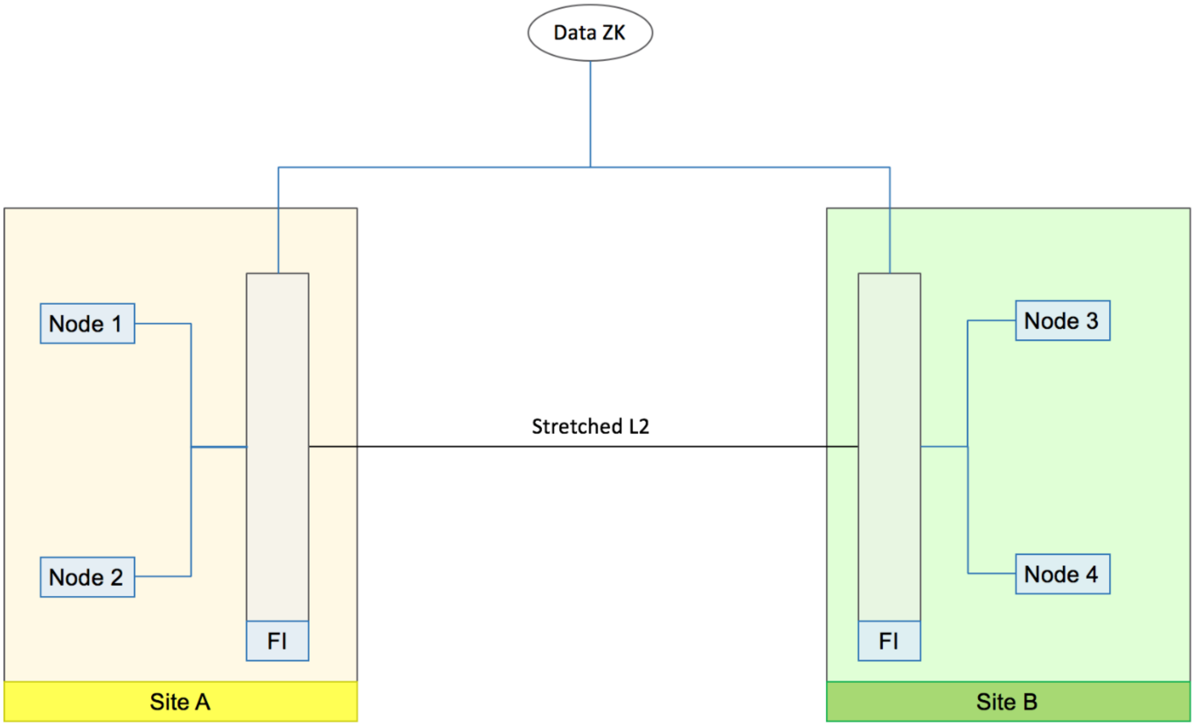
The management network is also extended over the L2 connection. It maintains its normal layer 3 (IP) connectivity for the purposes of external management access as allowed by firewall and routing rules at the border.

For site-to-site communication over the L2 networks, the respective VLANs are extended through a trunk connection from the Fabric Interconnects (FIs) to the site switches (usually Nexus 5k/7k/9k). The site switches are connected to each other over a dark fiber link. VLAN integrity is maintained throughout; i.e., routing to the management network is maintained and private data connection between sites for the storage traffic is maintained.

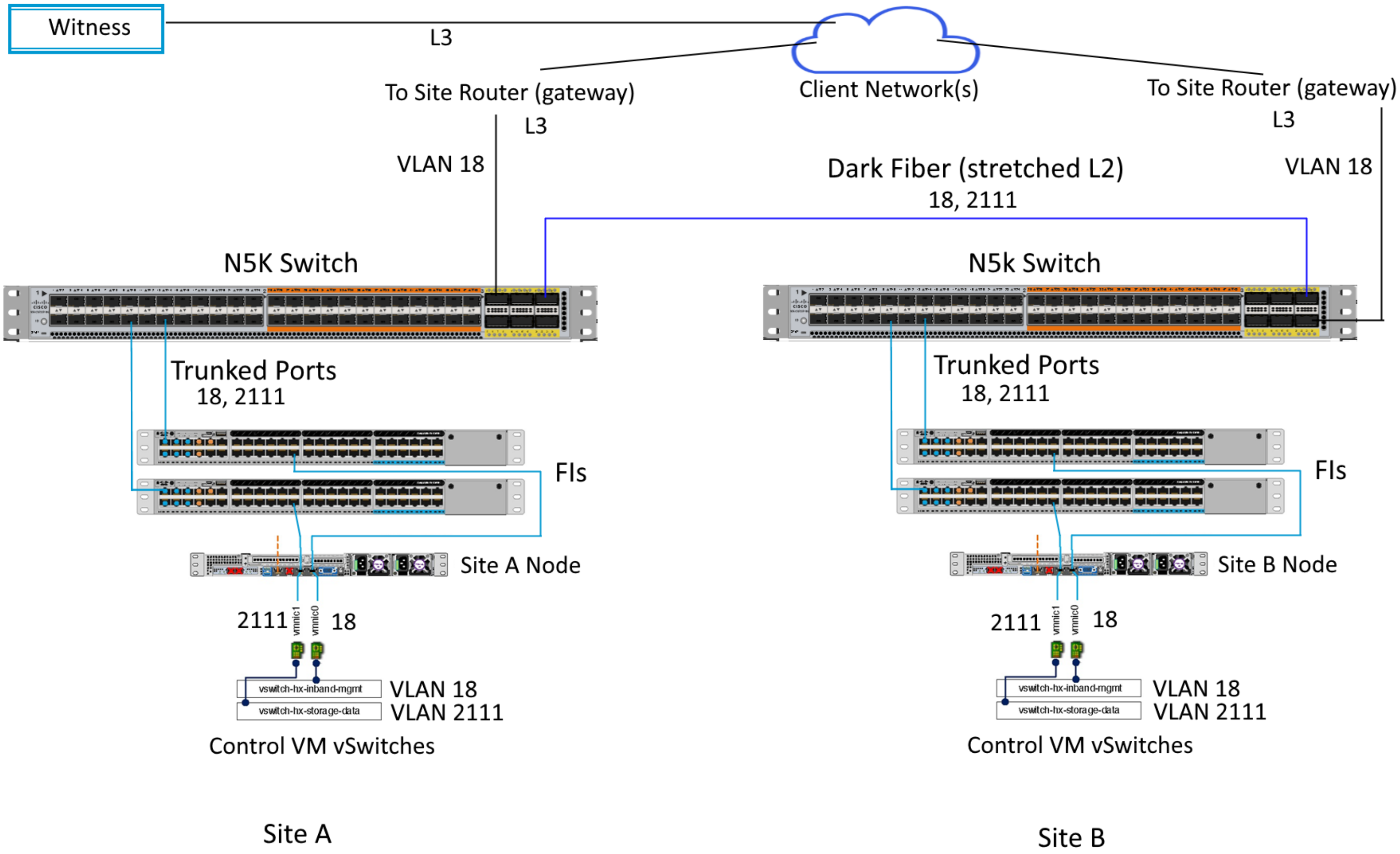
The diagram below shows the physical connectivity of the cluster node to the FI. The FI's have uplinks to the site physical switches that service the Stretched L2 connection. Only the vSwitches that are traversing the L2 site-to-site connection are shown.



The diagram below shows the logical connectivity between sites (with intervening switches omitted because the L2 connection is transparent to the nodes). The remote "Data ZK" icon is the witness.



The following is a detailed view of the Stretched connection taking the defined terms above into consideration. The management network VLAN is shared and accessible externally since it is routed. The data network (2111) is not useable or visible to anyone outside the non-routed 2111 VLAN. VLAN networking enforces the isolation per the VLAN RFC (https://tools.ietf.org/html/rfc5517).



Cisco HyperFlex installer

After you have reviewed the preinstallation checklist and met the prerequisites outlined in the previous section, you are ready to perform the installation. During the initial configuration process, the cluster is installed onsite using the Cisco HyperFlex installer. This installer can safely be removed from the environment immediately after cluster creation. It is typical for secure environments to isolate the deployment network during installation. In this scenario, the installer is never externally available during configuration. Removing the installer after deployment is complete reduces the threat exposure from the installer.

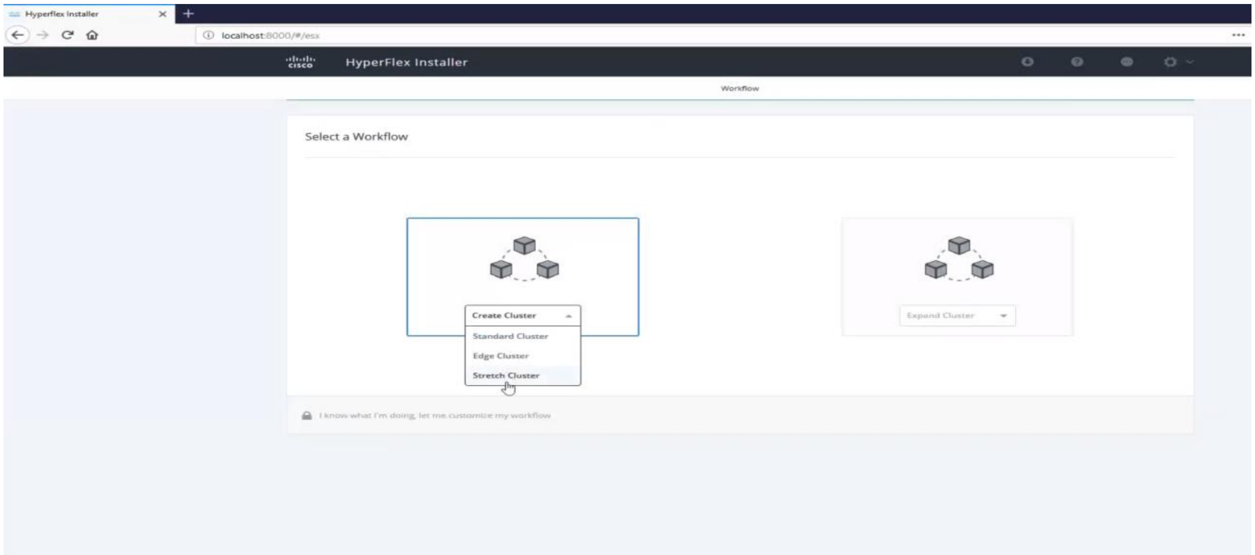
Cisco Intersight™ is a cloud-based installation, management, and upgrade platform that allows the creation and management of Cisco HyperFlex clusters. It supports:

- Cloud installation, with device connectors
- Device ownership
- Day-2 operations

Cisco Intersight is currently available for normal Cisco HyperFlex cluster installation and for Cisco HyperFlex Edge installation on ESXi. Cisco Intersight will be available as an installation and operation platform for Stretch clusters in a future Cisco HyperFlex release.

Boot the installer, open a browser, and enter the IP address of the running installer. You will be presented with a login screen. Use the default root login information detailed in the administration guide to begin. After you are logged in, select the Stretch Cluster workflow as shown in Figure 9.

Figure 9. Cisco hyperflex installer: choosing the Stretch cluster workflow



You will work through the installer site creation workflow twice: one time for each site. You will then run the workflow a final time to create the cluster. Each time through the workflow you will enter data you recorded on your preinstallation checklist. This process is similar to normal cluster creation.

The installer verifies that the cluster components are correct (model, quantity, etc.) and available as needed. This verification process helps ensure that the deployment has no gaps that could jeopardize security or supportability. The installer:

- Helps ensure firmware and BOM compliance
- Examines fabric interconnects through Cisco UCS Manager and generates an appropriate server selection
- Builds and applies service profiles to the nodes
  - VLANS
  - IP addresses
  - vNIC order
  - QoS configuration
  - MAC address pools
- Creates ESX vSwitches with appropriate VLANs and address spaces
- Deploys the HX Data Platform
- Deploys ESX plug-ins
- Deploys and creates the cluster
- Configures and starts the storage cluster
- Sets default passwords and generates secure certificates for node-to-node communication

Strong passwords are enforced on the Cisco HyperFlex user interfaces and HX Data Platform settings during the installation process. Be sure to record these passwords for future reference.

Default passwords

After deployment using the installer is complete, make sure that any default passwords are changed or updated. The ESX hypervisor default password is Cisco123. There are no default passwords for the HX Data Platform nodes because strong passwords are enforced during the installation process. Log in to each ESX node through the CLI and update the root password as needed using **passwd root.**

VLANs and vSwitches

VLANs are created for each type of traffic and for each vSwitch. Typically, four vSwitches are created during the installation process with associated VLANs for each. The vSwitches are for ESX management, Cisco HyperFlex management, ESX data (vMotion traffic), and Cisco HyperFlex data (storage traffic between nodes for the data stores). The HX Data Platform installer creates the vSwitches automatically.

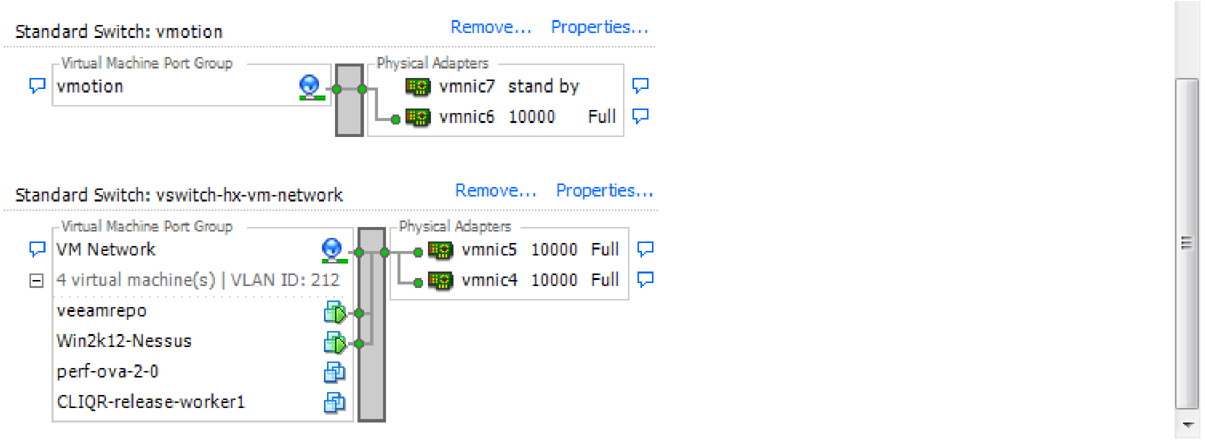
The zones that these switches handle are as follows:

- **Management zone:** This zone consists of the connections needed to manage the physical hardware, the hypervisor hosts, and the storage platform controller virtual machines (HX Data Platform). These interfaces and IP addresses must be available to all staff who will administer the Cisco HyperFlex system throughout the LAN and WAN. This zone must provide access to Domain Name System (DNS) and NTP services and allow Secure Shell (SSH) communication. The VLAN used for management traffic must be able to traverse the network uplinks from the Cisco UCS domain, reaching both fabric interconnect A and fabric interconnect B. This zone contains multiple physical and virtual components:
    - Fabric interconnect management ports
    - Cisco UCS external management interfaces used by the servers and blades, which communicate through the fabric interconnect management ports
    - ESXi host management interfaces
    - Storage controller virtual machine management interfaces
    - A roaming Cisco HyperFlex cluster management interface
- **Virtual machine zone:** This zone consists of the connections needed to service network I/O operations to the guest virtual machines that run inside the Cisco HyperFlex hyperconverged system. This zone typically contains multiple VLANs that are trunked to the Cisco UCS fabric interconnects through the network uplinks and tagged with IEEE 802.1Q VLAN IDs. These interfaces and IP addresses need to be available to all staff and other computer endpoints that communicate with the guest virtual machines in the Cisco HyperFlex system throughout the LAN and WAN.
- **Storage zone:** This zone consists of the connections used by the HX Data Platform software, ESXi hosts, and storage controller virtual machines to service the Cisco HyperFlex distributed file system. These interfaces and IP addresses need to be able to communicate with each other at all times for proper operation. During normal operation, this traffic all occurs within the Cisco UCS domain. However, in some hardware failure scenarios this traffic may need to traverse the network northbound of the Cisco UCS domain. For that reason, the VLAN used for Cisco HyperFlex storage traffic must be able to traverse the network uplinks from the Cisco UCS domain, reaching fabric interconnect A from fabric interconnect B, and fabric interconnect B from fabric interconnect A. This zone contains primarily jumbo frame traffic; therefore, jumbo frames should be enabled on the Cisco UCS uplinks. This zone contains multiple components:
    - A VMkernel interface on each ESXi host in the Cisco HyperFlex cluster, used for storage traffic
    - Storage controller virtual machine storage interfaces
    - A roaming Cisco HyperFlex cluster storage interface
- **vMotion zone:** This zone consists of the connections used by the ESXi hosts to enable vMotion movement of the guest virtual machines from host to host. During normal operation, this traffic all occurs within the Cisco UCS domain. However, in some hardware failure scenarios this traffic may need to traverse the network northbound of the Cisco UCS domain. For that reason, the VLAN used for Cisco HyperFlex storage traffic must be able to traverse the network uplinks from the Cisco UCS domain, reaching fabric interconnect A from fabric interconnect B, and fabric interconnect B from fabric interconnect A. This traffic must be able to traverse sites.

These vSwitches and their associated port groups are tied to a pair of vNICs at each node in an active-standby mode for high availability.

Figures 10 and 11 show the typical networking configuration for a node.

Figure 10. Screen showing virtual machine (user) and vMotion networks



Standard Switch: vmotion

Virtual Machine Port Group
vmotion

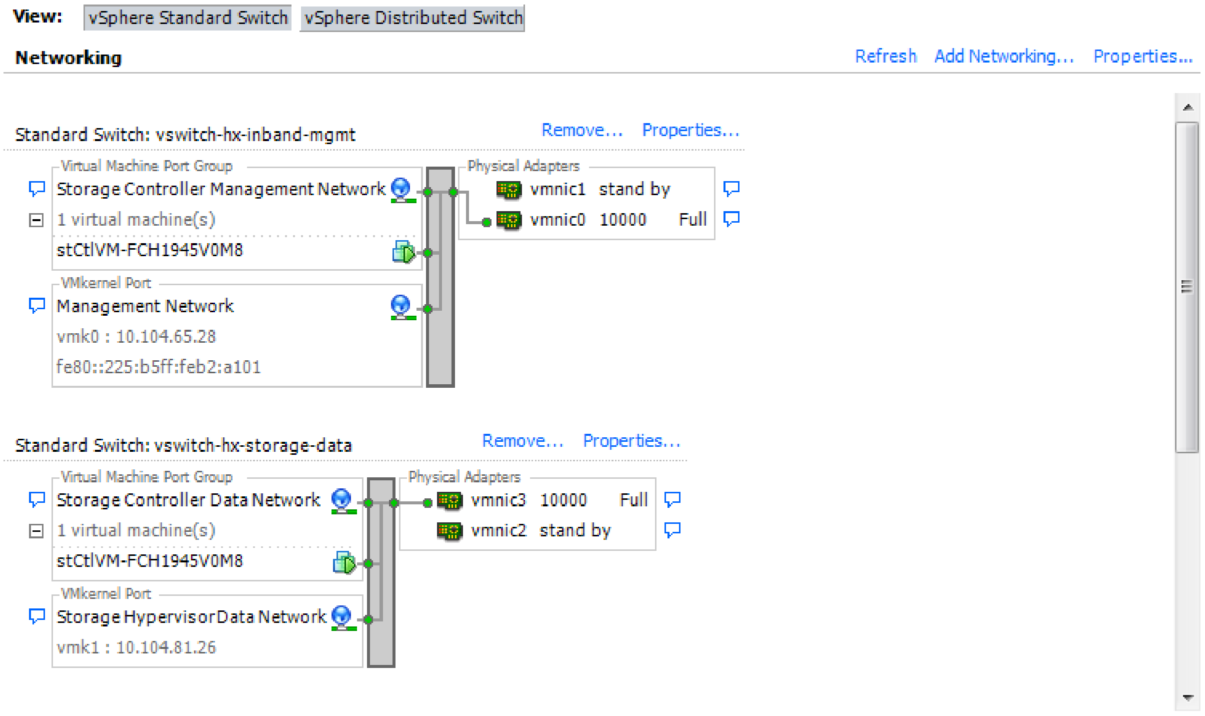
Physical Adapters
vmnic7 stand by
vmnic6 10000 Full

Standard Switch: vswitch-hx-vm-network

Virtual Machine Port Group
VM Network
4 virtual machine(s) | VLAN ID: 212
veeamrepo
Win2k12-Nessus
perf-ova-2-0
CLIQR-release-worker1

Physical Adapters
vmnic5 10000 Full
vmnic4 10000 Full

Figure 11. Screen showing management and Cisco HyperFlex storage data networks



Standard Switch: vswitch-hx-inband-mgmt

Virtual Machine Port Group
Storage Controller Management Network
1 virtual machine(s)
stCtlVM-FCH1945V0M8

VMkernel Port
Management Network
vmk0 : 10.104.65.28
fe80::225:b5ff:feb2:a101

Physical Adapters
vmnic1 stand by
vmnic0 10000 Full

Standard Switch: vswitch-hx-storage-data

Virtual Machine Port Group
Storage Controller Data Network
1 virtual machine(s)
stCtlVM-FCH1945V0M8

VMkernel Port
Storage Hypervisor Data Network
vmk1 : 10.104.81.26

Physical Adapters
vmnic3 10000 Full
vmnic2 stand by

For an in-depth discussion of virtual distributed switches (VDS) with Cisco HyperFlex systems, see the following resource:

http://www.cisco.com/c/dam/en/us/products/collateral/hyperconverged-infrastructure/whitepaper-c11-737724.pdf

Datastore Best Practices

Since the cluster is created with a set of affinity rules that define preferential resources per site, it is recommended to create two datastores. In HX Connect the datastore creation wizard allows (requires) the specification of a site affinity for the storage resource. Since reads are local to the site for VMs in a particular datastore, it is a best practice to create at least one datastore per site. When deploying a VM to a specific compute resource (site), the VM should be deployed in the corresponding datastore to maximize read efficiency.

CVM Best Practices

In an HX Stretch Cluster environment, the installer will deploy and configure the control VMs (CVMs) appropriately. There are no special settings that need to be made after the deployment is active. The CVM does not participate in the host group affinity settings and has no VM rules applied to it. HA and DRS will not function against the CVM since it is in its own local datastore on the node and utilizes PCI passthrough, so it has hardware dependence on the physical node.

Troubleshooting

In cases where a SC installation has not completed successfully, the usual cause is environmental. Double check the following:

- All component versions match the required builds as documented in the pre-installation checklist and the release notes
- Firewalls are not blocking essential ports as listed in the pre-installation checklist.
- Jumbo frames are enabled for the data storage Stretch VLAN
- The management VLAN is Stretch between sites
- The witness is deployed and reachable
- The witness is running the correct version for the build of HXDP you have deployed
    - To determine a running witness version, run the following from the witness CLI:
        - vmtoolsd --cmd "info-get guestinfo.hxwitness.version"
    -
- vCenter is deployed and reachable

In the event that a cluster build has failed and you need to restart, there are two primary rebuild modes: a cluster deployment-only rebuild or a complete rebuild. Open a TAC case if you are unsure about performing any of these operations.

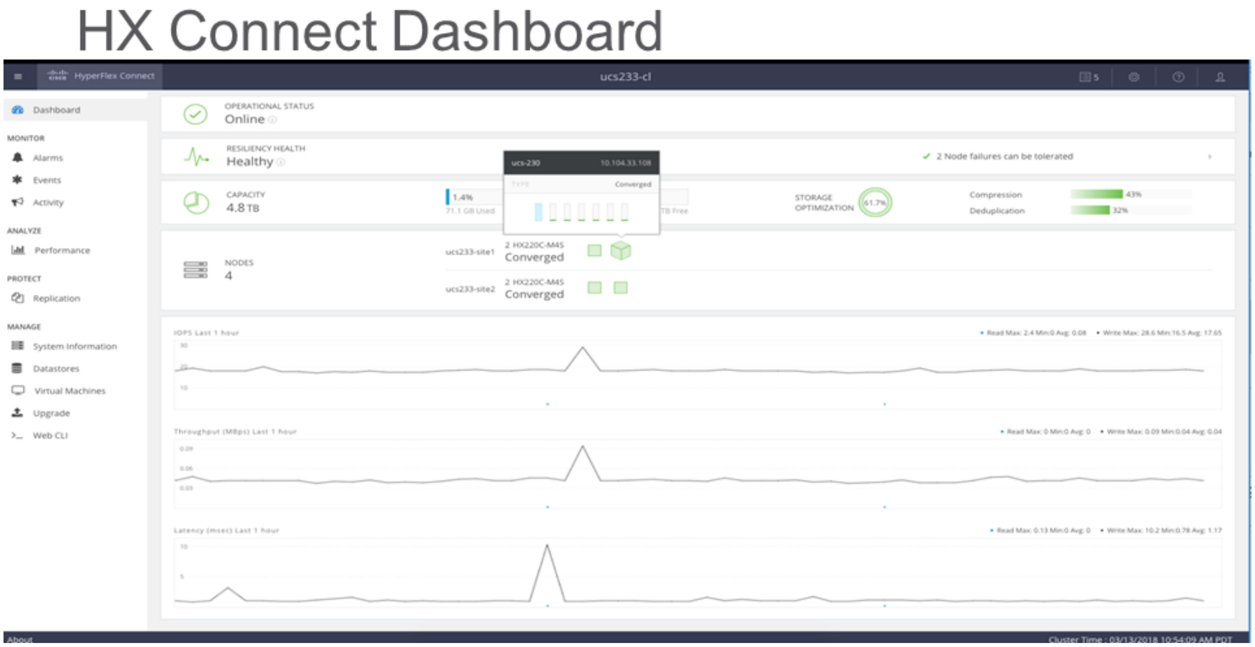
1. Restart the cluster deployment portion only
    - Be sure to clean the witness by running the cleanup script at /opt/springpath/cleanup.sh on the witness itself
    - Clean up the cluster nodes using the support tools to destroy and remove any partial cluster creations. Contact TAC for a procedure.
    - Re-run the cluster installer
2. Restart from scratch using the ESXi install image (downloadable from CCO) and the HX Installer.
    - Be sure to clean the witness by running the cleanup script at /opt/springpath/cleanup.sh on the witness itself
    - Be sure to clean up UCSM by removing the appropriate service profiles
    - Mount the ESXi image in UCSM on each node via KVM virtual storage and reinstall ESXi.
    - Re-run the cluster installer

Stretch cluster operations

After your cluster is successfully installed, you are ready to create data stores and deploy virtual machines. Cisco HyperFlex Connect is the HTML 5–based user interface native to the cluster. You access It either by clicking the button on the cluster creation summary screen after your installation is complete, or (usually) by entering the cluster management IP address (CIP-M) in a browser and logging in using the vCenter SSL administrative account or the local root account.

After you are logged in, you will be presented with an overview of the cluster status and performance in the Cisco HyperFlex Connect dashboard (Figure 12). From here you can view the node count and type, the overall space savings from deduplication and compression, performance at-a-glance (I/O operations per second [IOPs], throughput, and latency), and the site-based resiliency status using the arrows next to the health status.

