

Cisco HyperFlex Invisible Cloud Witness Powered by the Cisco Intersight Platform

Last Updated: April 21, 2022

Contents

The need for an Invisible Cloud Witness	3
Built-in resiliency with Cisco HyperFlex Edge Invisible Cloud Witness powered by the Cisco Intersight platform	5
Benefits of Cisco Invisible Cloud Witness architecture	6
Cisco Invisible Cloud Witness setup and operation	7
Cisco Invisible Cloud Witness operations	7
Cisco Invisible Cloud Witness requirements	8
Cisco Invisible Cloud Witness failure cases	8
Case 1: Loss of WAN connectivity to Cisco Intersight platform	9
Case 2: Cisco HyperFlex node failure	12
Case 3: Loss of connectivity between nodes	13
Case 4: Cisco HyperFlex node failure followed by a WAN failure	14
Case 5: WAN failure followed by a node failure	15
Local Witness Option for two-node HyperFlex Edge clusters	16
Cisco HyperFlex Containerized Witness	17
Local Witness on Schneider Electric NMC	18
Conclusion	19
For more information	19

The need for an Invisible Cloud Witness

When running a clustered file system such as the Cisco HyperFlex™ HX Data Platform, data consistency across nodes is of paramount importance. Cisco HyperFlex systems are built on a quorum mechanism that can help guarantee consistency whenever that quorum is available. A quorum in Cisco HyperFlex systems traditionally is based on a node majority in which each node in the system casts a single vote, with a simple majority required for the cluster to remain operational. This mechanism works well for three-node and larger clusters that can tolerate one or more node losses and still be able to obtain a majority consensus and continue operations.

However, fault tolerance and file system consistency becomes more challenging when only two nodes are deployed at a customer's Remote Office or Branch Office (ROBO) location. In this scenario, if one of the two nodes fails, a quorum can no longer be established using a node majority algorithm alone. In the unlikely event that the communication pathway between the two nodes is disrupted, a "split brain" condition may occur if both nodes continue to process data without obtaining a quorum. The opposite outcome—the loss of availability—is also possible. You must avoid both scenarios to help prevent the introduction of data inconsistency while also maintaining high availability.

For these reasons, hyperconverged two-node architectures require an additional component, sometimes referred to as a witness or arbiter, that can vote if a failure occurs within the cluster. This traditional and burdensome deployment architecture requires the additional witness to be provisioned on existing infrastructure and connected to the remote cluster over the customer's network. Typically, these external witnesses are packaged as either virtual machines or standalone software that is installable within a guest operating system. Figure 1 shows this classic topology.

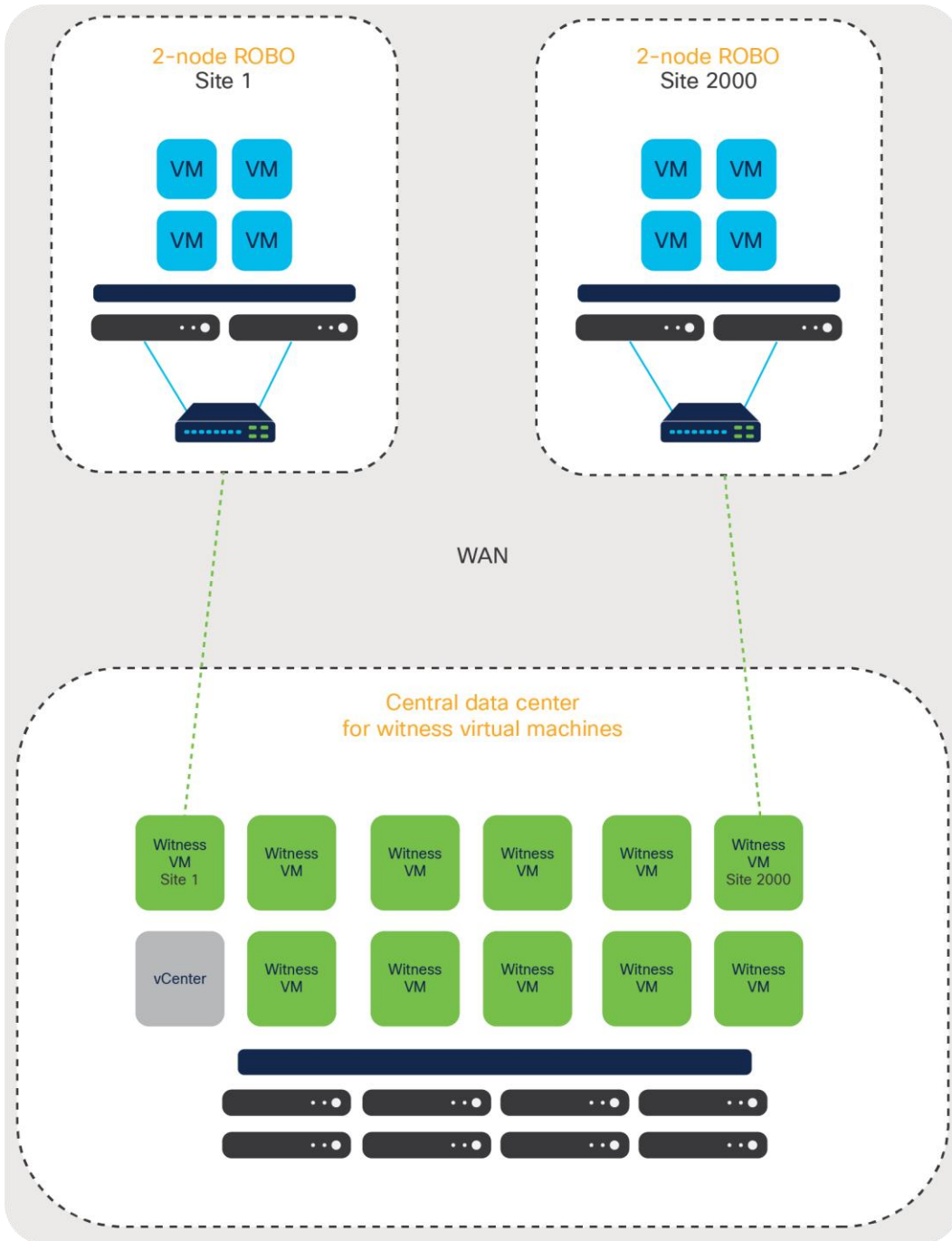


Figure 1.
Traditional witness architecture

Built-in resiliency with Cisco HyperFlex Edge Invisible Cloud Witness powered by the Cisco Intersight platform

In response to the capital expenditures (CapEx), ongoing operating expenses (OpEx), and long time to value when organizations need to deploy ROBO witness virtual machines, Cisco developed a next-generation invisible witness architecture for Cisco HyperFlex Edge deployments powered by the Cisco Intersight™ platform. This innovative new deployment methodology eliminates the need for witness virtual machines, ongoing patching and maintenance, and additional infrastructure at a third site. Figure 2 shows the Cisco HyperFlex Edge Invisible Cloud Witness architecture.

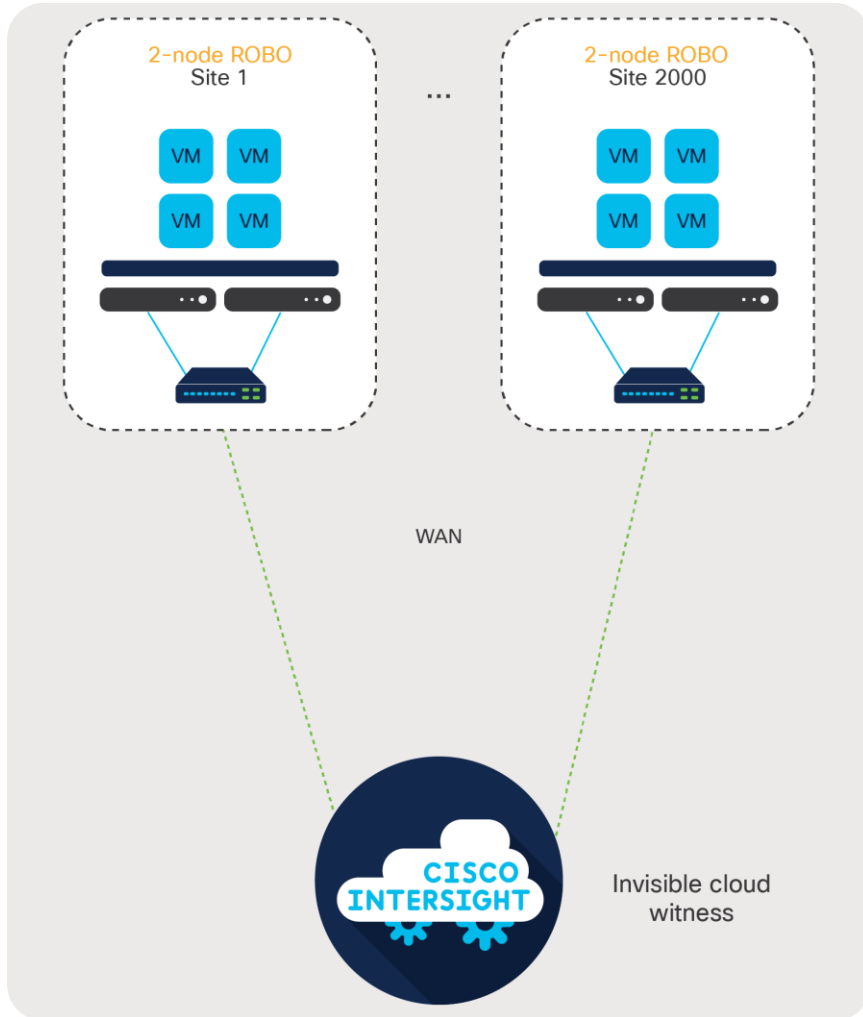


Figure 2.
Cisco Invisible Cloud Witness architecture

Benefits of Cisco Invisible Cloud Witness architecture

The next-generation Cisco® Invisible Cloud Witness architecture is designed to address some of the most difficult challenges facing customers when they deploy two-node Hyperconverged Infrastructure (HCI) clusters at scale. It offers these advantages:

- No additional license cost; the license is included in the Cisco HyperFlex Edge license subscription.
- No need for a third site or for existing computing, storage, and network infrastructure
- Cloud-like operations; with the Cisco Invisible Cloud Witness, there is nothing to manage:
 - No user interface to monitor
 - No user-based software updates
 - No setup, configuration, or backup required
 - No scale limitations
 - Built-in high availability and fault tolerance
- Security
 - Real-time updates to the Cisco Invisible Cloud Witness service with the latest security patches
 - All communications encrypted using Transport Layer Security (TLS) 1.2
 - Use of standard HTTPS port 443; no firewall configuration required
- Built on an efficient, silent protocol stack
 - No periodic heartbeats sent across the WAN
 - No cluster metadata or user data transferred to the Cisco Intersight platform
 - Tolerant for high latency and lossy WAN connections

Cisco Invisible Cloud Witness setup and operation

Getting started with the Cisco Invisible Cloud Witness is simple. The Invisible Cloud Witness does not require any configuration input. Instead, all components are configured transparently during Cisco HyperFlex Edge cluster installation. Before the Cisco Edge installation process begins, the physical servers must be securely claimed in the Cisco Intersight platform. From that point forward, the Invisible Cloud Witness is automatically deployed and configured.

Figure 3 shows an overview of the Invisible Cloud Witness architecture.

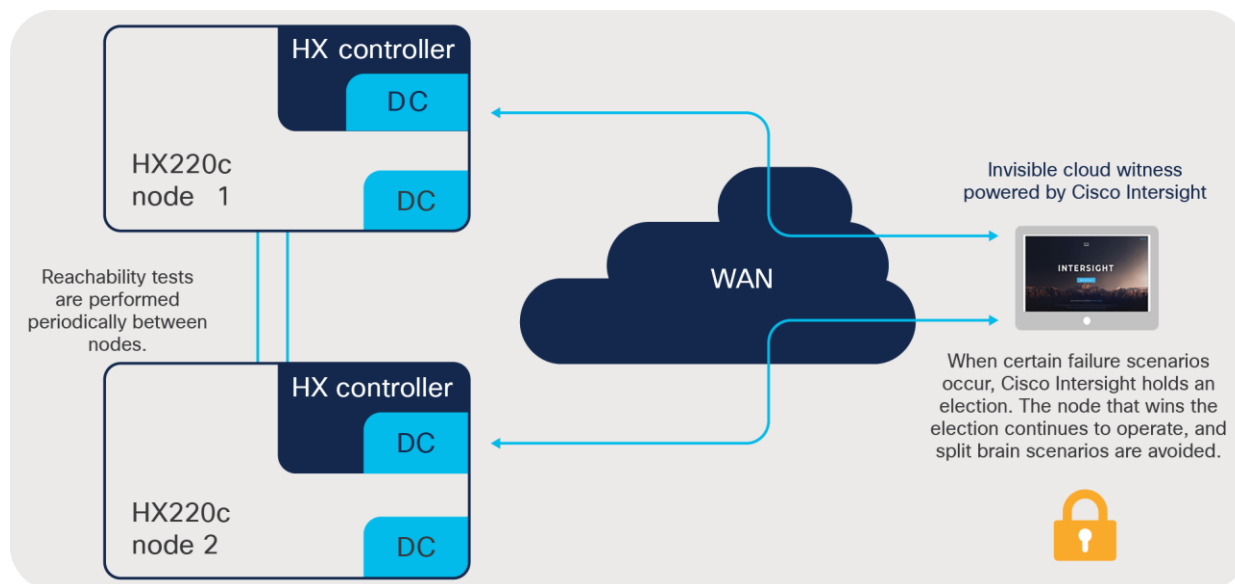


Figure 3.
Cisco Invisible Cloud Witness functional diagram

Cisco Invisible Cloud Witness operations

The Cisco Intersight platform uses an embedded device connector inside the Cisco HyperFlex controller virtual machines to deliver management capabilities to the platform. Features enabled by the device connector include the capability to monitor alarms, view cluster status, and cross-launch into the Cisco HyperFlex Connect user interface from anywhere.

Using this existing software component, the Cisco Invisible Cloud Witness communicates over the existing control channel: a secure, persistent web socket between the device connector and the Cisco Intersight service. For two-node clusters only, the Cisco Intersight platform also performs arbitration in the unlikely scenario that a node fails or the communication between nodes fails.

Arbitration and avoidance of a split-brain scenario are accomplished by using an internal election protocol that grants the right to continue operating to a single winner. The Cisco Intersight platform maintains the election state for each remote cluster and will, essentially, vote with one of the surviving nodes to help ensure continued cluster operation in the event of a failure.

When a failure is detected, the surviving nodes reach out to the Cisco Intersight platform for permission to continue I/O operations. The witness service grants privileges to only one of the nodes to continue operation. The purpose-built witness protocol is efficient in that it communicates only under failure conditions. This design helps ensure that even the most unreliable WAN connections are suitable for use with this Invisible Cloud Witness architecture.

The two Cisco HyperFlex clustered nodes periodically perform reachability tests over the local network. If reachability is interrupted because of either a node or a link failure, the process of reaching out to the Cisco Intersight platform will begin. The details of this process are described in the failure scenarios that follow.

Cisco Invisible Cloud Witness requirements

The Cisco Invisible Cloud Witness requires each Cisco HyperFlex controller virtual machine to have public Domain Name System (DNS) resolution capabilities and the capability to initiate an outbound TCP/HTTPS Internet connection on port 443 to the Cisco Intersight platform. Typically, these requirements are fulfilled by allowing outbound Network Address Translation (NAT) operations from the subnet assigned to the Cisco HyperFlex management traffic. If direct outbound connectivity is not available, an HTTPS proxy can be configured to enable use of this service.

Cisco Invisible Cloud Witness failure cases

Cisco HyperFlex systems are designed so that there is no single point of failure, and service is not disrupted even when individual components fail. Common points of failure include:

- Cisco HyperFlex nodes (power, hardware, and software failures)
- LAN links and switches
- WAN links and routers

It is expected that if any single component fails, the cluster and workloads will remain online and operational.

However, two-node hyperconverged clusters in general are not designed to handle all combinations of simultaneous failures at the same time. The system is, however, designed to be able to suffer a failure, eventually return to a healthy state (in many cases automatically), and then tolerate another failure at some point in the future.

A full discussion of the failure of Cisco HyperFlex drives, including boot drives, system drives, cache drives, and capacity drives, is outside the scope of this document. For detailed failure tolerances and recovery procedures, see the [Cisco HyperFlex Administration Guide](#). The capability of Cisco HyperFlex Edge two-node clusters to tolerate drive failures matches that of traditional Cisco HyperFlex clusters configured with Replication Factor 2 (RF2).

In many failure scenarios, VMware High Availability (HA) is responsible for restarting the virtual machines on the surviving nodes. After recovery is complete, the virtual machines should be redistributed across both nodes, either automatically through VMware Distributed Resource Scheduler (DRS), if licensed, or manually through VMware vMotion.

Case 1: Loss of WAN connectivity to Cisco Intersight platform

Figure 4 shows failure case 1.

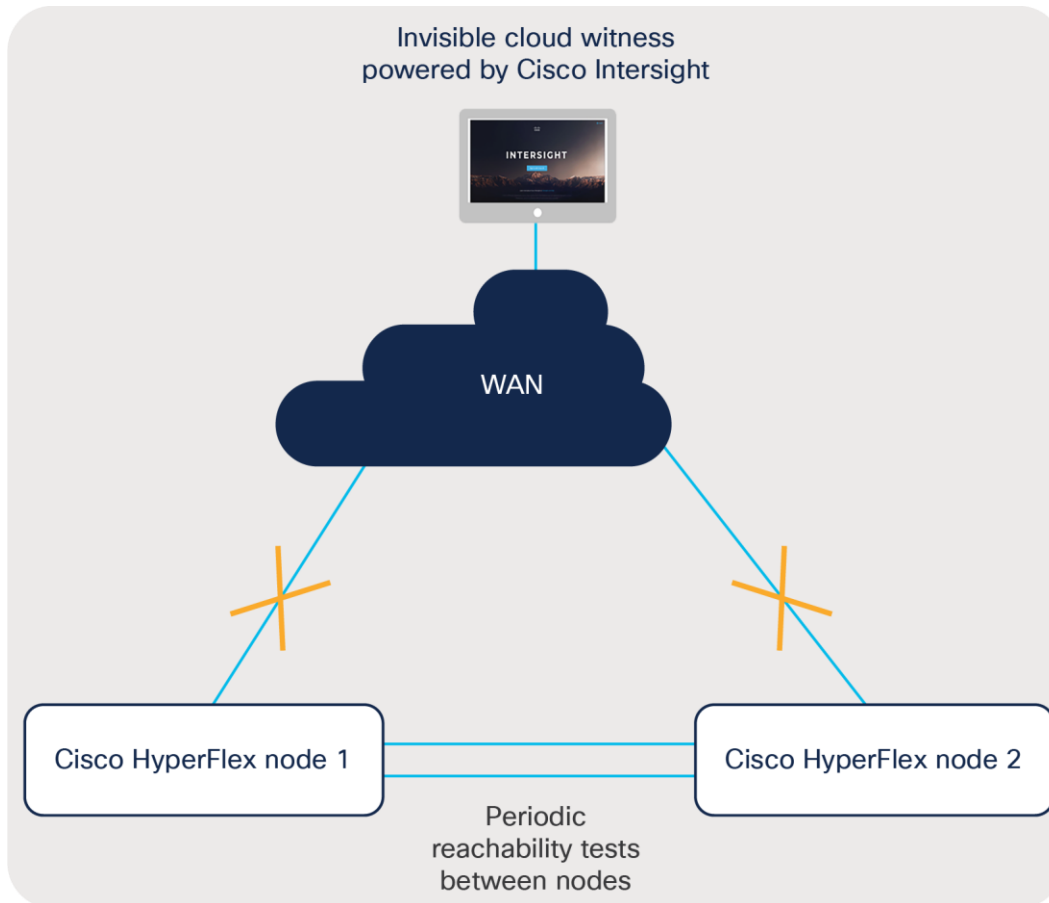


Figure 4.
Loss of WAN connectivity to Cisco Intersight platform

In this case, the entire site loses WAN connectivity to the Cisco Intersight platform. During the WAN outage, both Cisco HyperFlex nodes and all workloads continue to operate uninterrupted. Operation is not interrupted because communication to the Invisible Cloud Witness is required only when a local failure of a LAN link, switch, or Cisco HyperFlex node occurs.

When the WAN link is down and the Cisco Intersight platform is not reachable, additional LAN or node failures cannot be tolerated. It is a best practice to design WAN connectivity with reliability in mind.

When WAN connectivity is restored, the device connector automatically reconnects to help ensure that the cluster can tolerate a LAN or node failure. The device connectors have a built-in retry mechanism that quickly reconnects whenever WAN service is restored. You can monitor the current status of connectivity to the Invisible Cloud Witness by using `stcli` commands or launching Cisco HyperFlex Connect on the local cluster.

Note: Reachability to the Cisco Intersight Invisible Cloud Witness is required when you first deploy a Cisco HyperFlex two-node cluster. If the cluster enters an offline or administrative shutdown state, the cluster can be successfully restarted without connectivity to the Invisible Cloud Witness so long as both nodes can reach each other.

Figure 5 shows the view from the Cisco HyperFlex Connect dashboard when the cluster is fully healthy with fault tolerance.



Figure 5.
View from Cisco HyperFlex Connect dashboard when cluster is healthy with fault tolerance

Clicking the information icon under resiliency health brings up a window with more detailed information (Figure 6). In this case, the cluster is fully healthy and able to tolerate failures.

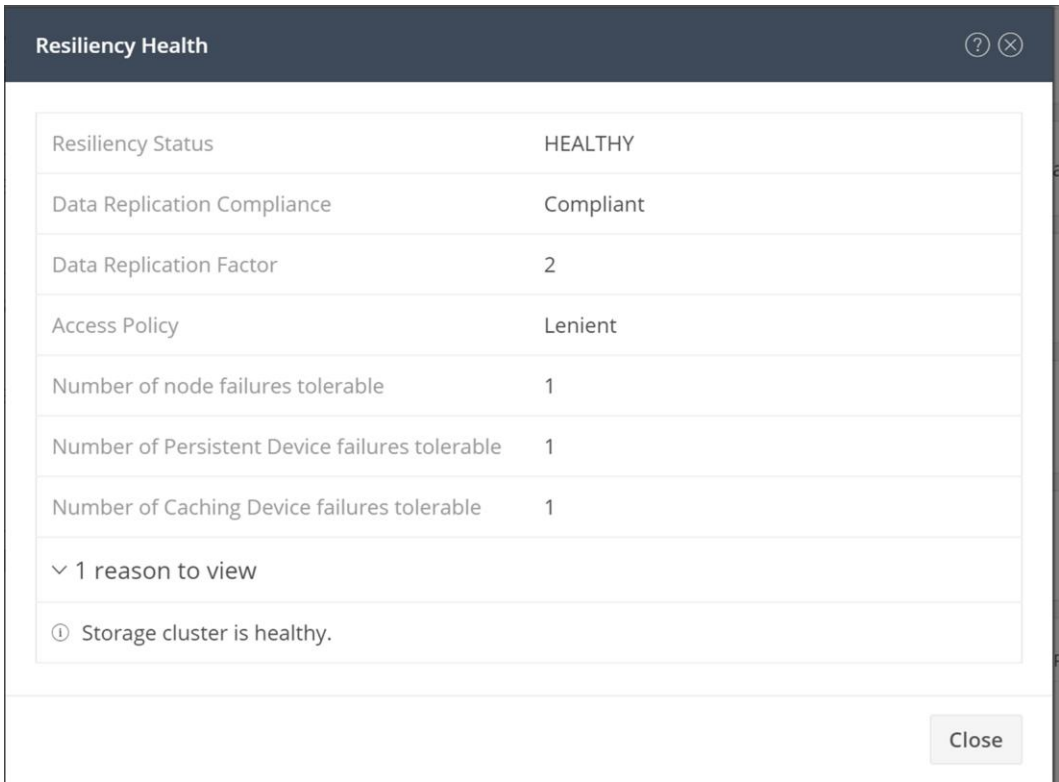


Figure 6.
Resiliency health information showing that the cluster is healthy and can tolerate failures

Figure 7 shows the view from the Cisco HyperFlex Connect dashboard when connectivity to the Cisco Intersight Invisible Cloud Witness is lost. Notice both the warning state and the inability to tolerate a node failure under these conditions.

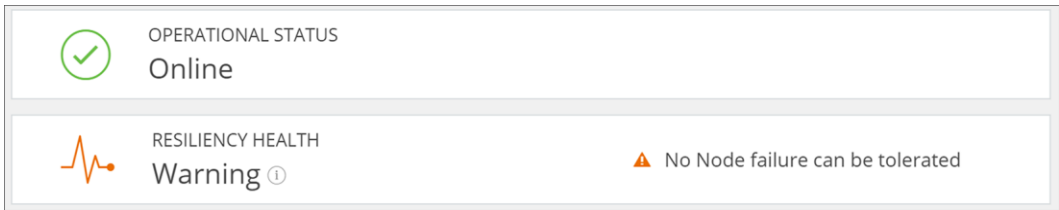
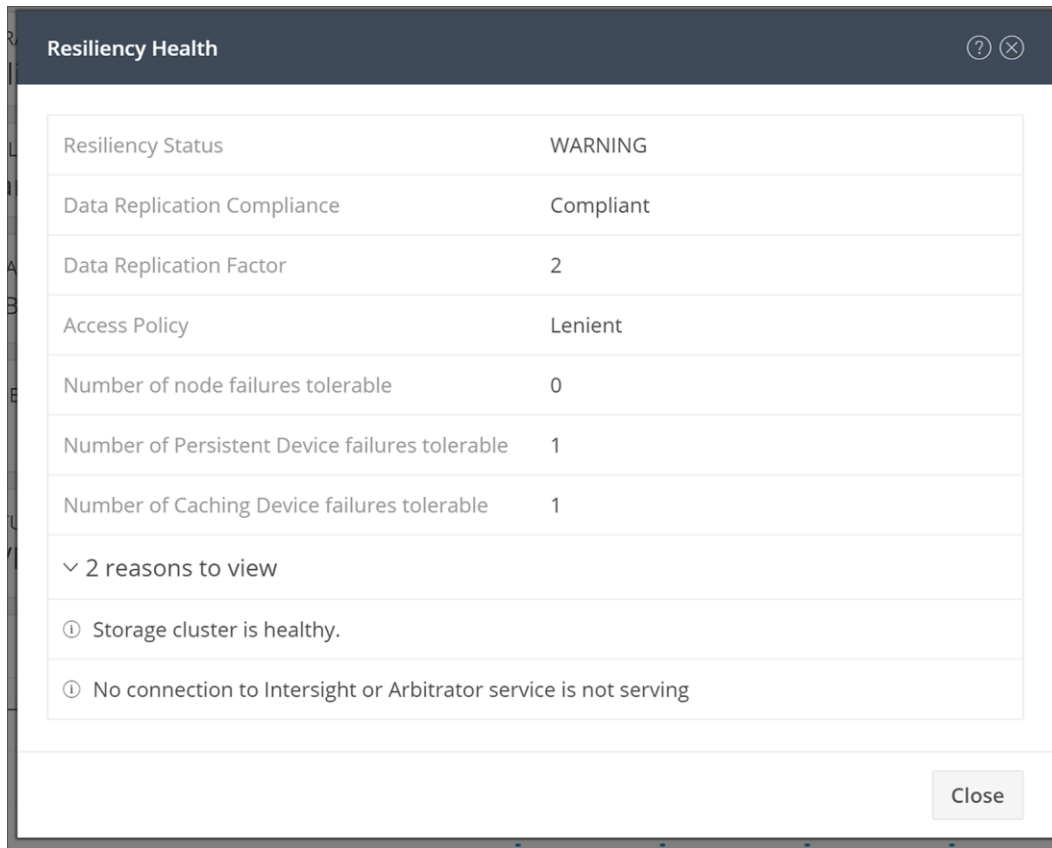


Figure 7.
View from Cisco HyperFlex Connect dashboard when connectivity to Cisco Intersight Invisible Cloud Witness is lost

Clicking the information icon under resiliency health brings up a window with more detailed information (Figure 8). In this case, there is no connection to the Cisco Intersight platform, which reduces the system's ability to tolerate a node failure.



The image shows a 'Resiliency Health' window with a dark header and a light body. The header contains the title 'Resiliency Health' and two icons: a question mark and a close button. The main content is a table with two columns: the left column lists various health metrics, and the right column shows their status. Below the table, there is a section titled '2 reasons to view' with a downward arrow, followed by two items, each with a red circle icon containing a white 'i'.

