Operating Cisco HyperFlex HX Data Platform Stretch Clusters

Version 1.2.
January 2019

Document information

<table>
<thead>
<tr>
<th>Document summary</th>
<th>Prepared for</th>
<th>Prepared by</th>
</tr>
</thead>
<tbody>
<tr>
<td>v1.2 Stretch Cluster information for HX 3.0 and 3.5</td>
<td>Cisco Field</td>
<td>Aaron Kapacinskas</td>
</tr>
</tbody>
</table>

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All information in this document is provided in confidence and shall not be published or disclosed, wholly or in part, to any other party without Cisco’s written permission.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

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document information</td>
<td>1</td>
</tr>
<tr>
<td>Intended use and audience</td>
<td>1</td>
</tr>
<tr>
<td>Legal notices</td>
<td>1</td>
</tr>
<tr>
<td>Prerequisites</td>
<td>3</td>
</tr>
<tr>
<td>Introduction</td>
<td>3</td>
</tr>
<tr>
<td>Cisco HyperFlex HX Data Platform general overview: Components and environment</td>
<td>3</td>
</tr>
<tr>
<td>Cisco Unified Computing System</td>
<td>3</td>
</tr>
<tr>
<td>Fabric interconnect traffic and architecture</td>
<td>4</td>
</tr>
<tr>
<td>Cisco HyperFlex HX stretch clusters</td>
<td>8</td>
</tr>
<tr>
<td>What is a stretch cluster</td>
<td>8</td>
</tr>
<tr>
<td>Business need for a stretch cluster</td>
<td>9</td>
</tr>
<tr>
<td>Stretch cluster physical limitations</td>
<td>9</td>
</tr>
<tr>
<td>Solution components</td>
<td>9</td>
</tr>
<tr>
<td>Stretch cluster architecture</td>
<td>11</td>
</tr>
<tr>
<td>Limitations</td>
<td>13</td>
</tr>
<tr>
<td>Fabric interconnects</td>
<td>14</td>
</tr>
<tr>
<td>VMware vCenter</td>
<td>14</td>
</tr>
<tr>
<td>Witness configuration</td>
<td>14</td>
</tr>
<tr>
<td>Sizing</td>
<td>15</td>
</tr>
<tr>
<td>Failure sizing</td>
<td>16</td>
</tr>
<tr>
<td>I/O path in a stretch cluster</td>
<td>16</td>
</tr>
<tr>
<td>Stretch cluster installation</td>
<td>17</td>
</tr>
<tr>
<td>Cisco HyperFlex installer</td>
<td>18</td>
</tr>
<tr>
<td>Default passwords</td>
<td>20</td>
</tr>
<tr>
<td>VLANs and vSwitches</td>
<td>20</td>
</tr>
<tr>
<td>Datastore Best Practices</td>
<td>22</td>
</tr>
<tr>
<td>Troubleshooting</td>
<td>22</td>
</tr>
<tr>
<td>Stretch cluster operations</td>
<td>22</td>
</tr>
<tr>
<td>Stretch cluster failure modes</td>
<td>24</td>
</tr>
<tr>
<td>Types of failures</td>
<td>24</td>
</tr>
<tr>
<td>Failure response summary</td>
<td>26</td>
</tr>
<tr>
<td>For more information</td>
<td>27</td>
</tr>
</tbody>
</table>
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 to 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 preintegrated 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 Cisco UCS 6248UP 48-Port or 6296UP 96-Port Fabric Interconnects and 6332 or 6332-16UP 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:

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.

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:

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:

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

Each node consists of the virtual machine NIC (VMNIC) and vSwitch architecture shown in Figure 3.
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. These 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 to 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, firewalls, 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 nonstretch, 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

![Figure 4: Three main components of a stretch cluster deployment](image)

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 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 write latency of less than 20 milliseconds (ms). Many other applications require latency of less 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.
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 nonhyperconverged infrastructure [HCI] or traditional storage) that implement stretch or metropolitan clusters on VMware.
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).
  - Stretch Layer 2 characteristics
  - The stretch data VLAN must use jumbo maximum transmission units (MTUs). This setting is validated by the installer, and the installation process will halt if this setting is not in place.
  - 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 `VMkp ping -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 20-ms RTT latency between the active sites and the witness site
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. 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
  - Eight small-form-factor (SFF) nodes per site (16 total)
  - Four large-form-factor (LFF) nodes per site (eight total)
  - 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.
- Computing-only nodes are supported in HyperFlex 3.5
- ESXi is the only supported hypervisor at this time. VMware vSphere Release 6.0 U3 or 6.5 U1 is required.
- Cisco HyperFlex native replication is supported in HyperFlex 3.5
- 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.
- 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.

**Overlay networks are not currently qualified for use with Stretch Clusters.** This means that VxLAN and OTV are not supported. ACI is in testing since it utilizes VxLAN and will be part of this qualification process as well. Contact your Business Unit representative for additional details and testing status.

The witness requires ESXi at the third site (cloud deployment is not currently supported).
**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 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 to 16 converged nodes per site with up to a 2:1 compute node ratio for a maximum mixed count of 32 per site.

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, Cisco UCS Manager Release 3.2(3e) is required.
- Existing fabric interconnects are supported as long as they work with Cisco M5 nodes.
- Nodes 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 is 16 converged nodes per site starting in 3.5 with a 2:1 maximum compute ratio (max 32 mixed nodes per site).

**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.

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.

**Witness configuration**

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) 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 200-ms RTT to each site is required.
- Bandwidth of 100 Mbps to each site is required.
- The node must be deployed separately before the Cisco HyperFlex installer stretch cluster workflow is run.

Even 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 is currently not supported in 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).