Figure 12. View of the Cisco hyperflex connect dashboard



HX Connect Dashboard

OPERATIONAL STATUS Online

RESILIENCY HEALTH Healthy

✓ 2 Node failures can be tolerated

ucs-230 10.104.33.108

TYPE Converged

CAPACITY 4.8 TB

1.4% 71.1 GB Used TB Free

STORAGE OPTIMIZATION 61.7%

Compression 43%

Deduplication 32%

NODES 4

ucs233-site1 2 HX220C-M4S Converged

ucs233-site2 2 HX220C-M4S Converged

IOPS Last 1 hour • Read Max: 2.4 Min:0 Avg: 0.08 • Write Max: 28.6 Min:16.5 Avg: 17.65

Throughput (MBps) Last 1 hour • Read Max: 0 Min:0 Avg: 0 • Write Max: 0.09 Min:0.04 Avg: 0.04

Latency (msec) Last 1 hour • Read Max: 0.13 Min:0 Avg: 0 • Write Max: 10.2 Min:0.78 Avg: 1.17

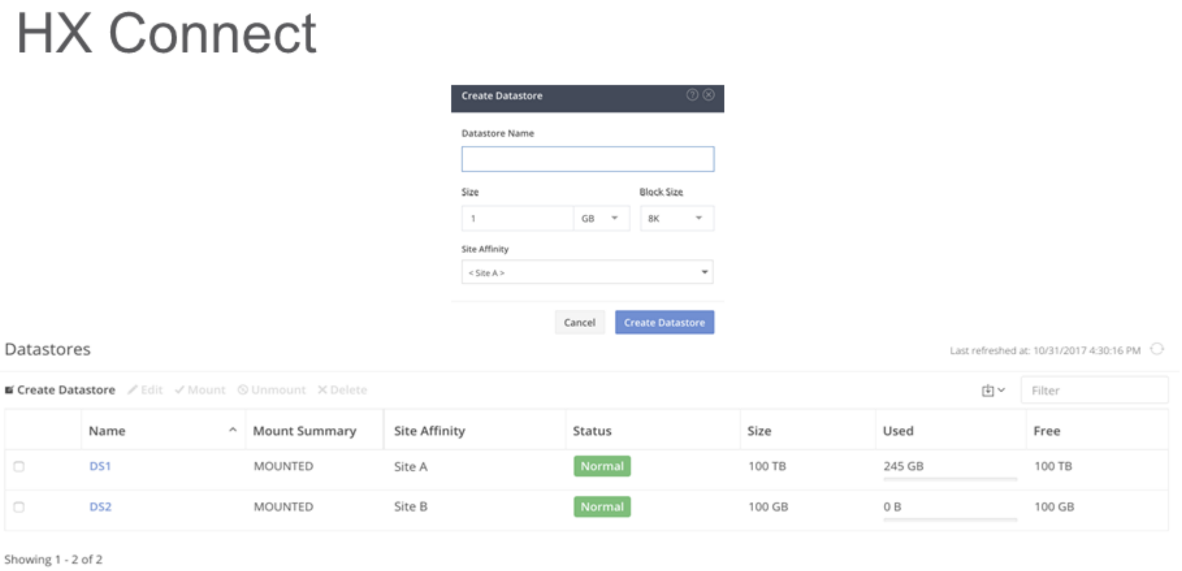
Cluster Time : 03/13/2018 10:54:09 AM PDT

You should use Cisco HyperFlex Connect for most, if not all, cluster management activities. In particular, be sure to use Cisco HyperFlex Connect to create data stores. Doing so helps ensure that site affinity is set appropriately for all data stores created. When you create a data store:

- Data store affinity will be set.
- Data stores created will be mounted on all nodes on both sites.
- Virtual machines will start on the appropriate site.

Figure 13 shows the data store creation wizard from Cisco HyperFlex Connect. Notice the site affinity setting.

Figure 13. Data store creation and site affinity



HX Connect

Create Datastore

Datastore Name:

Size: 1 GB

Block Size: 8K

Site Affinity: < Site A >

	Name	Mount Summary	Site Affinity	Status	Size	Used	Free
□	DS1	MOUNTED	Site A	Normal	100 TB	245 GB	100 TB
□	DS2	MOUNTED	Site B	Normal	100 GB	0 B	100 GB

Showing 1 - 2 of 2

Deploying a virtual machine in a Stretch cluster is no different than deploying one in a regular cluster using vCenter.

You can also use the Cisco HyperFlex HTML 5 vCenter Plugin for cluster management. As of HX 4.5.1a, this plugin interface supports Stretch cluster and offers an integrated single management portal through vCenter for daily cluster and VM operations. The plugin has the following benefits:

Cisco HyperFlex HTML5 Plugin for VMware vCenter—Provides users the ability to manage and monitor your HyperFlex clusters from the VMware vCenter Web UI. Additional functionally in version 2.1.0 includes:

- Snapshot Scheduler
- iSCSI Management
- Nodes and Disk View
- Virtual Machines Summary
- Events and Tasks
- VLAN Creation
- Rename Cluster
- HyperFlex Stretch Clusters
- Support for VMware vCenter Linked Mode

Shutting a Stretch Cluster Site down gracefully

There may be a need to shut a site down (maintenance, site move, site work, failure testing etc.). If you have removed the VMware EAM dependency from your cluster (default in 4.0.2b and above), then you can use HX maintenance mode in vCenter or through HX Connect to shut down the control VMs on each node in a site. When this occurs, the site will shut down gracefully and the guest VMs will failover to the "surviving" site. You can then also power down ESXi on the nodes if you need to. If EAM is enabled on your CVMs, see the appendix for a more comprehensive shutdown procedure.

Stretch cluster upgrades

Prior to HX 4.0.2a, upgrading a Stretch cluster is a more manual process. Using HX Connect you can upgrade only HXDP software itself, however this often requires an accompanying upgrade of ESXi and/or UCSM. In order to upgrade UCSM, follow the documentation for UCS firmware upgrades. Regardless of version, upgrade ordering should be done in the following order in general:

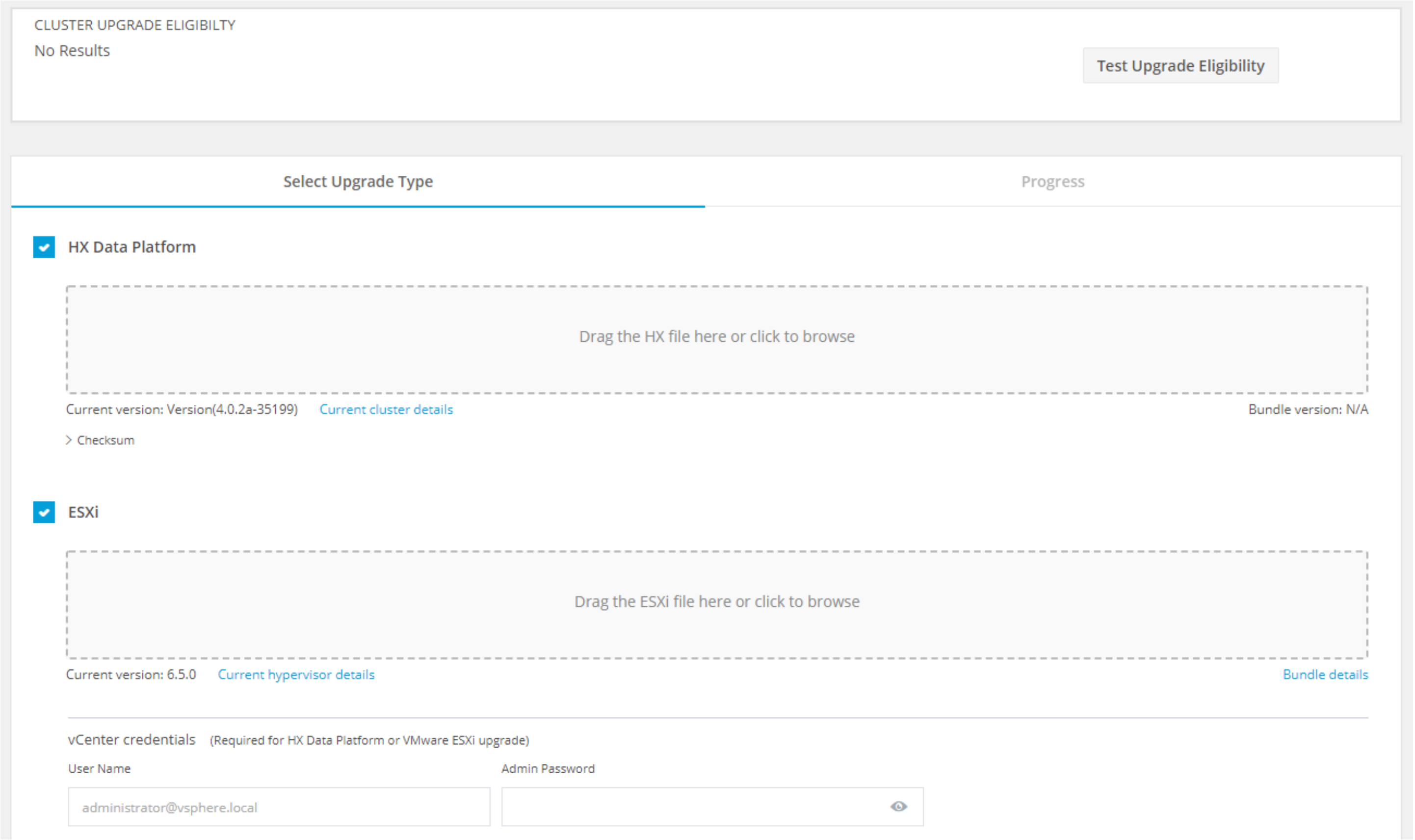
- Upgrade Cisco UCS Infrastructure
- Upgrade Cisco UCS firmware
- Upgrade Cisco HX Data Platform
- Upgrade Cisco customized VMware ESXi

You only need to upgrade the firmware if there is a compelling reason to do so and your current version meets the compatibility requirements of the version of HX and ESXi that you are running or plan to run. Note that upgrading UCSM on each pair of FIs should be conducted first (the infrastructure upgrade). Upgrade only one site at a time. Once UCSM is upgraded, you can proceed with the rolling node firmware upgrade. These are the B and C packages that you download from Cisco. Keep in mind that these upgrades, performed via UCSM, require monitoring since there are 2 acknowledgement steps in the upgrade that require user intervention. Upgrade the firmware on only one site at a time. Once this is complete, you can proceed with the HX and/or ESXi upgrades as needed.

To upgrade ESXi in these environments, the following steps should be conducted:

- Download the ESXi upgrade packages for your desired version from Cisco CCO
- Stage these upgrade packages on each ESXi node in an appropriate location (e.g., /tmp)
- Perform the upgrade one node at a time, while the node is in HX maintenance mode
- Execute the following (sample) to verify the sources:
    - esxcli software sources profile list -d /vmfs/volumes/ISO/HX-ESXi-6.5U3-15256549-Cisco-Custom-6.5.3.6-upgrade-bundle.zip
- Execute the following to update the ESXi version for that node:
    - esxcli software profile update -d /vmfs/volumes/ISO/HX-ESXi-6.5U3-15256549-Cisco-Custom-6.5.3.6-upgrade-bundle.zip -p HX-ESXi-6.5U3-15256549-Cisco-Custom-6.5.3.6
- Exit the node from HX maintenance mode
- Verify version with vmware -v1 from esxcli
- Continue to the next node

For versions subsequent to 4.0.2a, the ESXi upgrade component is part of the HX Connect upgrade workflow. The screen below shows this additional functionality in the upgrade section of HX Connect.


Test Upgrade Eligibility
Select Upgrade Type
Progress
Drag the HX file here or click to browse
Current cluster details
> Checksum
Drag the ESXi file here or click to browse
Current hypervisor details
Bundle details

CLUSTER UPGRADE ELIGIBILTY
No Results

☑ HX Data Platform

Current version: Version(4.0.2a-35199)

Bundle version: N/A

☑ ESXi

Current version: 6.5.0

vCenter credentials (Required for HX Data Platform or VMware ESXi upgrade)

User Name: administrator@vsphere.local

Admin Password

Notice the “Test Validation” button in the top right of the upgrade screen. Pre-upgrade validations are also now part of the upgrade workflow. The system will verify that all components for an HXDP upgrade are in place and correct. It may be that you do not need to update components other than just HXDP. This workflow will verify that condition.

As of HX 4.5.x, upgrade of Stretch Cluster is supported using Intersight (IS). Once the cluster is claimed in IS, you can proceed with a cluster upgrade using the cloud interface presented there.

Stretch cluster failure modes

For Arbitrator deployment scenarios and behaviors, see the Witness Configuration – Intersight Arbitrator section above.

One of the main precautions required for using a Stretch or metropolitan (multisite) single cluster is the need to avoid a split-brain scenario. A split-brain condition indicates data or availability inconsistencies originating from the maintenance of two separate data sets with overlap in scope, either because of the loss of a site or a failure condition based on servers not communicating and synchronizing their data with each other (site link loss). The witness exists to prevent this scenario, and it is discussed in the various failure modes presented here.

Because a Stretch cluster is a single cluster, for most failure situations you can simply ask yourself: How would a single cluster with a replication factor of 2 behave here? It is when you experience site losses (or more than two simultaneous node failures on a single site) that the behavior diverges from that of the single-location RF 2 cluster because you actually have the advantage of an effective RF4.

To appreciate the failover mechanics of a Stretch cluster, take a closer look at ZooKeeper. Architecturally, a Stretch cluster contains five instances of ZooKeeper: two at each site and one on the witness server. So, in total there is one master ZK node and 4 followers. Only a storage node can be

ZK nodes. Compute-only nodes will never be created with a ZK instance. The function of ZooKeeper is to maintain the cluster membership and a consistent cluster-wide file system configuration. So, if there are eight nodes at each site (a 16-node cluster), there will still be two ZooKeeper instances running on two nodes at each site and one more on the witness server.

Whenever a failure occurs, at least three ZooKeeper instances must be present to re-create the cluster membership and help ensure a consistent file system configuration. ZooKeeper achieves this behavior by using its built-in voting algorithm (based on the well-known Paxos algorithm).

If the witness goes down, then one ZooKeeper instance is lost. However, four more ZooKeeper instances are still running (2 at each site), which is more than the minimum of three ZooKeeper instances needed. Hence, the cluster will not be affected (no virtual machine failover or internal I/O hand-off occurs).

If a site goes offline, two ZooKeeper instances will go down. However, three more ZooKeeper instances are still running, which again is greater than or equal to the minimum of three ZooKeeper instances required. Hence, the cluster will not be affected. Virtual machines will automatically failover to the surviving site because of the presence of VMware HA. This failure will be treated as if half (minus one since the witness is still online) the number of nodes are lost in a single cluster.

If a ZK node at a site goes down that was hosting the ZooKeeper master, the ZooKeeper algorithm will elect another ZK node to be promoted to ZooKeeper master. The promotion of another ZK node happens only if the failed node is a master ZK node and the failover target is part of the ZK ensemble (which is always the case for a ZK master). If the failed node was a ZK follower (i.e., a stand-alone ZK node), then no election occurs, and you are running with one less ZK instance. Either way, four more ZooKeeper instances are still running, which is more than the minimum of three ZooKeeper instances required. The site and the cluster will still be online. Only the affected virtual machines will be restarted on the surviving nodes at the same site (with Stretch cluster DRS rules managing the movement). The failure will be treated like a node lost in a single cluster.

It is expected that you will recover or replace any failed nodes. New ZK nodes will not be automatically created. If those nodes happen to be ZK nodes, then there is a manual process to reassign ZK membership if the node needs to be completely replaced. See your support representative for assistance. A recovered node will simply resume its previous role unless it was master (since a new master is now elected) in which case it will join as a follower.

Survivability while maintaining online status after node losses requires a majority zookeeper quorum and more than 50% of any nodes (the witness counts as an active zookeeper node). If one site has suffered multiple losses, it is possible that the surviving site could tolerate a node or disk loss (in a cluster greater than 2+2) if that node is not a zookeeper node, but it is not guaranteed. See the scenario walk throughs below.

Zookeeper has a notable dependence on NTP from the nodes to maintain cluster synchronization. The allowable ZK time drift between nodes is 300ms. If the skew exceeds this the cluster is subject to ZK errors and may not function properly. It is advisable to monitor NTP skew between CVMs using the HX APIs and alert on time drift issues.

Recovery of ZK Nodes That Have Failed

As mentioned above, there are two types of ZK nodes: a single ensemble master and 4 followers. Zookeeper nodes that fail have to undergo a special process to be replaced. If they are recovered, then they resume their previous role unless they were the master since a new master is now in place. This node will become an ensemble follower.

- If the master ZK node fails, zookeeper will automatically elect a new master from the remaining ZK follower nodes. This will leave you with a cluster having 4 ZK instances.
- If a follower fails, a new follower is not created on rebuild. If the node that failed is recovered, your ZK instance will return, and you will be back to normal (5 ZK instances). If you are unable to recover the node, the manual node-remove workflow and node replacement will result in a new follower being created. Contact support for assistance.

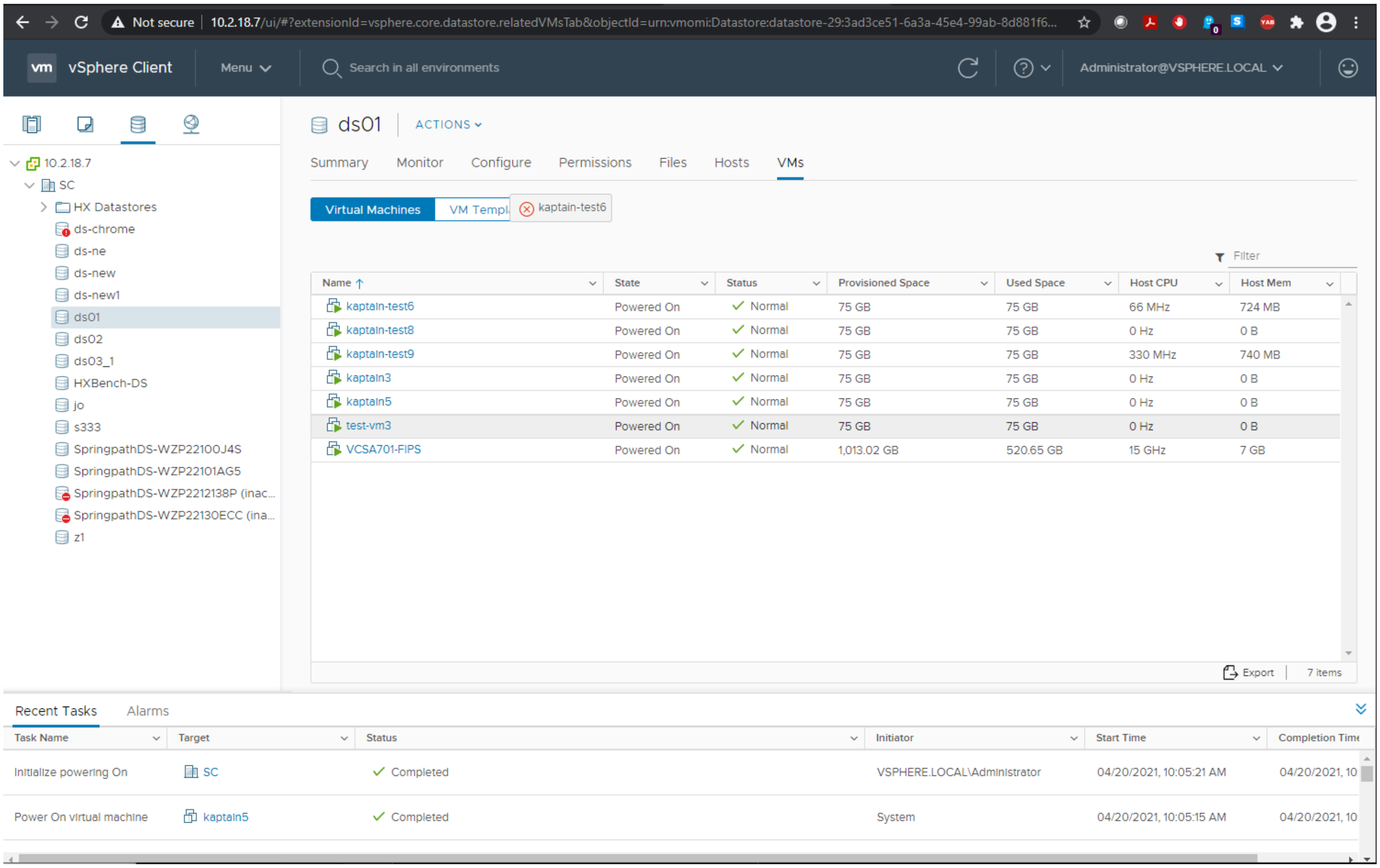
Types of failures

The typical timeout value for a node or site is about 17 seconds. Failure timing is dependent on zookeeper quorum reporting and cluster resource manager (CRM) mapping updates. This timeout determines when a node is considered offline or when a site is considered unavailable via the data center interconnect (DCI) for the storage and management networks. VMs are restarted on surviving nodes in a site or on the surviving site after this timeout value.

The types of failures and the responses to each are summarized here:

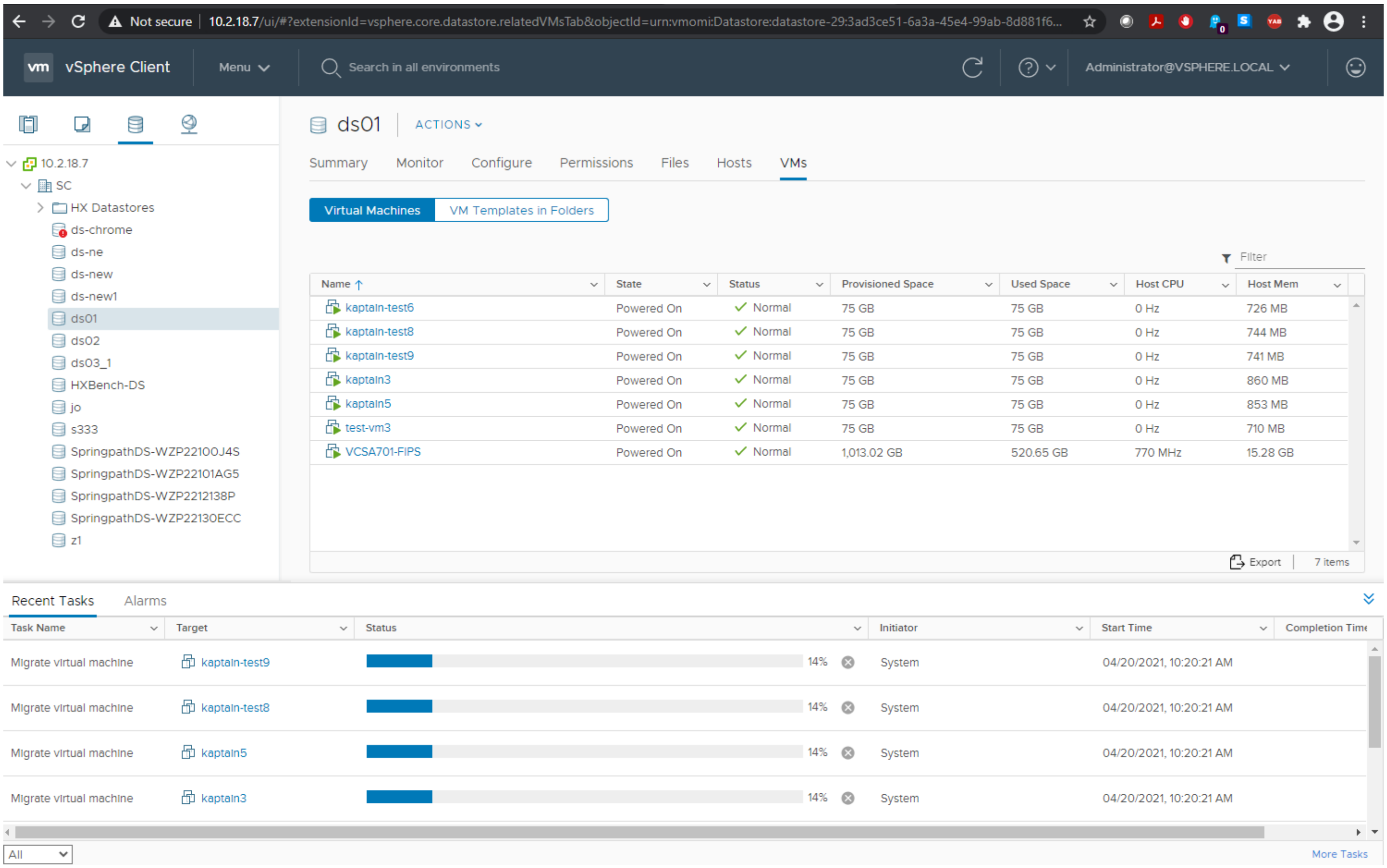
- Disk loss
    - Cache disk: This failure is treated the same way as in a normal cluster. Other cache disks in the site service requests, and overall cache capacity is reduced until the failed component is replaced.
    - Persistent disk: This failure is treated the same way as in a normal cluster. After a 2-minute timeout interval, the data from the failed disk is rebuilt using the remaining capacity.
    - System disk: This failure results in node failure on nodes that do not have redundant RAID 1 system disks, otherwise it is tolerated without incident.
- Node loss
    - 1x: The site will rebuild the failed node after a 2-hour timeout or earlier through manual intervention.
    - Nx: If the node losses are simultaneous and are not all the nodes in a site (e.g., lose 3 nodes out of 5 on a site), the site will remain online, and site failover will not occur. For example, if you have a 3+3 cluster, and you lose 2 nodes on site 1 (regardless of ZK type), then site 1 will still be active, VMs will migrate to the surviving node and the site will still function. There may not be enough resources on the surviving 1 node to restart all the VMs from the 2 failed nodes on the site. In that case, since the host affinity rules are “should run” and not “must run”, DRS will restart the VMs that exceed the site capacity at the other site.
- Fabric interconnect loss
    - 1x: The redundant fabric interconnect at the site will handle data until its partner is recovered.
    - 2x: The site will be offline, and site failover will occur.
- Witness loss
    - Nothing happens; the cluster is not affected. Bring the witness back online after it is repaired.
    - Since the Witness is a ZK node, you are guaranteed to survive one node failure at either site (since in worst case that failed node will be ZK as well). This leaves 3 ZK’s left. You cannot be guaranteed another failure because if that is a ZK node as well, you no longer have majority ZK surviving. You ***may*** survive these failures if you get lucky with no additional ZK node failures, but you are not guaranteed this condition. Only worst-case survivability is reported.
- Accidental deletion of the witness virtual machine
    - Restore from backup with identical networking. The cluster will discover the witness and resynchronize.
    - Contact Cisco TAC for a recovery process otherwise.
- Switch loss (single site)
    - 1x: For redundant switches at a site, the partner switch will handle data until the failure is repaired. If there is a single uplink switch per site, site failover will occur.
    - 2x: The site will be offline, and site failover will occur.
- Site loss
    - The site will be offline, and site failover will occur.
- Site link loss
    - For a scenario in which a fault occurs in the network between the two sites (a cable is damaged, a network port on either site fails, etc.) but the nodes on the two sites are still alive, the following process is implemented:

1. When a Stretch cluster is created, one site is biased to establish a ZooKeeper master. This is done by assigning a higher node ID. For the purpose of this discussion, the quorum site is site A.
2. When the network disconnect occurs, the witness and the nodes of the site that have the ZooKeeper master form the quorum.
3. The nodes at the other site (site B) will still stay powered on, and I/O operations from the local IO Visor instance from this site (site B) will not be able to perform write I/O operations successfully, which this will guarantee the isolated site's consistency. The stcli cluster-info command will show these nodes as unavailable in the cluster, even though physically they may be powered on.
4. Because site A is the ZooKeeper quorum site, the updates to ZooKeeper will eventually (after a failure-detection timeout) be visible to site B. Eventually, the IO Visor on ESX at site B will see that it needs to talk to a different node, which is the actual I/O primary node (which is in the ZooKeeper quorum at site A). Because there is no network connection, site B will keep retrying those I/O operations and will eventually see "All Paths Down" (APD), assuming that there are still user virtual machines on this site (site B). Your intervention should verify that eventually no virtual machines remain on this site (because they have been failed over to other ESX hosts).
5. Virtual machines fail over to the site having the ZooKeeper leader. VMware HA and DRS are responsible for the failover of virtual machines.
6. If the network is restored, the nodes of site B that were fenced out will become available again in the cluster. Automatic resynchronization between the sites should occur. Virtual machine failback is fully automatic. The screens below show (in order) failing a site with automatic restart, followed by automatic failback on site restoration.



Name ↑	State	Status	Provisioned Space	Used Space	Host CPU	Host Mem
kaptain-test6	Powered On	✓ Normal	75 GB	75 GB	66 MHz	724 MB
kaptain-test8	Powered On	✓ Normal	75 GB	75 GB	0 Hz	0 B
kaptain-test9	Powered On	✓ Normal	75 GB	75 GB	330 MHz	740 MB
kaptain3	Powered On	✓ Normal	75 GB	75 GB	0 Hz	0 B
kaptain5	Powered On	✓ Normal	75 GB	75 GB	0 Hz	0 B
test-vm3	Powered On	✓ Normal	75 GB	75 GB	0 Hz	0 B
VCSA701-FIPS	Powered On	✓ Normal	1,013.02 GB	520.65 GB	15 GHz	7 GB

Task Name	Target	Status	Initiator	Start Time	Completion Time
Initialize powering On	SC	✓ Completed	VSPHERE.LOCAL\Administrator	04/20/2021, 10:05:21 AM	04/20/2021, 10
Power On virtual machine	kaptain5	✓ Completed	System	04/20/2021, 10:05:15 AM	04/20/2021, 10



Name ↑	State	Status	Provisioned Space	Used Space	Host CPU	Host Mem
kaptain-test6	Powered On	✓ Normal	75 GB	75 GB	0 Hz	726 MB
kaptain-test8	Powered On	✓ Normal	75 GB	75 GB	0 Hz	744 MB
kaptain-test9	Powered On	✓ Normal	75 GB	75 GB	0 Hz	741 MB
kaptain3	Powered On	✓ Normal	75 GB	75 GB	0 Hz	860 MB
kaptain5	Powered On	✓ Normal	75 GB	75 GB	0 Hz	853 MB
test-vm3	Powered On	✓ Normal	75 GB	75 GB	0 Hz	710 MB
VCSA701-FIPS	Powered On	✓ Normal	1,013.02 GB	520.65 GB	770 MHz	15.28 GB

Task Name	Target	Status	Initiator	Start Time	Completion Time
Migrate virtual machine	kaptain-test9	14%	System	04/20/2021, 10:20:21 AM	
Migrate virtual machine	kaptain-test8	14%	System	04/20/2021, 10:20:21 AM	
Migrate virtual machine	kaptain5	14%	System	04/20/2021, 10:20:21 AM	
Migrate virtual machine	kaptain3	14%	System	04/20/2021, 10:20:21 AM	

Scenario Walk Through - Failure of Multiple Nodes in a Site

If you have an 8-node cluster (4 nodes on each site) and you lose 2 nodes on site A, what happens to the remaining nodes? Do the VMs still continue to run on site A's remaining nodes?

In this scenario, the remaining nodes on site A restart the VMs that were running on the failed nodes. The site is still online and serving data, however, since the site is locally RF2 (globally RF4) you will have lost some part of the distributed local primary write logs that were running on the 2 nodes that failed. You will also have lost some local persistent data. HX will recognize this and switch the primaries over to Site B. This will incur a read penalty for these VMs. Note also, depending on how heavily loaded the system is, the surviving nodes on Site A may be either at capacity or unable to restart all of the VMs. In that case, DRS can ignore the affinity rules (HX uses "should" rules for affinity not "must" rules) and restart the VMs on available resources in Site B. If there is capacity in Site A, rebuild will begin and attempt to reestablish local RF2.

This behavior is the same if you have a 10+10 cluster and lose, say, 5 nodes on Site A.

If you were to suffer multiple node failure at each site simultaneously, that would constitute a catastrophic loss and your cluster would be offline pending recovery.

Scenario Walk Through - Failure of a Site

In the event of site failover, operations should continue as intended after the virtual machines from the failed site boot on the surviving site. Virtual machine and IO Visor behavior is as described for site link loss in the preceding discussion. For example:

If you have an 8-node cluster (4 nodes on each site) and you lose site A, vCenter HA initiates a restart of all of Site A's VMs on nodes in site B. Failback is automatic due to site affinity. If sized at maximum capacity (50% per site) then Site B is now running at maximum capacity (100%).

After your downed site has been recovered and communications with the remaining site and the witness have been reestablished, you can move your virtual machines back to their original compute resource (based on site affinity) at the recovered site. Use vMotion for this process so that affinity and proper IO Visor routing occurs after the virtual machines are moved back to their preferred locations. Storage vMotion is not required, since the datastore is mounted on all nodes. Only a migration of the compute resource is needed to re-establish site storage affinity and compute resource parity. In short, during a site failover the affinity rule will not change so that you can quickly migrate back once the impacted site is recovered. However, if you manually Storage vMotion (SVMotion) VMs around outside of a failover event, the affinity rules will automatically be updated to reflect residence in the datastore with the correct rules.

Failure response summary

Table 1 summarizes the failure modes discussed previously, with some additional information for particular situations. Note that double, separate catastrophic failures are not considered here (for example, both site loss and witness loss) because such failures always result in a cluster offline status.

Table 1. Failure responses

Component failure	Cluster behavior	Quorum update	Virtual machine restart	Site status	Cluster status
Single site single cache disk	Site is online, with diminished cache capacity.	No	No	Online	Online
Single site single persistent disk	Site is online, with diminished capacity, and is rebuilt after 2 minutes using the remaining capacity.	No	No	Online	Online
Single site double cache disk	Site is online, with diminished cache capacity.	No	No	Online	Online
Single site double persistent disk	If failure is simultaneous and on different nodes. The site is still online but some VMs will switch primary write logs to the opposite site. If the failure is not simultaneous on different nodes at different times, then the cluster behaves as with a single-disk failure with reduced capacity. If the failure is simultaneous on the same node.	No No No	No No No	Online Online Online	Online Online Online
Single site single node loss	Node is rebuilt after 2 hours or through manual intervention.	Maybe	Yes	Online	Online
Single site multiple node loss	Site is online and will restart VMs on surviving nodes in the site.	Maybe	Yes	Online	Online
Single site single fabric interconnect loss	No impact on the site; recover the fabric interconnect.	No	No	Online	Online
Single site double fabric interconnect loss	Site is offline.	Yes	Yes	Offline	Online
Double site single fabric interconnect loss	No impact on the site; recover the fabric interconnects.	No	No	Online	Online
Double site double fabric interconnect loss	Both sites are offline.	No	No	Offline	Offline
Witness loss	No impact on the site; recover the witness.	No	No	Online	Online
Single site single switch loss	If redundant switching exists at the site, there is no impact; recover the switch. If the site has only a single switch, site is offline.	No Yes	No Yes	Online Offline	Online
Single site double switch loss	Site is offline.	Yes	Yes	Offline	Online
Double site single switch loss	If redundant switching exists at the sites, there is no impact; recover the switches. If the sites have only a single switch, the sites are offline.	No No	No No	Online Offline	Online Offline
Double site double switch loss	Both sites are offline.	No	No	Offline	Offline
Site loss	ZooKeeper instance maintains information about cluster groups and forms the quorum. When a site is lost, ZooKeeper communications disappear, site fencing is enforced, and the cluster quorum is redefined. ZooKeeper with DRS rules (affinity, groups, etc.) makes sure that the same virtual machine is never running on both sites simultaneously.	Yes	Yes	Offline	Online

Site link loss – sites still online and witness is reachable, but site-to-site link is down	In this case, if both sites are still online and they can communicate with witness, the site with the master zk node will become primary and VMs from the second site will failover. Replicas will sync up once the link is re-established. Failback is automatic.	No	Yes	Online	Online
Site loss and disk or node loss on the surviving site in a 2+2 node cluster	In the special case of a 2+2 cluster, the surviving site will go offline due to a zookeeper loss and a loss of quorum if a node is lost on the surviving site. A disk loss can be tolerated at the surviving site if it is a cache, persistent, or RAID enabled system disk. If it is a system disk on a system without RAID, then the node will go offline and the cluster will be down due to loss of a 3rd zk node.	No	No	Depends – see note	Depends – see note
Site loss and disk or node loss on the surviving site in a n+n node cluster	This situation is identical to the previous case, however, it is possible that the site may survive. This will depend on which node or disk fails. The failure must not occur on a zookeeper node. Site survivability is NOT guaranteed.	No	No	Offline (possibly online but not guaranteed)	Offline (possibly online but not guaranteed)
Site Loss and witness loss (connectivity, VM failure etc.)	If a site is lost and the witness is lost, regardless of mechanism (link failure, VM failure, witness hypervisor failure etc.) the cluster will go offline since it is now below 50% ZK nodes (<3).	No	No	Offline	Offline

Witness Failure and Restore from Backup

To increase the resiliency of a Stretch cluster deployment, many users will back up their witness VM. If the witness were to fail, you can restore from back up (retaining identical network settings). The witness ZK instance will be stale but will re-synchronize with the surviving site(s). As mentioned in the failure scenario section, in the event that the witness fails after a site goes offline and the cluster fails over, the system will be offline. If the witness is subsequently restored from backup and communication is available between witness and either site, the cluster will synchronize zookeeper and come back online.

It is also possible in this scenario to maintain a cold witness stand-by VM and promote it when needed. In order to properly integrate with the cluster, it will need to be an identical copy of the original witness and will have to retain the same networking. It will synchronize with the cluster when brought online.

Failure Response Times

Failure response times for disk loss are near-instantaneous. It is the same for active/passive standby links to the FIs. Node failures in a site are they typical node timeout values (approximately 17 seconds). For a site failover to occur, the timeout must happen for multiple nodes at the same time. Since connectivity loss to a site typically involves multiple simultaneous losses, the site time out is at best the same as a node time out. This value can increase due to other factors, such as heavy workloads (affecting ZK updates), latency to the witness, and inter-site link latency.

Failure Capacity Considerations and Example

Failure of a node in a site reduces the overall capacity by the free space on the lost node symmetrically; that is, times two. In other words, the cluster capacity in general equals twice the minimum site cluster capacity at the cluster RF (4). This can be expressed as 2[min(site A capacity, site B capacity)]/4.

Take the following example for capacity after node loss on a 5+5 Stretch cluster:

siteA/siteB

5 nodes/5 nodes = 10 nodes

At 10TB/node useable that is 100 TB usable total

50TB usable/ 50 TB usable per site

RF 2+2 usable / RF 2+2 usable data protection (RF 4 equivalent)

25TB usable after data protection (100 TB/4)

50 TB/4=12.5TB per site usable.

If site A loses a node, it has dropped by 20% capacity from 12.5 TB to 10 TB. Since the total cluster usable capacity is defined as 2[min(site A capacity, site B capacity)]/4. Site A is now the cluster site minimum at 10 TB (site B is still at 12.5TB) which means the total cluster capacity is now 2(10TB)/4 = 20 TB usable.

In the event of a site failure, the surviving site capacity is the total cluster capacity divided by two, however the remaining capacity is filled in a RF 2 manner until the failed site is recovered so the total free capacity remains constant before and after a site failure. Once the failed site is recovered changes made since the failure are synchronized across the cluster and RF 4 is re-established. Capacity reporting during this transitional interval (site loss, surviving site stabilization, surviving site production usage, failed site recovery) is in flux. The reported capacity will be variable as things like the file system cleaner, rebuilds, and transitions to temporary RF 2 for all VMs take place. In the steady state where the failed site is not recovered in a timely fashion, the surviving capacity will approach RF 2 for the free capacity that was available on the surviving site before the secondary site failure occurred.

For more information

For additional information, see the following resources:

- Cisco HyperFlex 3.0 with VSI:
  https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/UCS_CVDs/HyperFlex_30_vsi_esxi.html#_Toc514225504
- Cisco Hyperflex 3.5 with Multipod ACI
  https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/UCS_CVDs/hx_35_vsi_aci_multipod_design.html
- Cisco HyperFlex 4.0 Stretched Cluster with Cisco ACI 4.2 Multi-Pod Fabric Design Guide
  https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/UCS_CVDs/hx_40_vsi_aci_multipod_design.html
- Cisco HyperFlex 4.5.1a with VXLAN CVD
  https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/UCS_CVDs/hyperflex_4_vsphere6_7_vxlan.html
- Cisco HyperFlex sizer:
  https://HyperFlexsizer.cloudapps.cisco.com/ui/index.html#/scenario
- Cisco HyperFlex preinstallation checklist:
  http://www.cisco.com/c/dam/en/us/td/docs/hyperconverged_systems/HyperFlex_HX_DataPlatformSoftware/HyperFlex_preinstall_checklist/Cisco_HX_Data_Platform_Preinstallation_Checklist_form.pdf
- Cisco HyperFlex release notes:
  https://www.cisco.com/c/en/us/support/hyperconverged-systems/HyperFlex-hx-data-platform-software/products-release-notes-list.html
- Cisco HyperFlex with VDSs:
  http://www.cisco.com/c/dam/en/us/products/collateral/hyperconverged-infrastructure/whitepaper-c11-737724.pdf
- ZooKeeper:
  https://en.wikipedia.org/wiki/Apache_ZooKeeper
- Cisco HyperFlex with disjointed Layer 2 networking:
  https://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/unified-computing/white_paper_c11-692008.html
- vNICS and vHBAs in Cisco UCS:

https://supportforums.cisco.com/document/29931/what-concept-behind-vnic-and-vhba-ucs

- Intersight Containerized Arbitration:

  https://intersight.com/help/saas/resources/deploy_local_container__hyperflex_witness_servers

APPENDIX A – Shutting down a site

This procedure is valid for both EAM enabled CVMs and newer systems with EAM disabled. If you just need to shut the site down and not move it and EAM is disabled, then you do not need to do this. See the section on Gracefully Shutting Down a Site. This procedure is for shutting down and moving Site B to a new location while leaving Site A up and running after failing over the guest VMs***. Please note that this procedure only works for systems prior to 4.5 since the secure admin shell introduced in 4.5+ disallows certain commands. This method will work in 4.5+ if you have performed the root token workflow and gained root access to the system.***

1. Leave witness alone.
2. Log in to each storage controller VM on Site B and issue a power off command.
3. Via the vSphere Web Client, put each ESX server in Site B in maintenance mode.
4. Via the vSphere Web Client, shutdown all ESX servers in Site B.
5. Power off FI's and TOR switches at Site B.
6. Physically remove equipment and re-rack at new site.
7. Cable everything back up and bring online the TOR switches, then FI's.
8. Wait for FI's to fully boot and then via UCSM power up all the servers.
9. Wait for all ESX servers to fully boot.
10. Via vSphere Web Client exit maintenance mode.
11. Wait for all storage controller VMs to fully boot.
12. Check and monitor the health of the Stretch cluster either via HX Connect or via SSH into the HX Connect CIP-M using stcli cluster info.

There will be transient system errors and APD warnings as the system stabilizes and re-establishes connectivity. Once the system is whole again, you can reset these errors in vCenter. There will also be a re-synchronization time for both zookeeper and any forward progress made on the datastores while the site was shutdown. You will need to manually vMotion the fail-over VMs on Site A back to their proper compute resources on site B.

For systems running 4.5+ with secure admin shell, you will need to use a different procedure. This method is non-root, non-CLI:

1. Migrate VMs from the first node to shut down in Site B to a node in Site A
2. Place the first node to shut down in maintenance mode using vCenter
3. Power off the node in vCenter
4. Repeat for all nodes in Site B

APPENDIX B – Rebooting an entire cluster

1. Check the health of the cluster.
2. Power down the guest VM's.
3. Shutdown the HX cluster using command "stcli cluster shutdown"
4. Power off SCVMs
5. Put the hosts in ESX maintenance mode
6. Power down the hosts.

Power off the FIs

Power on the FIs. Once both are UP and UCSM is accessible, proceed to next step

7. Power on the hosts; Ensure that all the Host get discovered by UCSM properly. They should also automatically re-associate the Service-profile and Boot to the ESXi OS
8. Ensure the hosts are connected to VC
9. Exit the hosts out of maintenance mode.
10. Power on the SCVM. The SCVMs will need some time to finish reading the disks
11. Start the HX cluster using command “stcli cluster start”
12. Check the cluster health state
13. Power on guest VM’s
14. Check the datastores

APPENDIX C – Changeable quantities

The following quantities are changeable in the cluster, with TAC assistance, for the following combinations. Currently only CVM IP changes are not supported.

Supported

1. Witness VM Change only
2. Witness + ESX
3. ESX IPs
4. Cluster mgmt. ip
5. Witness + #3 and #4

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at www.cisco.com/go/offices.