Metric	Status
Resiliency Status	WARNING
Data Replication Compliance	Compliant
Data Replication Factor	2
Access Policy	Lenient
Number of node failures tolerable	0
Number of Persistent Device failures tolerable	1
Number of Caching Device failures tolerable	1

2 reasons to view

- Storage cluster is healthy.
- No connection to Intersight or Arbitrator service is not serving

Close

Figure 8. Resiliency health information showing that there is no connection to the Cisco Intersight platform

Case 2: Cisco HyperFlex node failure

Figure 9 shows failure case 2.

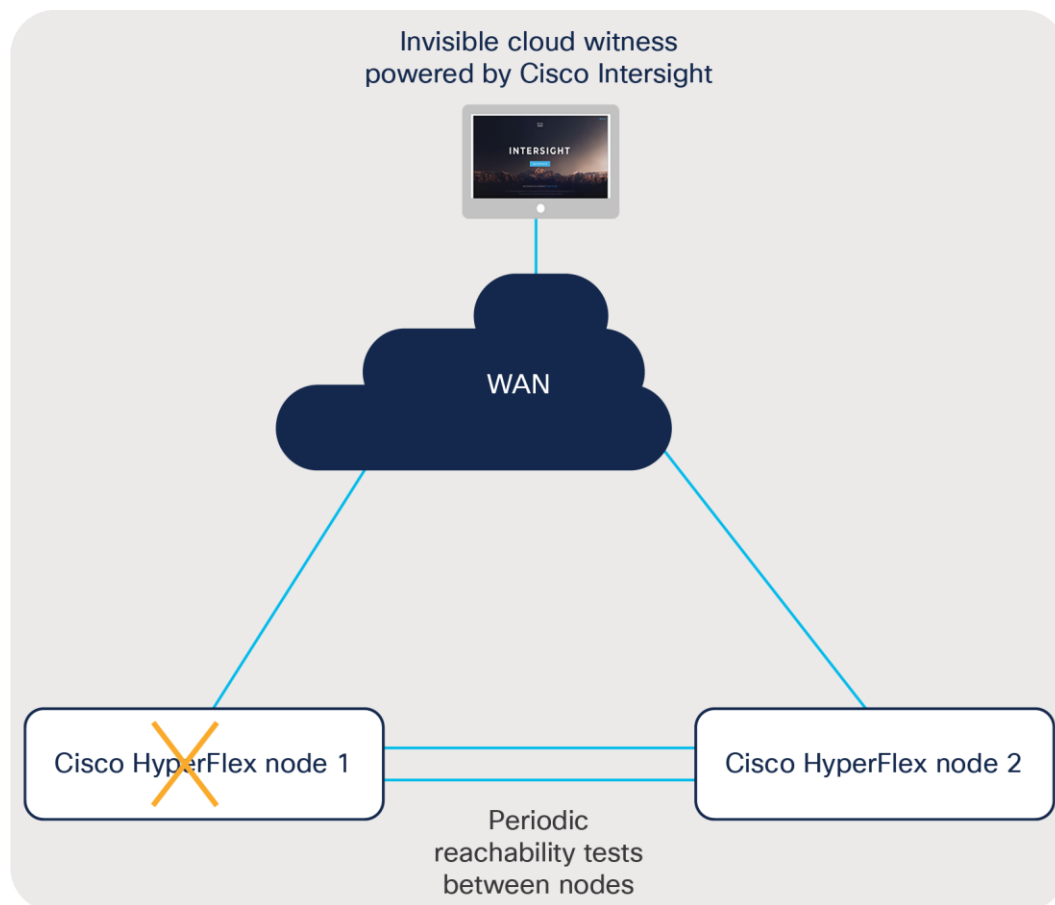


Figure 9.
Cisco HyperFlex node failure

This case shows what happens when a Cisco HyperFlex node fails as a result of a hardware or software failure.

When the local reachability test fails, the surviving node immediately reaches out to the Cisco Intersight platform to obtain permission to continue operating. If the election state is clean for this cluster, meaning that the other node has not been granted the right to continue operating, the Cisco Intersight platform will return a success response, and the cluster will continue to operate on the single surviving node.

The virtual machines that were running on the failed node will be automatically restarted on the surviving node by VMware HA. The virtual machines that were on the surviving node will continue to operate without interruption.

When the failed Cisco HyperFlex node is recovered, the reachability test between nodes will succeed, and the surviving node will relinquish control. This process will clear the state from the Cisco Intersight platform so that future failures can be tolerated. Next, the restored node is resynchronized with the surviving node, and the cluster will become fully healthy again. From this healthy state, fault tolerance is restored, and any new failure scenario can be tolerated again.

The cluster's resiliency health can be viewed at any time in the Cisco HyperFlex Connect dashboard. Figure 5 shows a fully healthy two-node cluster that can tolerate a node failure.

Figure 10 shows the Cisco HyperFlex Connect dashboard during a node failure. Notice that the resiliency health is shown in a warning state with no node failures tolerable. The red node icon indicates that one of the two nodes is currently offline. Despite the node failure, the cluster is still online and serving I/O to applications running on the surviving node.

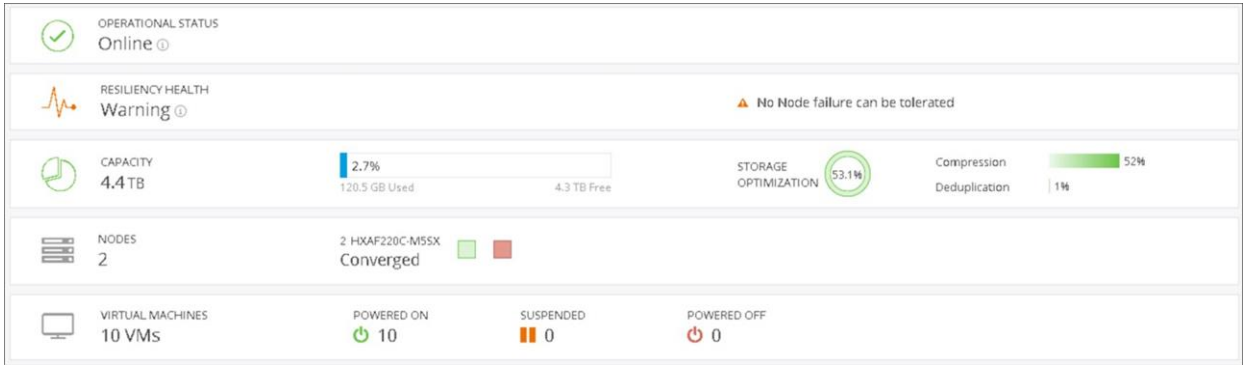


Figure 10.
Cisco HyperFlex Connect dashboard during node failure

Case 3: Loss of connectivity between nodes

Figure 11 shows failure case 3.

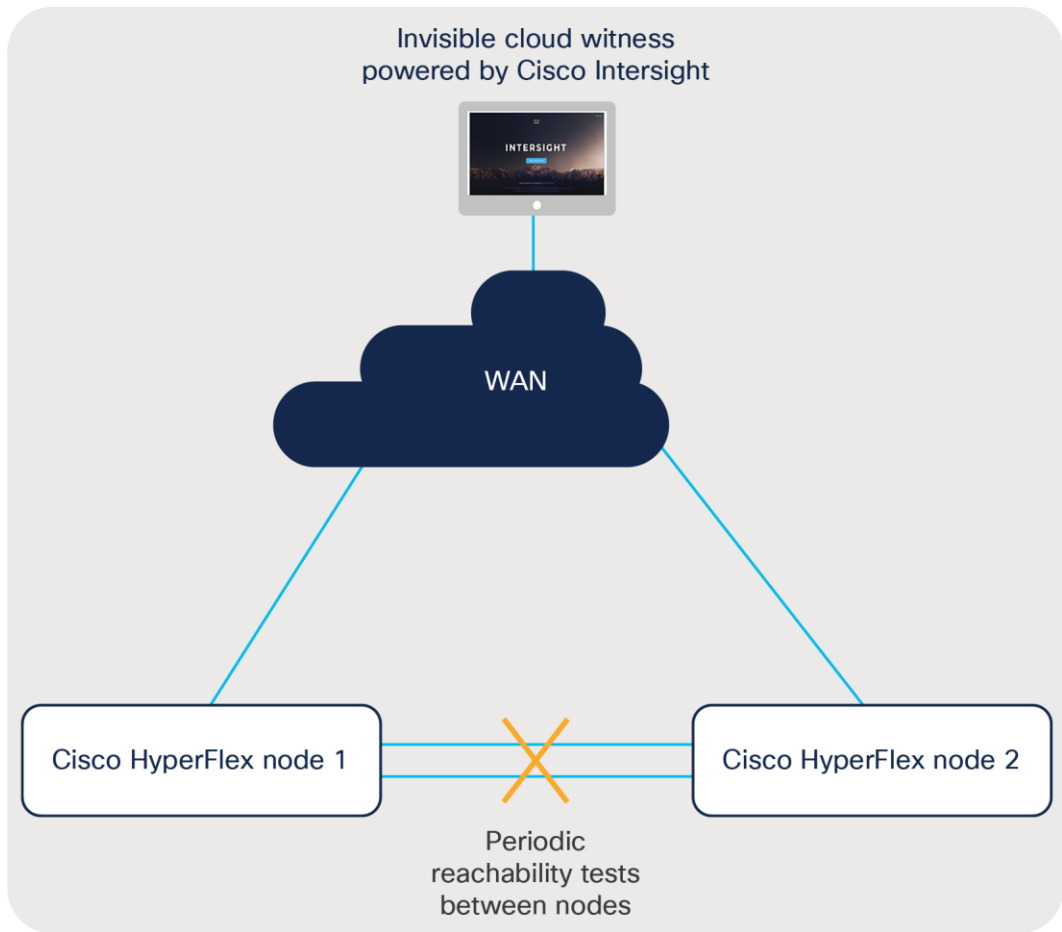


Figure 11.
Loss of connectivity between nodes

- **Scenario A: Single link loss between nodes**

If a single direct-connect cable (1 Gigabit Ethernet topology) or a single network uplink (10 Gigabit Ethernet topology) is severed, all services will immediately fail over to the secondary link, and no additional action is required. The running cluster is not affected.

After the link is restored, the services that failed over will fail back to their original failover priority, and network fault tolerance will be restored.

- **Scenario B: All links between Cisco HyperFlex nodes are lost**

If both direct-connect cables (1 Gigabit Ethernet topology) are lost, or if switching infrastructure or both uplinks from any Cisco HyperFlex node to the switch (10 Gigabit Ethernet topology) are lost, an election process occurs. When the local reachability test fails, both nodes will immediately reach out to the Cisco Intersight platform. Only one node will win the election. This winning node will continue to operate, and the rejected node will enter a suspended state and any Cisco HyperFlex datastores will go offline. VMware HA can be configured to automatically restart the virtual machine on the online node.

After connectivity is restored, the reachability test between nodes will succeed and the operating node will clear the election state from the Cisco Intersight platform so that future failures can be tolerated. Next, the previously offline node is resynchronized, and the cluster will become fully healthy. From this healthy state, any new failure scenario can now be tolerated.

Case 4: Cisco HyperFlex node failure followed by a WAN failure

Figure 12 shows failure case 4.

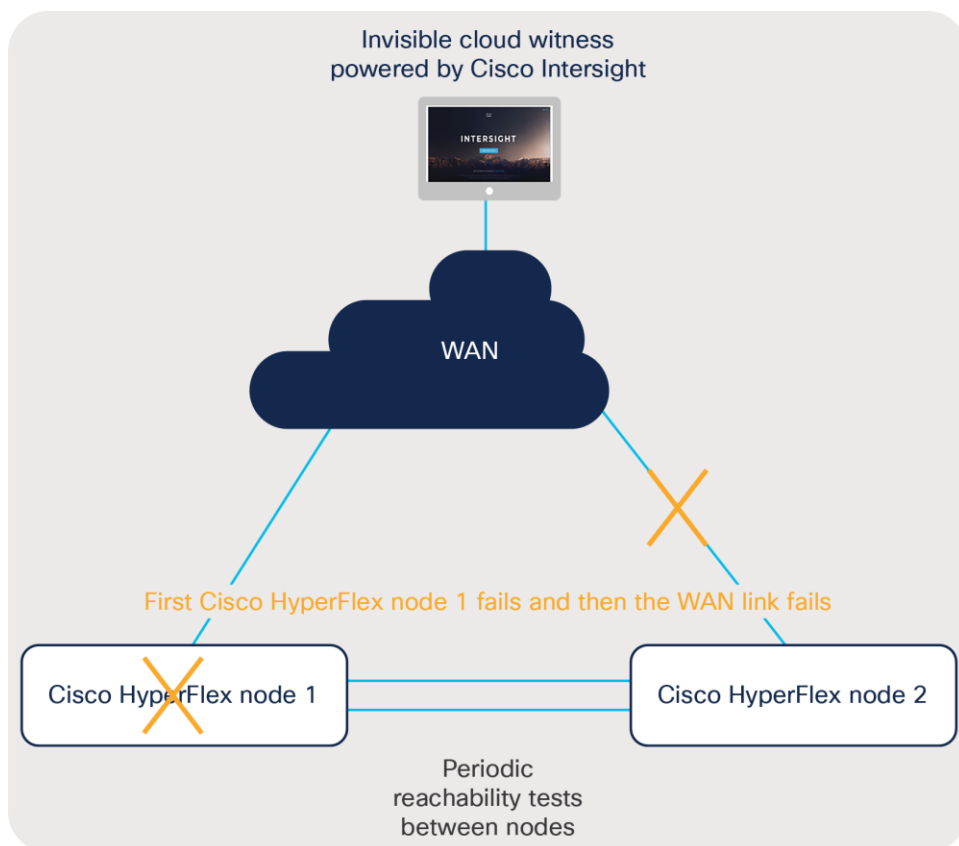


Figure 12.
Cisco HyperFlex node failure followed by a WAN failure

This case shows what happens when first a Cisco HyperFlex node fails as a result of a hardware or software failure and then the WAN fails while the cluster is operating with a single node.

When the node fails, the behavior will be the same as in failure case 2. After the cluster has reached a steady state and is operating on a single node, if the WAN subsequently suffers an outage, there is no impact on the system. This situation is explained in failure case 1. Because the cluster was able to failover to a single node before the WAN outage, there is no need for additional communication with the Invisible Cloud Witness until the failed node is recovered.

- **Recovery scenario 1:** If the WAN comes back online first, the recovery procedure is the same as in failure case 2.
- **Recovery scenario 2:** If the failed node recovers first, the recovered node will synchronize with the surviving node and resume I/O operations. In this state, the cluster cannot tolerate any node failures without affecting I/O operations. After connectivity to Intersight is reestablished, the cluster will synchronize with the Invisible Cloud Witness and regain the default high availability state for a healthy cluster.

Case 5: WAN failure followed by a node failure

Figure 13 shows failure case 5.

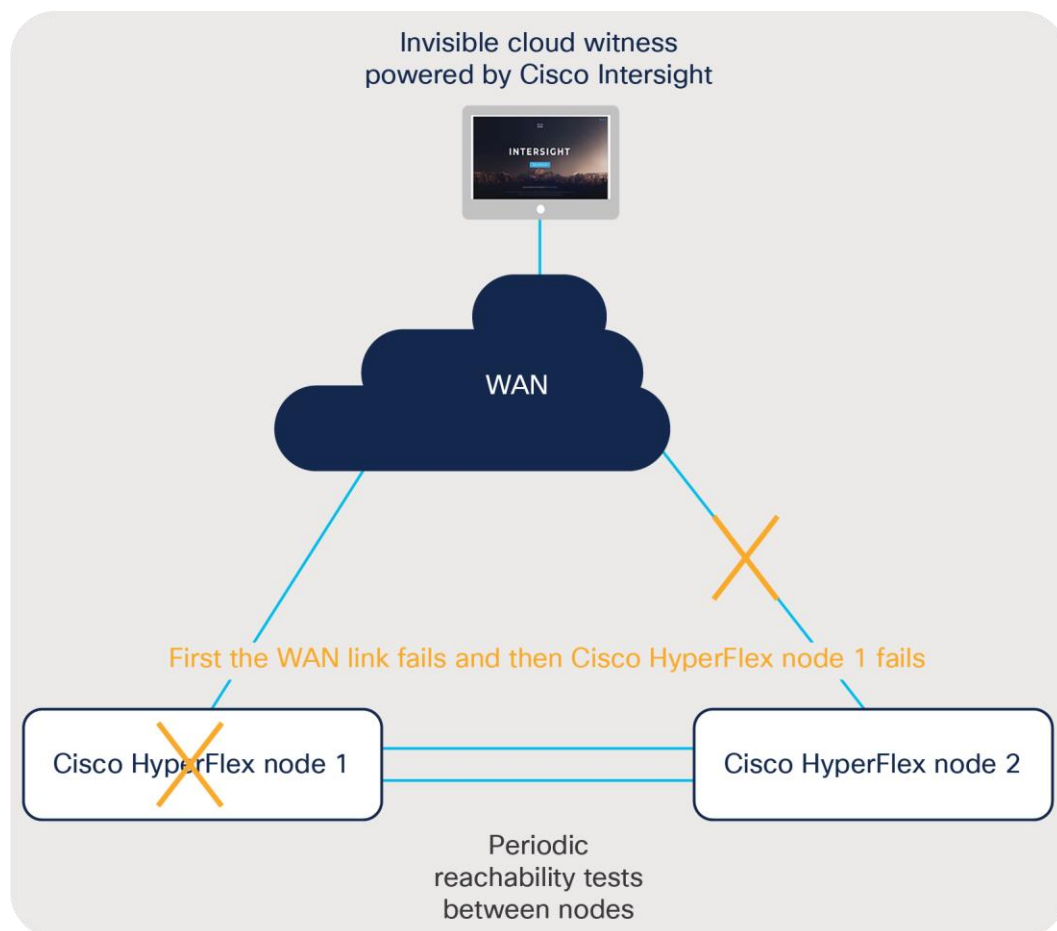


Figure 13.
WAN failure followed by a node failure

As explained in failure case 1, if WAN reachability to the Invisible Cloud Witness is down, another subsequent failure cannot be safely achieved without interrupting I/O operations. To maintain data consistency, the cluster will stall temporarily and be subject to the following recovery scenarios:

- **Recovery scenario 1:** If the WAN recovers first, the surviving node will resume operations serving I/O, the same as in failure case 2, by achieving a quorum with the Cisco Intersight platform. Depending on the length of the WAN outage, some virtual machines may experience an All-Paths Down (APD) condition. The Cisco HyperFlex datastore will come back online automatically. However, some virtual machines may require manual remediation depending on the application's tolerance of I/O interruption.

The cluster is now fully operational on a single node. At some later time, the failed node will be recovered. In this case, the same node recovery process as described in failure case 2 is followed to admit the node back into the cluster and achieve fault tolerance again.

- **Recovery scenario 2:** If the failed node recovers first, the recovered node will synchronize with the surviving node and resume I/O operations. In this state, the cluster cannot tolerate any future node failures without affecting I/O operations. After connectivity to Intersight is reestablished, the cluster will synchronize with the Invisible Cloud Witness and regain the default high availability state for a healthy cluster.

Depending on the length of the WAN outage, some virtual machines may experience an APD condition. The Cisco HyperFlex datastore will come back online automatically. However, some virtual machines may require manual remediation depending on the application's tolerance of I/O interruption.

Local Witness Option for two-node HyperFlex Edge clusters

HyperFlex Edge also provides a flexible, secondary option to use an external (also called a local) witness for clusters that demand the highest levels of availability. Based on the same robust protocol powering the Cisco Intersight Invisible Cloud Witness, this external, or local, witness option operates without any dependency on network connectivity back to Intersight.

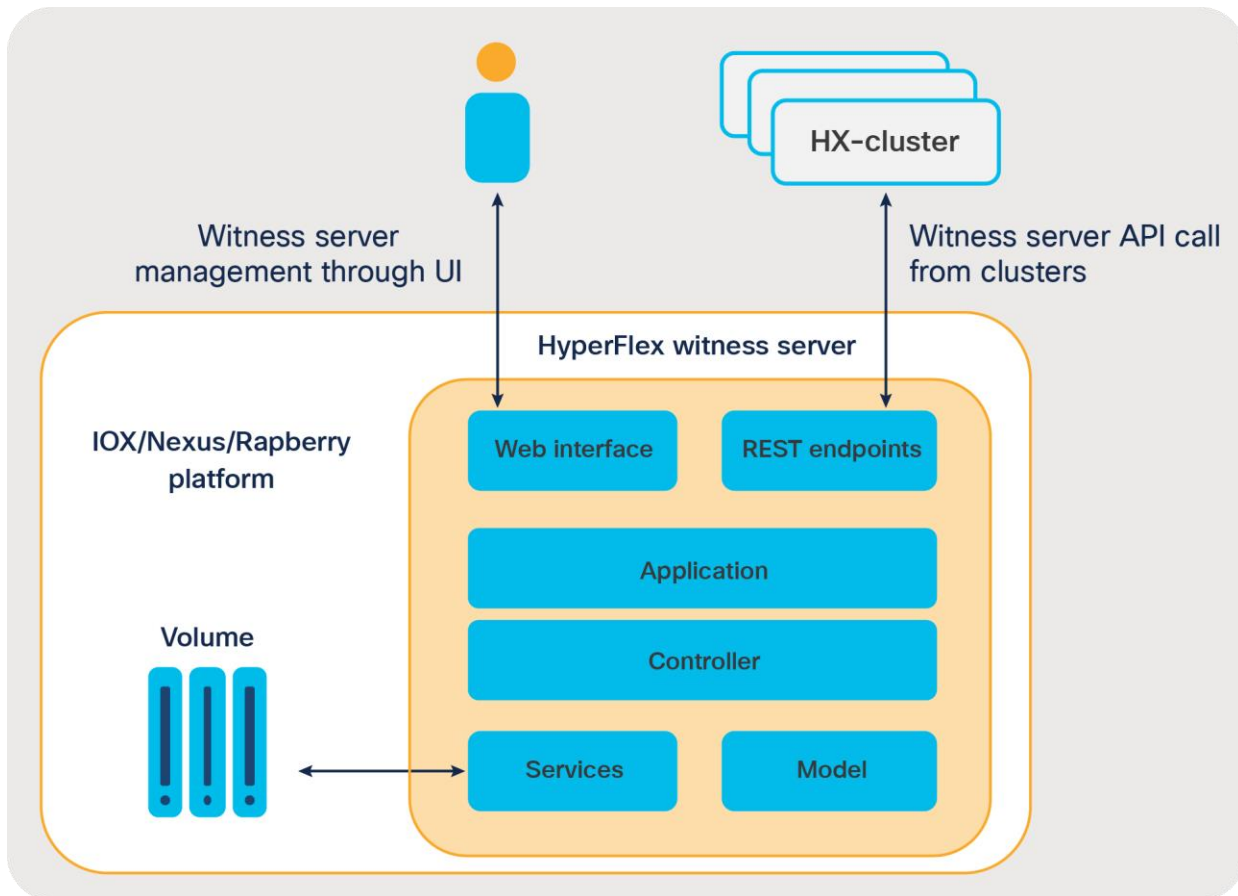
As the HyperFlex Edge clusters continue to expand into new use cases, there have been requests for a secondary option to be able to place the witness locally to the cluster. The goal is to eliminate any dependency on the WAN or any network backhaul to a witness running in another physical location. While the cluster remains healthy and operational when faced with a WAN or network outage, it does affect the ability of the cluster to tolerate a node failure. The purpose of the local witness is to provide a more reliable mechanism to be able to tolerate cascading failure scenarios – for example, losing both the WAN and a node – without causing disruption to the cluster and the workloads hosted there. Using a local witness, Cisco can leverage the same lightweight protocol and design used in the centralized Intersight architecture, but provide higher availability guarantees when required by the customer.

Cisco HyperFlex Containerized Witness

The Cisco HyperFlex Containerized Witness Server is a web application that provides necessary arbitration APIs while providing management functions through a web UI. The witness server supports multiple HX clusters to lock and unlock each cluster ID with the corresponding node ID. The server also provides a UI interface where end users can manage the witness server and perform the following operations:

- View cluster lock information
- Update internal listening port (disabled on wireless APs)
- Generate a new self-signed certificate (new certificate generated on first startup)
- Upload a custom server certificate
- Change password

Below is a high-level architecture of the witness server application:



To learn more information about this innovative lightweight arbitrator and deployment procedure, refer to [HyperFlex Local Containerized witness deployment guide](#).

Local Witness on Schneider Electric NMC

Cisco has partnered with the Schneider Electric team to bring to market an industry first: the ability to run a local witness server on an Uninterruptible Power Supply (UPS) Network Management Card (NMC). A UPS is a fundamental investment for edge locations to protect equipment and data against any unexpected power failures or power anomalies. Cisco and Schneider Electric have worked closely together to embed the witness protocol for HyperFlex Edge two-node clusters directly into the NMC card, thus operating as a third vote for a quorum. The end goal is a turnkey solution that is validated by both Cisco and Schneider Electric, with no need for new infrastructure, dedicated software patching, or another device needed on the network. Furthermore, the Schneider Electric PowerChute Network Shutdown software is fully integrated with the HyperFlex Data Platform APIs to orchestrate orderly shut-down and power-up of all infrastructure and applications running on the cluster.

The same cloud-like manageability benefits of the Cisco Intersight Invisible Cloud Witness are realized in this integrated local witness design. The solution is ready to use with no deployment or complex setup other than enabling the witness server in the NMC and then configuring the local HyperFlex cluster to use the NMC. Security patching is handled at the NMC firmware level without the need to monitor and patch a dedicated witness server on the network. Additionally, there is no extra infrastructure to invest in to host witness VMs. Each NMC can support up to 64 clusters for demanding environments with many clusters. Finally, the solution is built on a hardened, embedded operating system native to the NMC and is powered by the UPS itself, ensuring the highest levels of resiliency for a local witness. The NMC not only functions as a witness server but brings numerous additional benefits to the remote environment: power protection, orchestrated shutdown, environmental monitoring, power monitoring, alerting, and more.

This integrated solution is supported on the Schneider Electric UPS NMC3 card that operates in a wide variety of APC UPS models. Switching a HyperFlex Edge two-node cluster from using the Cisco Intersight Invisible Cloud Witness to an external local witness like the NMC3 is a simple and nondisruptive operation (Figure 14). Simply configure the IP address, username, and password from the HyperFlex Connect UI.

Note: Connectivity to Cisco Intersight is still required for cluster life-cycle management operations including upgrades. Using an external witness helps to ensure a running cluster maintains high availability and is protected against a node failure even when the WAN/network backhaul to Intersight is disconnected. This combination of local witness with centralized management provides the best of both worlds: wielding the power of Cisco Intersight for management while decoupling the data plane for maximum availability.

Newly deployed HyperFlex Edge clusters default to the Cisco Intersight Invisible Cloud Witness. Using the external local witness is an option for environments that demand the highest levels of availability.

To learn more information about this innovative integration with Schneider Electric, review the following resources:

- [APC Application Note: Network Management Card 3 Local Witness for Cisco HyperFlex 2-Node Edge Clusters](#)
- [Validate APC PowerChute Network Shutdown Integration with Cisco HyperFlex Edge Cluster White Paper](#)
- [Intersight Online Help - Configuring Device Connector](#)
- [Cisco Data Center Blog Announcement - Bridging the data center and edge dichotomy](#)

Conclusion

The Invisible Cloud Witness powered by the Cisco Intersight platform is an innovation that can significantly ease the burden of deploying and managing two-node HCI clusters at a large scale. The architecture provides numerous customer benefits compared to traditional implementations, including elimination of the need for costly hardware and the benefits of a truly invisible operational model.

As explained in the failure cases, the system is designed to be able to tolerate any single point of failure, and in some cases, many failures, while ensuring data consistency. In the end, customers can achieve the high availability they expect for their edge or ROBO locations, with full confidence that their data is well protected even under the most severe failure conditions.

For more information

<https://www.cisco.com/site/us/en/products/computing/hyperconverged-infrastructure/index.html>

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 <https://www.cisco.com/go/offices>.

Cisco and the Cisco logo are trademarks or registered trademarks of Cisco and/or its affiliates in the U.S. and other countries. To view a list of Cisco trademarks, go to this URL: <https://www.cisco.com/go/trademarks>. Third-party trademarks mentioned are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (1110R)