**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](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 is 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 and can tolerate a local disk or node failure.

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 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 continuity 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.

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 be 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.

**Stretch cluster installation**

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


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:


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 must support jumbo frames.
- Test the RTT ping using VMkping -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 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 should be 100 Mbps with a 200-ms RTT latency.
- 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.

VMware vCenter has some installation preinstallation and postinstallation 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.)
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

---

**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**

![Hyperflex Installer Workflow](image)

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 must 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

**Figure 11.** Screen showing management and Cisco HyperFlex storage data networks
For an in-depth discussion of virtual distributed switches (VDS) with Cisco HyperFlex systems, see the following resource:


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.

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
- 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.
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.

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

One of the main reasons 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 cluster.

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. The function of ZooKeeper is to maintain the cluster membership and a consistent clusterwide 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, 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 more than 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 the number of nodes are lost in a single cluster.

If a node goes down that was hosting a ZooKeeper instance, the ZooKeeper algorithm will reelect another node for ZooKeeper. However, four more ZooKeeper instances are still running, which is more than the minimum of three ZooKeeper instances required. Hence, the cluster will not be affected. Only the affected virtual machines will be failed over to 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.

Types of failures

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.

- **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, the site will be offline, and site failover will occur.

- **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.
- **Accidental deletion of the witness virtual machine**
  - Contact the Cisco TAC for a recovery process.
- **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 follow process is implemented:
    1. When a stretch cluster is created, one site is biased to establish a ZooKeeper leader. 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 leader form the quorum at site A.
    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. However, virtual machine failback is not automatic.
**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>
<thead>
<tr>
<th>Component failure</th>
<th>Cluster behavior</th>
<th>Quorum update</th>
<th>Virtual machine restart</th>
<th>Site status</th>
<th>Cluster status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-site cache disk</td>
<td>Site is online, with diminished cache capacity.</td>
<td>No</td>
<td>No</td>
<td>Online</td>
<td>Online</td>
</tr>
<tr>
<td>Single-site persistent disk</td>
<td>Site is online, with diminished capacity, and is rebuilt after 2 minutes using the remaining capacity.</td>
<td>No</td>
<td>No</td>
<td>Online</td>
<td>Online</td>
</tr>
<tr>
<td>Double-site cache disk</td>
<td>Site is online, with diminished cache capacity.</td>
<td>No</td>
<td>No</td>
<td>Online</td>
<td>Online</td>
</tr>
<tr>
<td>Double-site persistent disk</td>
<td>If failure is simultaneous and on different nodes, it is a double-node failure, and the site is offline. 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, this is a node failure.</td>
<td>Yes</td>
<td>Yes</td>
<td>Offline</td>
<td>Online</td>
</tr>
<tr>
<td>Single-site single node loss</td>
<td>No impact on the site; recover the fabric interconnect.</td>
<td>No</td>
<td>No</td>
<td>Online</td>
<td>Online</td>
</tr>
<tr>
<td>Single-site multiple node loss</td>
<td>Site is offline.</td>
<td>Yes</td>
<td>Yes</td>
<td>Offline</td>
<td>Online</td>
</tr>
<tr>
<td>Single-site single fabric interconnect loss</td>
<td>Site is offline.</td>
<td>Yes</td>
<td>Yes</td>
<td>Offline</td>
<td>Online</td>
</tr>
<tr>
<td>Single-site double fabric interconnect loss</td>
<td>Site is offline.</td>
<td>Yes</td>
<td>Yes</td>
<td>Offline</td>
<td>Online</td>
</tr>
<tr>
<td>Double-site single fabric interconnect loss</td>
<td>Site is offline.</td>
<td>Yes</td>
<td>Yes</td>
<td>Offline</td>
<td>Online</td>
</tr>
<tr>
<td>Double-site double fabric interconnect loss</td>
<td>Site is offline.</td>
<td>Yes</td>
<td>Yes</td>
<td>Offline</td>
<td>Online</td>
</tr>
<tr>
<td>Witness loss</td>
<td>No impact on the site; recover the witness.</td>
<td>No</td>
<td>No</td>
<td>Online</td>
<td>Online</td>
</tr>
<tr>
<td>Single-site single switch loss</td>
<td>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.</td>
<td>No</td>
<td>No</td>
<td>Online</td>
<td>Online</td>
</tr>
<tr>
<td>Single-site double switch loss</td>
<td>Site is offline.</td>
<td>Yes</td>
<td>Yes</td>
<td>Offline</td>
<td>Online</td>
</tr>
<tr>
<td>Double-site single switch loss</td>
<td>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.</td>
<td>No</td>
<td>No</td>
<td>Online</td>
<td>Online</td>
</tr>
<tr>
<td>Double-site double switch loss</td>
<td>Both sites are offline.</td>
<td>–</td>
<td>–</td>
<td>Offline</td>
<td>Offline</td>
</tr>
<tr>
<td>Site loss</td>
<td>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.</td>
<td>Yes</td>
<td>Yes</td>
<td>Offline</td>
<td>Online</td>
</tr>
<tr>
<td>Site link loss</td>
<td>See the preceding detailed discussion.</td>
<td>Yes</td>
<td>Yes</td>
<td>Online</td>
<td>Online</td>
</tr>
</tbody>
</table>

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. 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.

For more information

For additional information, see the following resources:

- Cisco HyperFlex 3.0 with VSI:

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

- Cisco HyperFlex preinstallation checklist:

- Cisco HyperFlex release notes:

- Cisco HyperFlex with VDSs:

- ZooKeeper:
  https://en.wikipedia.org/wiki/Apache_ZooKeeper

- Cisco HyperFlex with disjointed Layer 2 networking:

- vNICS and vHBAs in Cisco UCS: