Disaster recovery

Enterprise applications

NetApp

December 17, 2024

Table of Contents

Disaster recovery

Disaster recovery

Enterprise databases and application infrastructures often require replication to protect from natural disaster or unexpected business disruption with minimal downtime.

The SQL Server Always-On availability group replication feature can be an excellent option, and NetApp offers options to integrate data protection with Always-On. In some cases, however, you might want to consider ONTAP replication technology. There are three basic options.

SnapMirror

SnapMirror technology offers a fast and flexible enterprise solution for replicating data over LANs and WANs. SnapMirror technology transfers only changed data blocks to the destination after the initial mirror is created, significantly reducing network bandwidth requirements. It can be configured in either synchronous or asynchronous mode.

NetApp MetroCluster and SnapMirror active sync

For many customers, DR requires more than just possessing a remote copy of data, it requires the ability to rapidly make use of that data. NetApp offers two technologies that address this need - MetroCluster and SnapMirror active sync

MetroCluster refers to ONTAP in a hardware configuration that includes low-level synchronously mirrored storage and numerous additional features. Integrated solutions such as MetroCluster simplify today's complicated, scale-out database, application, and virtualization infrastructures. It replaces multiple, external data protection products and strategies with one simple, central storage array. It also provides integrated backup, recovery, disaster recovery, and high availability (HA) within a single clustered storage system.

SnapMirror active sync is based on SnapMirror Synchronous. With MetroCluster, each ONTAP controller is responsible for replicating its drive data to a remote location. With SnapMirror active sync, you essentially have two different ONTAP systems maintaining independent copies of your LUN data, but cooperating to present a single instance of that LUN. From a host point of view, it's a single LUN entity.

SM-as and MCC comparison

SM-as and MetroCluster are similar in overall functionality, but there are important differences in the way in which RPO=0 replication was implemented and how it is managed. SnapMirror asynchronous and synchronous can also be used as part of a DR plan, but they are not designed as HA repliation technologies.

- A MetroCluster configuration is more like one integrated cluster with nodes distributed across sites. SM-as behaves like two otherwise independent clusters that are cooperating in serving up select RPO=0 synchronously replicated LUNs.
- The data in a MetroCluster configuration is only accessible from one particular site at any given time. A second copy of the data is present on the opposite site, but the data is passive. It cannot be accessed without a storage system failover.
- MetroCluster and SM-as perform mirroring occurs at different levels. MetroCluster mirroring is performed at the RAID layer. The low-level data is stored in a mirrored format using SyncMirror. The use of mirroring is virtually invisible up at the LUN, volume, and protocol layers.

- In contrast, SM-as mirroring occurs at the protocol layer. The two clusters are overall independent clusters. Once the two copies of data are in sync, the two clusters only need to mirror writes. When a write occurs on one cluster, it is replicated to the other cluster. The write is only acknowledged to the host when the write has completed on both sites. Other than this protocol splitting behavior, the two clusters are otherwise normal ONTAP clusters.
- The primary role for MetroCluster is large-scale replication. You can replicate an entire array with RPO=0 and near-zero RTO. This simplifies the failover process because there is only one "thing" to fail over, and it scales extremely well in terms of capacity and IOPS.
- One key use case for SM-as is granular replication. Sometimes you don’t want to replicate all data as a single unit, or you need to be able to selectively fail over certain workloads.
- Another key use case for SM-as is for active-active operations, where you want fully usable copies of data to be available on two different clusters located in two different locations with identical performance characteristics and, if desired, no requirement to stretch the SAN across sites. You can have your applications already running on both sites, which reduces the overall RTO during failover operations.

SnapMirror

The following are recommendations for SnapMirror for SQL Server:

- If SMB is used, the destination SVM must be a member of the same Active Directory domain of which the source SVM is a member so that the access control lists (ACLs) stored within NAS files are not broken during recovery from a disaster.
- Using destination volume names that are the same as the source volume names is not required but can make the process of mounting destination volumes into the destination simpler to manage. If SMB is used, you must make the destination NAS namespace identical in paths and directory structure to the source namespace.
- For consistency purposes, do not schedule SnapMirror updates from the controllers. Instead, enable SnapMirror updates from SnapCenter to update SnapMirror after either full or log backup is completed.
- Distribute volumes that contain SQL Server data across different nodes in the cluster to allow all cluster nodes to share SnapMirror replication activity. This distribution optimizes the use of node resources.
- Use synchronous replication where demand for quick data recovery is higher and asynchronous solutions for flexibility in RPO.

For more information about SnapMirror, see TR-4015: SnapMirror Configuration and Best Practices Guide for ONTAP 9.

MetroCluster

Architecture

Microsoft SQL Server deployment with MetroCluster environment requires some explanation of physical design of a MetroCluster system.

MetroCluster synchronously mirrors data and configuration between two ONTAP clusters in separate locations or failure domains. MetroCluster provides continuously available storage for applications by automatically managing two objectives:

- Zero recovery point objective (RPO) by synchronously mirroring data written to the cluster.

- Near zero recovery time objective (RTO) by mirroring configuration and automating access to data at the second site.

MetroCluster provides simplicity with automatic mirroring of data and configuration between the two independent clusters located in the two sites. As storage is provisioned within one cluster, it is automatically mirrored to the second cluster at the second site. NetApp SyncMirror® provides a complete copy of all data with a zero RPO. This means that workloads from one site could switch over at any time to the opposite site and continue serving data without data loss. MetroCluster manages the switchover process of providing access to NAS and SAN-provisioned data at the second site. The design of MetroCluster as a validated solution contains sizing and configuration that enables a switchover to be performed within the protocol timeout periods or sooner (typically less than 120 seconds). This results in a near zero RPO and applications can continue accessing data without incurring failures.MetroCluster is available in several variations defined by the back-end storage fabric.

MetroCluster is avialable in 3 different configurations

- HA pairs with IP connectivity
- HA pairs with FC connectivity
- Single controller with FC connectivity

> The term 'connectivity' refers to the cluster connection used for cross-site replication. It does not refer to the host protocols. All host-side protocols are supported as usual in a MetroCluster configuration irrespective of the type of connection used for inter-cluster communication.

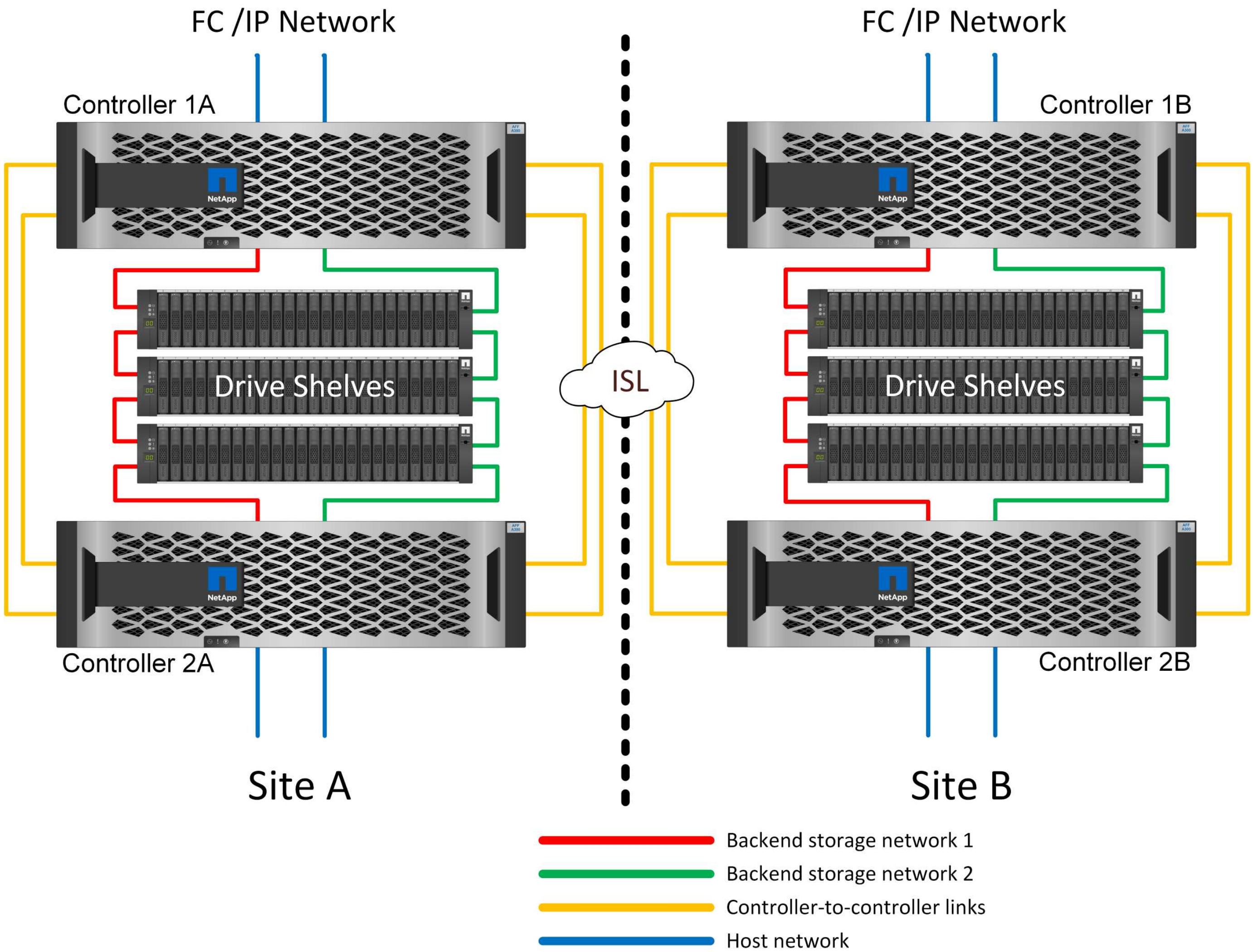
MetroCluster IP

The HA-pair MetroCluster IP configuration uses two or four nodes per site. This configuration option increases the complexity and costs relative to the two-node option, but it delivers an important benefit: intrasite redundancy. A simple controller failure does not require data access across the WAN. Data access remains local through the alternate local controller.

Most customers are choosing IP connectivity because the infrastructure requirements are simpler. In the past, high-speed cross-site connectivity was generally easier to provision using dark fibre and FC switches, but today high-speed, low latency IP circuits are more readily available.

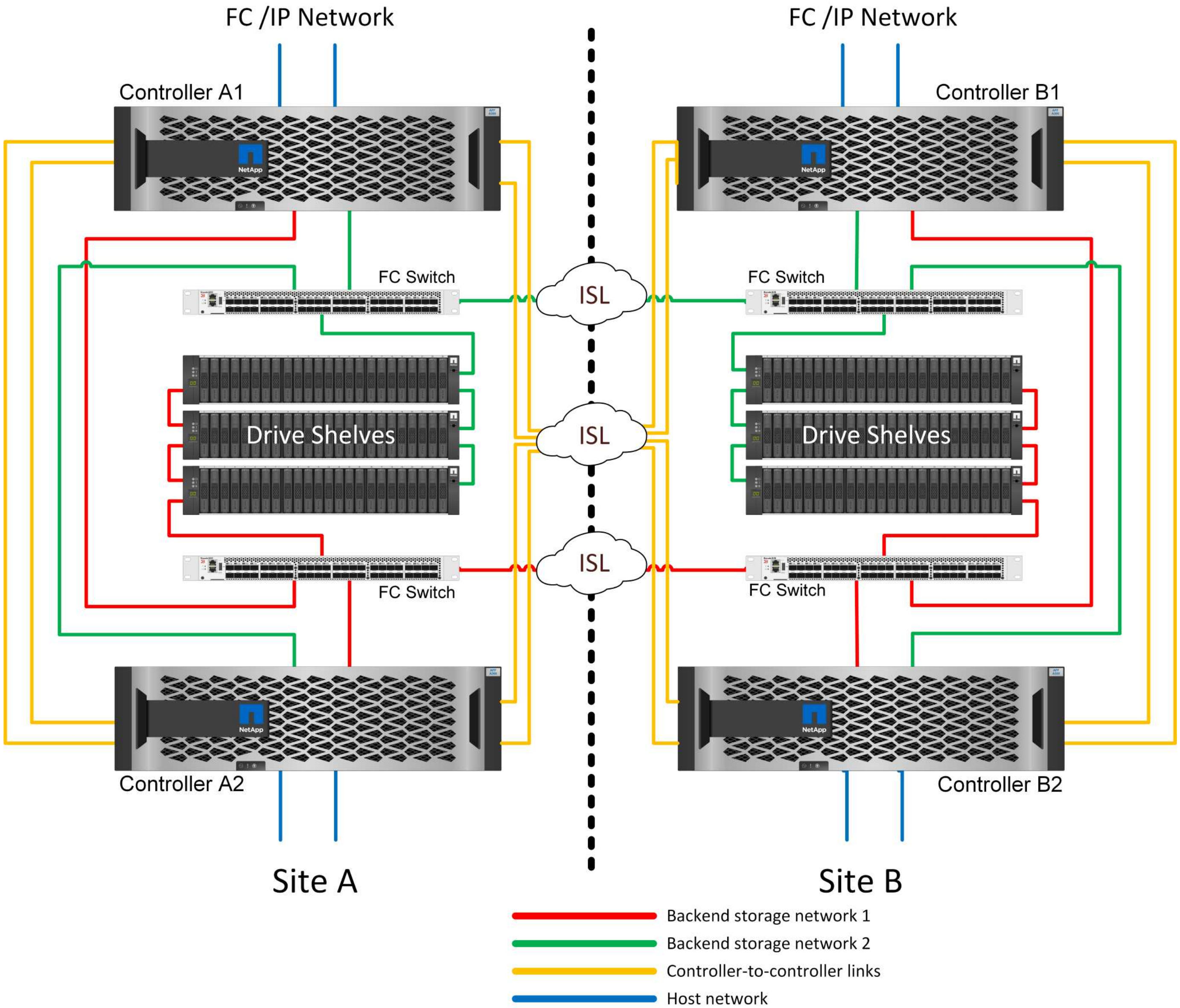
The architecture is also simpler because the only cross-site connections are for the controllers. In FC SAN attached MetroClusters, a controller writes directly to the drives on the opposite site and thus requires additional SAN connections, switches, and bridges. In contrast, a controller in an IP configuration writes to the opposite drives via the controller.

For additional information, refer to the official ONTAP documentation and MetroCluster IP Solution Architecture and Design.



HA-Pair FC SAN-attached MetroCluster

The HA-pair MetroCluster FC configuration uses two or four nodes per site. This configuration option increases the complexity and costs relative to the two-node option, but it delivers an important benefit: intrasite redundancy. A simple controller failure does not require data access across the WAN. Data access remains local through the alternate local controller.

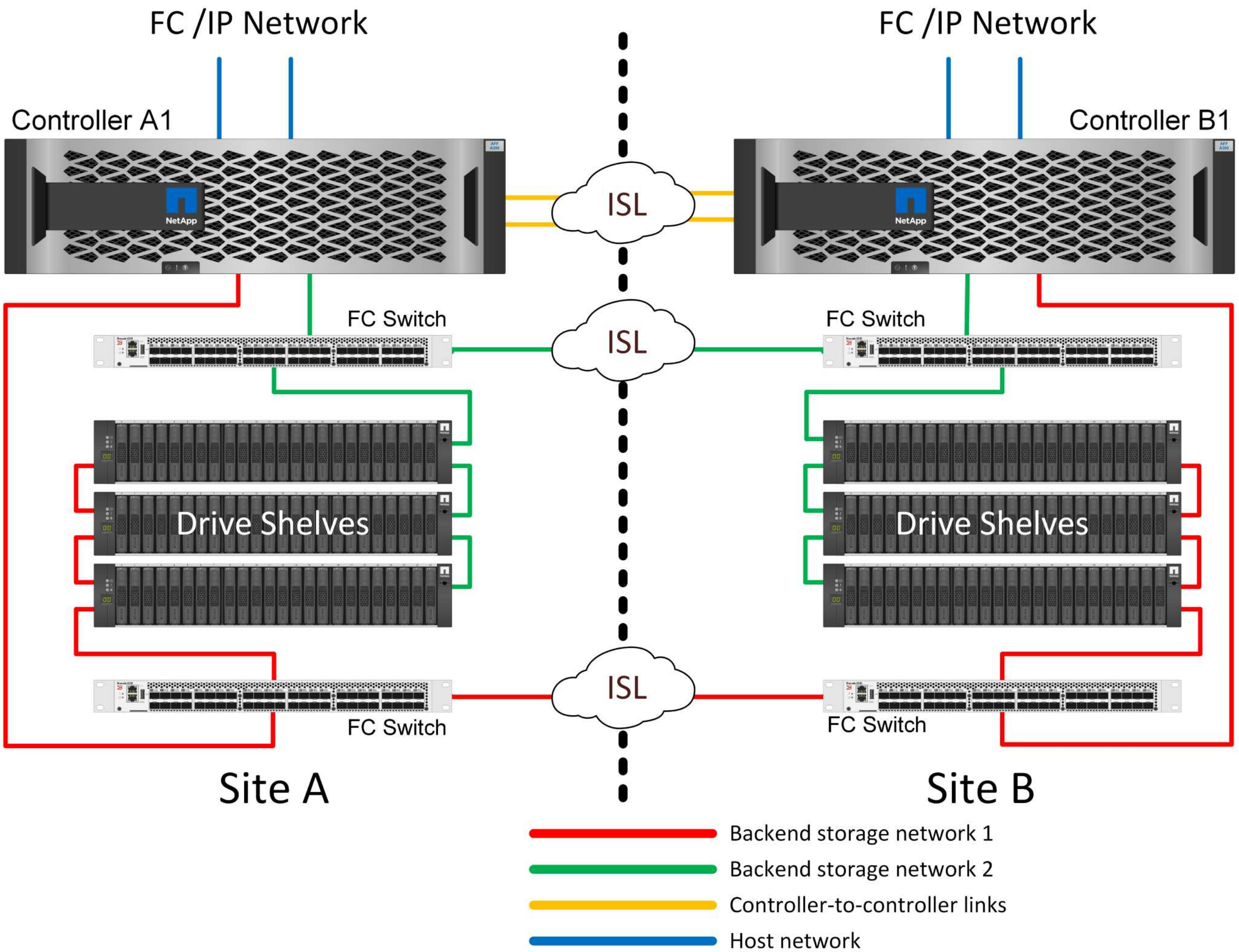


Some multisite infrastructures are not designed for active-active operations, but rather are used more as a primary site and disaster recovery site. In this situation, an HA-pair MetroCluster option is generally preferable for the following reasons:

- Although a two-node MetroCluster cluster is an HA system, unexpected failure of a controller or planned maintenance requires that data services must come online on the opposite site. If the network connectivity between sites cannot support the required bandwidth, performance is affected. The only option would be to also fail over the various host OSs and associated services to the alternate site. The HA-pair MetroCluster cluster eliminates this problem because loss of a controller results in simple failover within the same site.
- Some network topologies are not designed for cross-site access, but instead use different subnets or isolated FC SANs. In these cases, the two-node MetroCluster cluster no longer functions as an HA system because the alternate controller cannot serve data to the servers on the opposite site. The HA-pair MetroCluster option is required to deliver complete redundancy.
- If a two-site infrastructure is viewed as a single highly available infrastructure, the two-node MetroCluster configuration is suitable. However, if the system must function for an extended period of time after site failure, then an HA pair is preferred because it continues to provide HA within a single site.

Two-node FC SAN-attached MetroCluster

The two-node MetroCluster configuration uses only one node per site. This design is simpler than the HA-pair option because there are fewer components to configure and maintain. It also has reduced infrastructure demands in terms of cabling and FC switching. Finally, it reduces costs.



The obvious impact of this design is that controller failure on a single site means that data is available from the opposite site. This restriction is not necessarily a problem. Many enterprises have multisite data center operations with stretched, high-speed, low-latency networks that function essentially as a single infrastructure. In these cases, the two-node version of MetroCluster is the preferred configuration. Two-node systems are currently used at petabyte scale by several service providers.

MetroCluster resiliency features

There are no single points of failure in a MetroCluster solution:

- Each controller has two independent paths to the drive shelves on the local site.
- Each controller has two independent paths to the drive shelves on the remote site.
- Each controller has two independent paths to the controllers on the opposite site.
- In the HA-pair configuration, each controller has two paths to its local partner.

In summary, any one component in the configuration can be removed without compromising the ability of MetroCluster to serve data. The only difference in terms of resiliency between the two options is that the HA-pair version is still an overall HA storage system after a site failure.

SyncMirror

Protection for SQL Server with MetroCluster is based on SyncMirror, which gives a maximum-performance, scale-out synchronous mirroring technology.

Data protection with SyncMirror

At the simplest level, synchronous replication means any change must be made to both sides of mirrored storage before it is acknowledged. For example, if a database is writing a log, or a VMware guest is being patched, a write must never be lost. As a protocol level, the storage system must not acknowledge the write until it has been committed to nonvolatile media on both sites. Only then is it safe to proceed without the risk of data loss.

The use of a synchronous replication technology is the first step in designing and managing a synchronous replication solution. The most important consideration is understanding what could happen during various planned and unplanned failure scenarios. Not all synchronous replication solutions offer the same capabilities. If you need a solution that delivers a recovery point objective (RPO) of zero, meaning zero data loss, all failure scenarios must be considered. In particular, what is the expected result when replication is impossible due to loss of connectivity between sites?

SyncMirror data availability

MetroCluster replication is based on NetApp SyncMirror technology, which is designed to efficiently switch into and out of synchronous mode. This capability meets the requirements of customers who demand synchronous replication, but who also need high availability for their data services. For example, if connectivity to a remote site is severed, it is generally preferable to have the storage system continue operating in a non-replicated state.

Many synchronous replication solutions are only capable of operating in synchronous mode. This type of all-or-nothing replication is sometimes called domino mode. Such storage systems stop serving data rather than allowing the local and remote copies of data to become un-synchronized. If replication is forcibly broken, resynchronization can be extremely time consuming and can leave a customer exposed to complete data loss during the time that mirroring is reestablished.

Not only can SyncMirror seamlessly switch out of synchronous mode if the remote site is unreachable, it can also rapidly resync to an RPO = 0 state when connectivity is restored. The stale copy of data at the remote site can also be preserved in a usable state during resynchronization, which ensures that local and remote copies of data exist at all times.

Where domino mode is required, NetApp offers SnapMirror Synchronous (SM-S). Application-level options also exist, such as Oracle DataGuard or SQL Server Always On Availability Groups. OS-level disk mirroring can be an option. Consult your NetApp or partner account team for additional information and options.

SQL Server with MetroCluster

One option for protecting SQL Server databases with a zero RPO is MetroCluster. MetroCluster is a simple, high-performance RPO=0 replication technology that allows you to easily replicate an entire infrastructure across sites.

SQL Server can scale up to thousands of databases on a single MetroCluster system. There could be SQL

Server standalone instances or failover cluster instances, MetroCluster system does not necessarily add to or change any best practices for managing a database.

A complete explanation of MetroCluster is beyond the scope of this document, but the principles are simple. MetroCluster can provide an RPO=0 replication solution with rapid failover. What you build on top of this foundation depends on your requirements.

For example, a basic rapid DR procedure after sudden site loss could use the following basic steps:

- Force a MetroCluster switchover
- Performing discovery of FC/iSCSI LUNs (SAN only)
- Mount file systems
- Start SQL Services

The primary requirement of this approach is a running OS in place on the remote site. It must be preconfigured with SQL Server setup and should be updated with equivalent build version. SQL Server system databases can also be mirrored to the remote site and mounted if a disaster is declared.

If the volumes, file systems and datastore hosting virtualized databases are not in use at the disaster recovery site prior to the switchover, there is no requirement to set `dr-force- nvfail` on associated volumes.

SnapMirror active sync

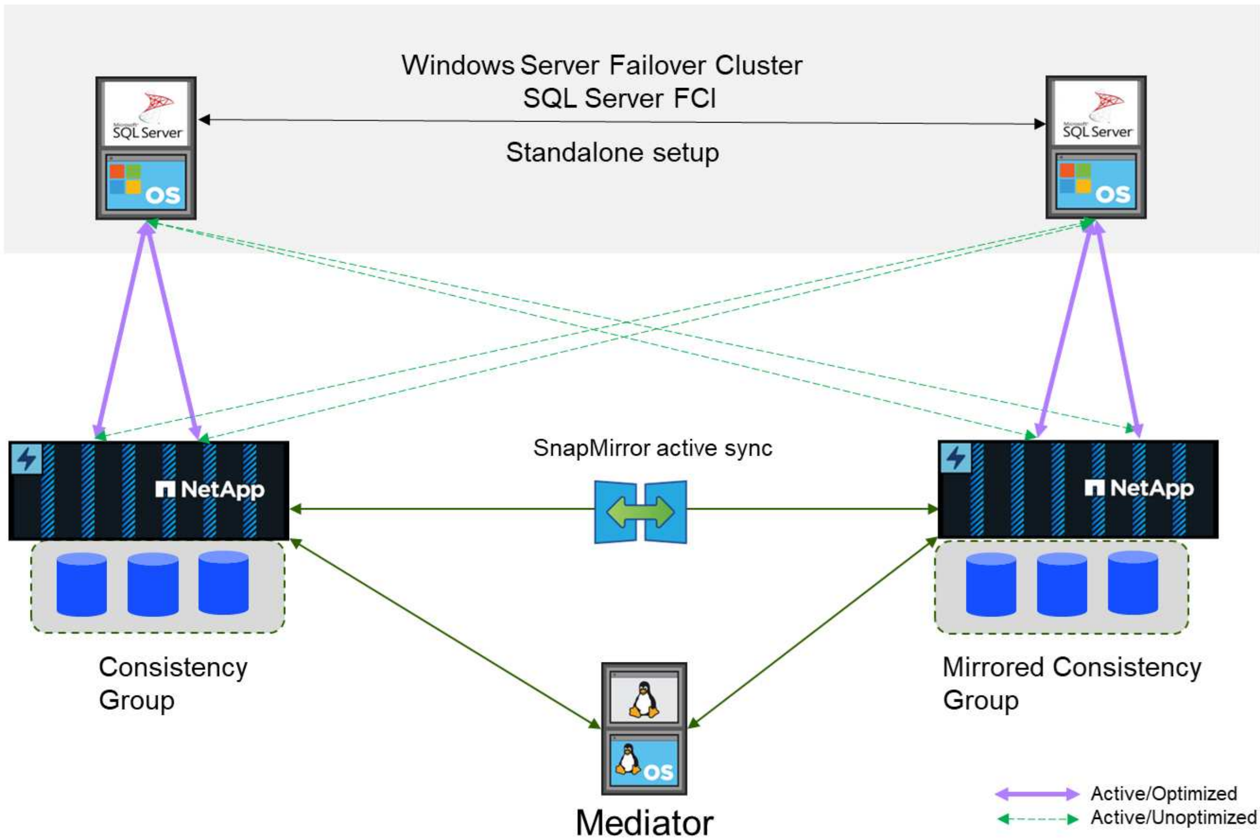
Overview

SnapMirror active sync enables individual SQL Server databases and applications to continue operations during storage and network disruptions, with transparent storage failover without any manual intervention.

Starting ONTAP 9.15.1, SnapMirror active sync supports symmetric active/active architecture in addition to the existing asymmetric configuration. Symmetric active/active capability provides synchronous bi-directional replication for business continuity and disaster recovery. It helps you protect your data access for critical SAN workloads with simultaneous read and write access to data across multiple failure domains, ensuring uninterrupted operations and minimizing downtime during disasters or system failures.

SQL Server hosts access storage using either Fiber Channel(FC) or iSCSI LUNs. Replication between each cluster hosting a copy of the replicated data. Since this feature is storage level replication, SQL Server instances running on standalone host or failover cluster instances can perform read/write operations either cluster. For planning and configuration steps, refer ONTAP documentation on SnapMirror active sync .

Architecture of SnapMirror active with symmetric active/active



Synchronous replication

In normal operation, each copy is an RPO=0 synchronous replica at all times, with one exception. If data cannot be replicated, ONTAP will release the requirement to replicate data and resume serving IO on one site while the LUNs on the other site are taken offline.

Storage hardware

Unlike other storage disaster recovery solutions, SnapMirror active sync offers asymmetric platform flexibility. The hardware at each site does not need to be identical. This capability allows you to right-size the hardware used to support SnapMirror active sync. The remote storage system can be identical to the primary site if it needs to support a full production workload, but if a disaster results in reduced I/O, than a smaller system at the remote site might be more cost-effective.

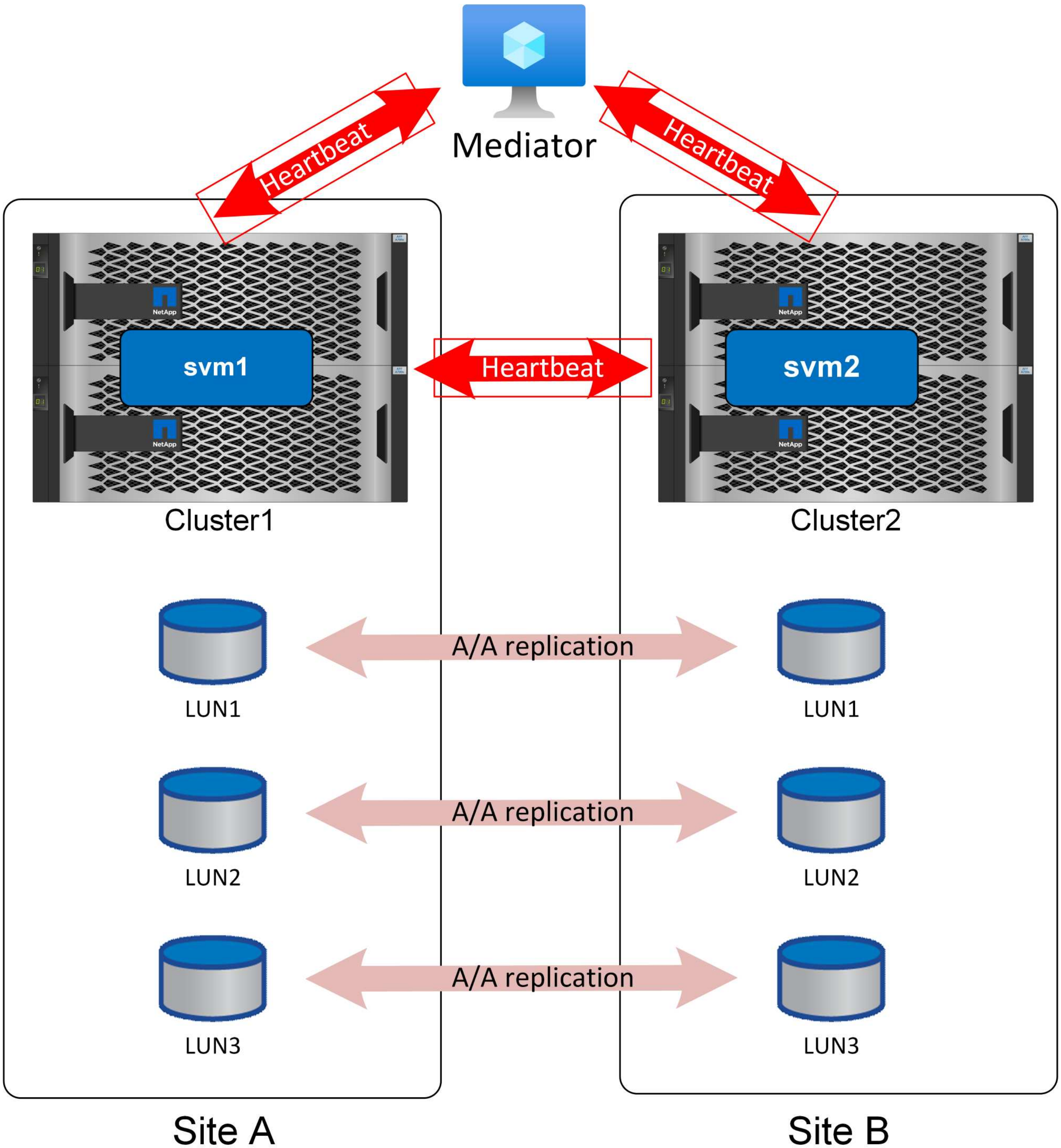
ONTAP mediator

The ONTAP Mediator is a software application that is downloaded from NetApp support, and is typically deployed on a small virtual machine. The ONTAP Mediator is not a tiebreaker. It is an alternate communication channel for the two clusters that participate in SnapMirror active sync replication. Automated operations are driven by ONTAP based on the responses received from the partner via direct connections and via the mediator.

ONTAP mediator

The mediator is required for safely automating failover. Ideally, it would be placed on an independent 3rd site, but it can still function for most needs if colocated with one of the clusters participating in replication.

The mediator is not really a tiebreaker, although that is effectively the function it provides. It does not take any actions; instead it provides an alternate communication channel for cluster to cluster communication.



The #1 challenge with automated failover is the split-brain problem, and that problem arises if your two sites lose connectivity with each other. What should happen? You do not want to have two different sites designate themselves as the surviving copies of the data, but how can a single site tell the difference between actual loss of the opposite site and an inability to communicate with the opposite site?

This is where the mediator enters the picture. If placed on a 3rd site, and each site has a separate network connection to that site, then you have an additional path for each site to validate the health of the other. Look at the picture above again and consider the following scenarios.

- What happens if the mediator fails or is unreachable from one or both sites?
  - The two clusters can still communicate with each other over the same link used for replication services.
  - Data is still served with RPO=0 protection
- What happens if Site A fails?
  - Site B will see both of the communication channels go down.
  - Site B will take over data services, but without RPO=0 mirroring
- What happens if Site B fails?
  - Site A will see both of the communication channels go down.
  - Site A will take over data services, but without RPO=0 mirroring

There is one other scenario to consider: Loss of the data replication link. If the replication link between sites is lost, RPO=0 mirroring will obviously be impossible. What should happen then?

This is controlled by the preferred site status. In an SM-as relationship, one of the sites is secondary to the other. This has no effect on normal operations, and all data access is symmetric, but if replication is interrupted then the tie will have to be broken to resume operations. The result is the preferred site will continue operations without mirroring and the secondary site will halt IO processing until replication communication is restored.

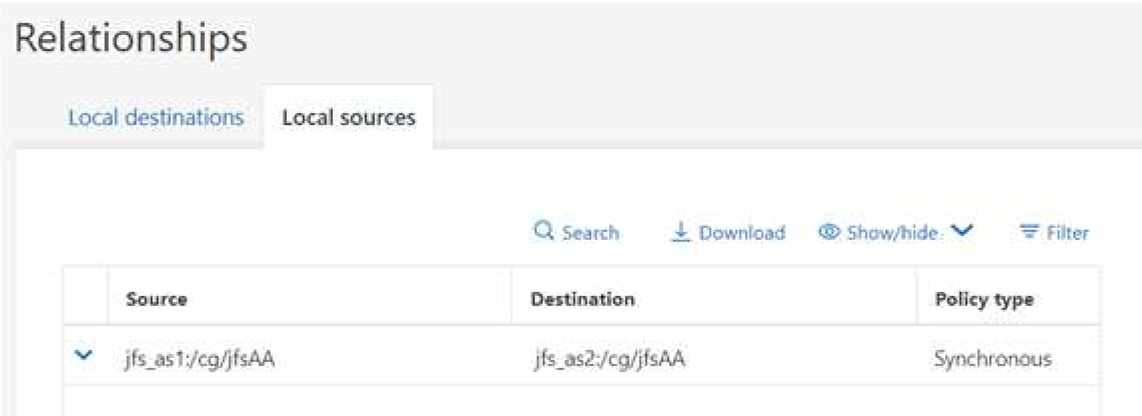
Preferred site

SnapMirror active sync behavior is symmetric, with one important exception - preferred site configuration.

SnapMirror active sync will consider one site the "source" and the other the "destination". This implies a one-way replication relationship, but this does not apply to IO behavior. Replication is bidirectional and symmetric and IO response times are the same on either side of the mirror.

The `source` designation is controls the preferred site. If the replication link is lost, the LUN paths on the source copy will continue to serve data while the LUN paths on the destination copy will become unavailable until replication is reestablished and SnapMirror reenters a synchronous state. The paths will then resume serving data.

The sourced/destination configuration can be viewed via SystemManager:



or at the CLI:

```
Cluster2::> snapmirror show -destination-path jfs_as2:/cg/jfsAA

                            Source Path: jfs_as1:/cg/jfsAA
                       Destination Path: jfs_as2:/cg/jfsAA
                      Relationship Type: XDP
                Relationship Group Type: consistencygroup
                    SnapMirror Schedule: -
                 SnapMirror Policy Type: automated-failover-duplex
                      SnapMirror Policy: AutomatedFailOverDuplex
                            Tries Limit: -
                      Throttle (KB/sec): -
                           Mirror State: Snapmirrored
                    Relationship Status: InSync
```

The key is that the source is the SVM on cluster1. As mentioned above, the terms "source" and "destination" don’t describe the flow of replicated data. Both sites can process a write and replicate it to the opposite site. In effect, both clusters are sources and destinations. The effect of designating one cluster as a source simply controls which cluster survives as a read-write storage system if the replication link is lost.

Network topology

Uniform access

Uniform access networking means hosts are able to access paths on both sites (or failure domains within the same site).

An important feature of SM-as is the ability to configure the storage systems to know where the hosts are located. When you map the LUNs to a given host, you can indicate whether or not they are proximal to a given storage system.

Proximity settings

Proximity refers to a per-cluster configuration that indicates a particular host WWN or iSCSI initiator ID belongs to a local host. It is a second, optional step for configuring LUN access.

The first step is the usual igroup configuration. Each LUN must be mapped to an igroup that contains the WWN/iSCSi IDs of the hosts that need access to that LUN. This controls which host has *access* to a LUN.

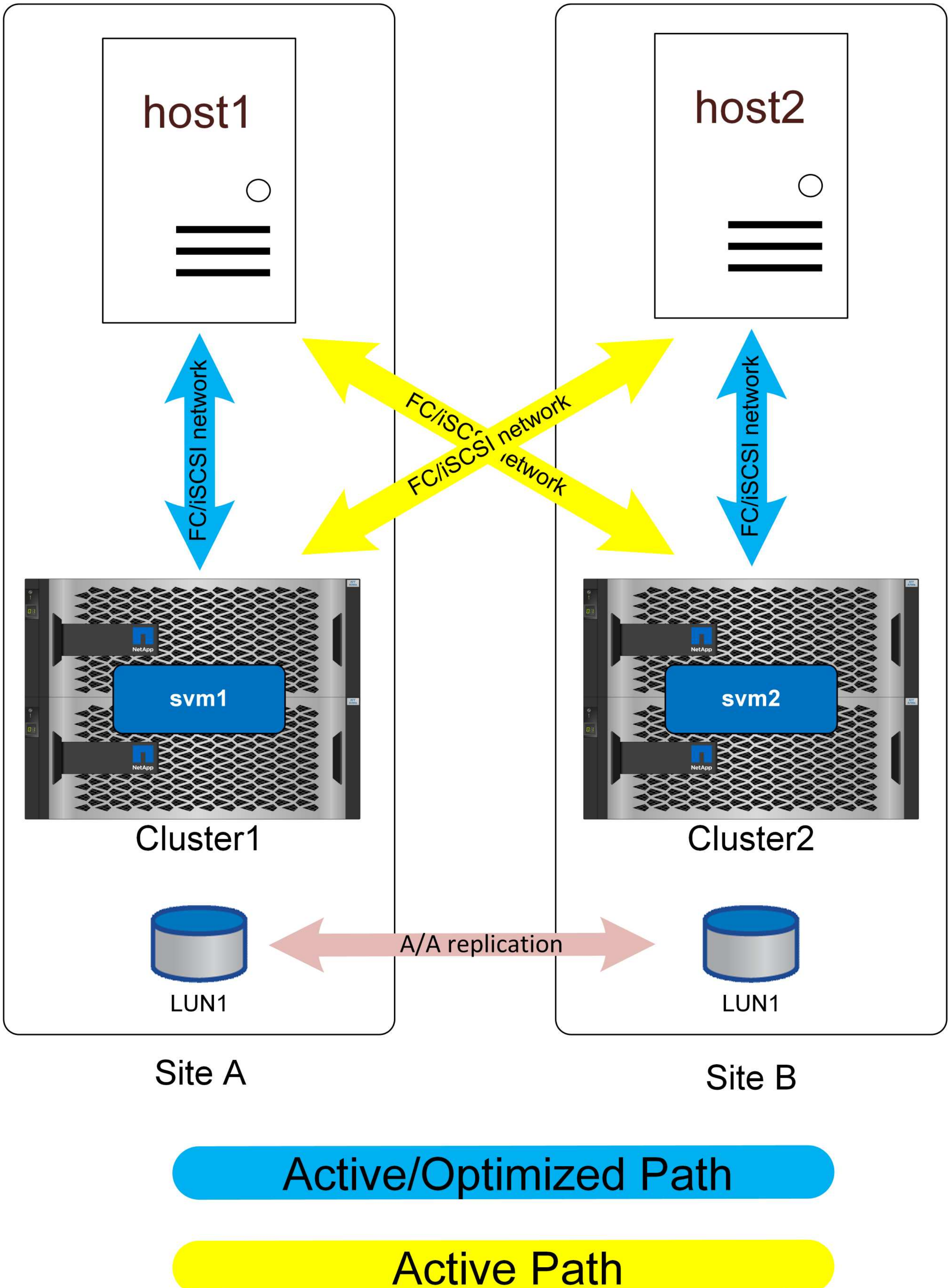
The second, optional step is to configure host proximity. This does not control access, it controls *priority*.

For example, a host at site A might be configured to access a LUN that is protected by SnapMirror active sync, and since the SAN is extended across sites, paths are available to that LUN using storage on site A or storage on site B.

Without proximity settings, that host will use both storage systems equally because both storage systems will advertise active/optimized paths. If the SAN latency and/or bandwidth between sites is limited, this may not be desireable, and you may wish to ensure that during normal operation each host preferentially uses paths to the local storage system. This is configured by adding the host WWN/iSCSI ID to the local cluster as a proximal host. This can be done at the CLI or SystemManager.

AFF

With an AFF system, the paths would appear as shown below when host proximity has been configured.


host1
host2
FC/iSCSI network
FC/iSCSI network
FC/iSCSI network
FC/iSCSI network
svm1
svm2
Cluster1
Cluster2
A/A replication
LUN1
LUN1
Site A
Site B
Active/Optimized Path
Active Path

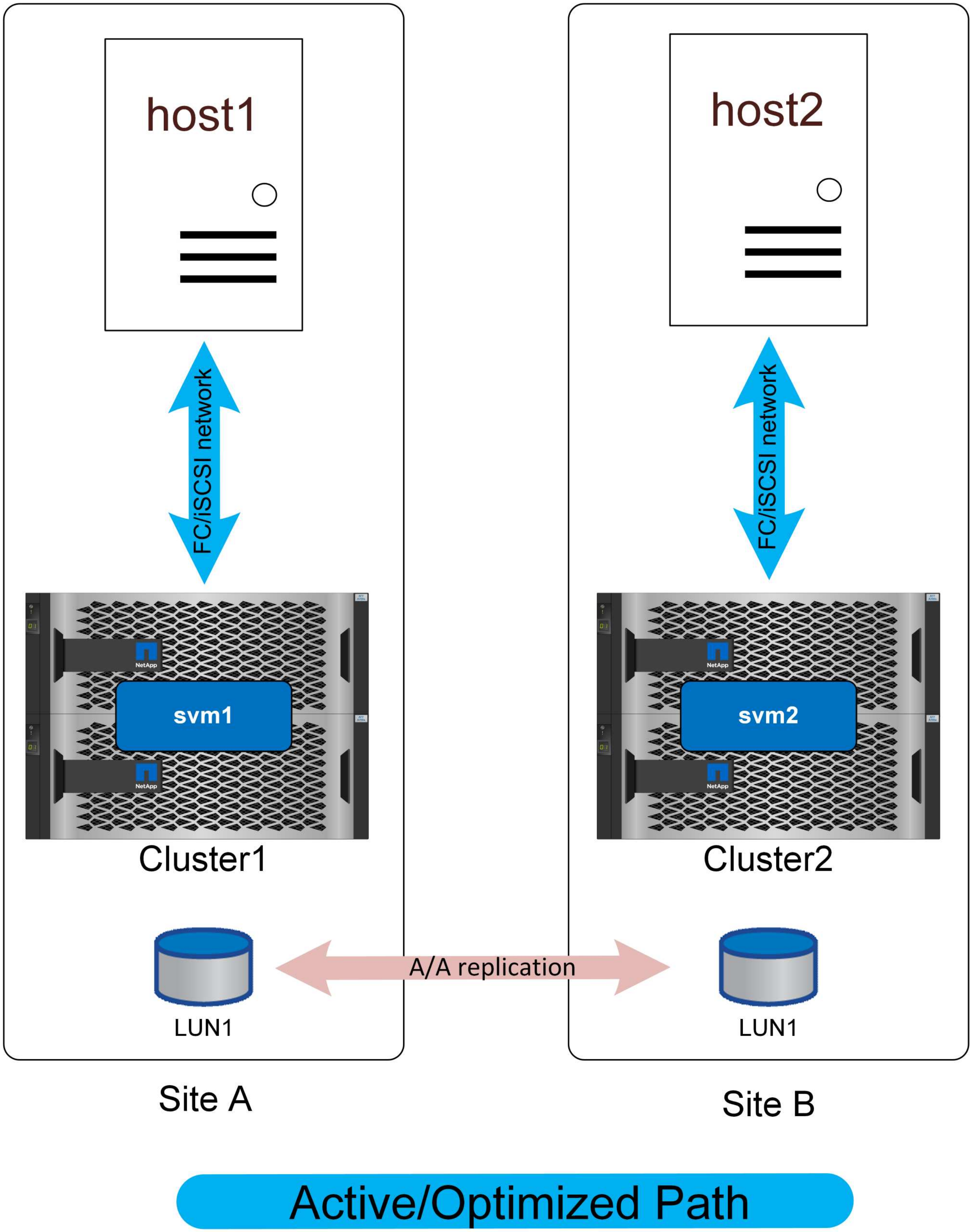
In normal operation, all IO is local IO. Reads and writes are serviced from the local storage array. Write IO will, of course, also need to be replicated by the local controller to the remote system before being acknowledged, but all read IO will be serviced locally and will not incur extra latency by traversing the SAN link between sites.

The only time the nonoptimized paths will be used is when all active/optimized paths are lost. For example, if the entire array on site A lost power, the hosts at site A would still be able to access paths to the array on site B and therefore remain operational, although they would be experiencing higher latency.

There are redundant paths through the local cluster that are not shown on these diagrams for the sake of simplicity. ONTAP storage systems are HA themselves, so a controller failure should not result in site failure. It should merely result in a change in which local paths are used on the affected site.

ASA

NetApp ASA systems offer active-active multipathing across all paths on a cluster. This also applies to SM-as configurations.



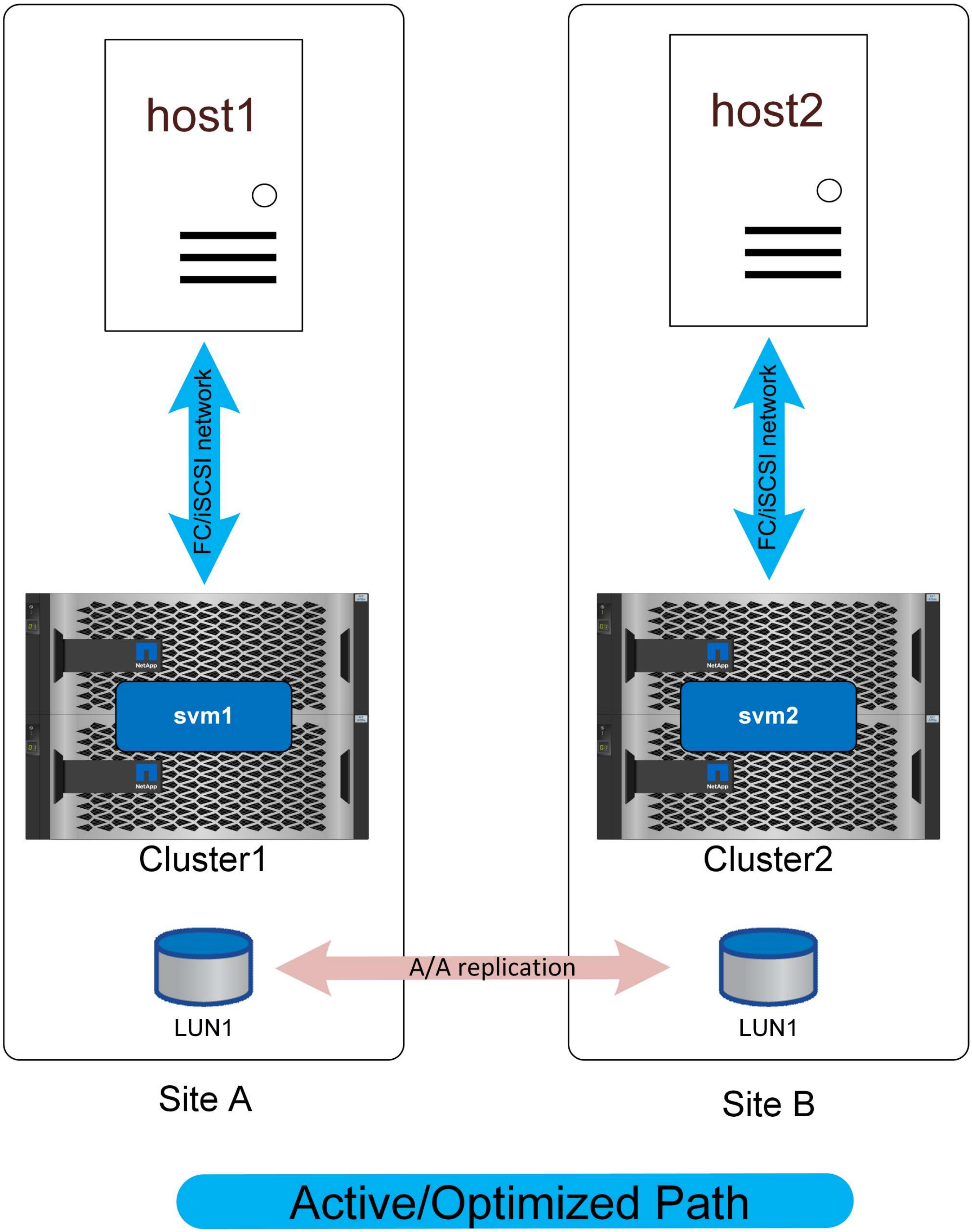
An ASA configuration with non-uniform access would work largely the same as it would with AFF. With uniform access, IO would be crossing the WAN. This may or may not be desirable.

If the two sites were 100 meters apart with fiber connectivity there should be no detectable additional latency crossing the WAN, but if the sites were a long distance apart then read performance would suffer on both sites. In contrast, with AFF those WAN-crossing paths would only be used if there were no local paths available and day-to-day performance would be better because all IO would be local IO. ASA with nonuniform access network would be an option to gain the cost and feature benefits of ASA without incurring a cross-site latency access penalty.

ASA with SM-as in a low-latency configuration offers two interesting benefits. First, it essentially **doubles** the performance for any single host because IO can be serviced by twice as many controllers using twice as many paths. Second, in a single-site environment it offers extreme availability because an entire storage system could be lost without interrupting host access.

Nonuniform access

Nonuniform access networking means each host only has access to ports on the local storage system. The SAN is not extended across sites (or failure domains within the same site).



The primary benefit to this approach is SAN simplicity - you remove the need to stretch a SAN over the network. Some customers don’t have sufficiently low-latency connectivity between sites or lack the infrastructure to tunnel FC SAN traffic over an intersite network.

The disadvantage to nonuniform access is that certain failure scenarios, including loss of the replication link, will result some hosts losing access to storage. Applications that run as single instances, such as a non-clustered database that is inherently only running on a single host at any given mount would fail if local storage connectivity was lost. The data would still be protected, but the database server would no longer have access. It would need to be restarted on a remote site, preferably through an automated process. For example, VMware HA can detect an all-paths-down situation on one server and restart a VM on another server where paths are available.

In contrast, a clustered application such as Oracle RAC can deliver a service that is simultaneously available at two different sites. Losing a site doesn't mean loss of the application service as a whole. Instances are still available and running at the surviving site.

In many cases, the additional latency overhead of an application accessing storage across a site-to-site link would be unacceptable. This means that the improved availability of uniform networking is minimal, since loss of storage on a site would lead to the need to shut down services on that failed site anyway.

There are redundant paths through the local cluster that are not shown on these diagrams for the sake of simplicity. ONTAP storage systems are HA themselves, so a controller failure should not result in site failure. It should merely result in a change in which local paths are used on the affected site.

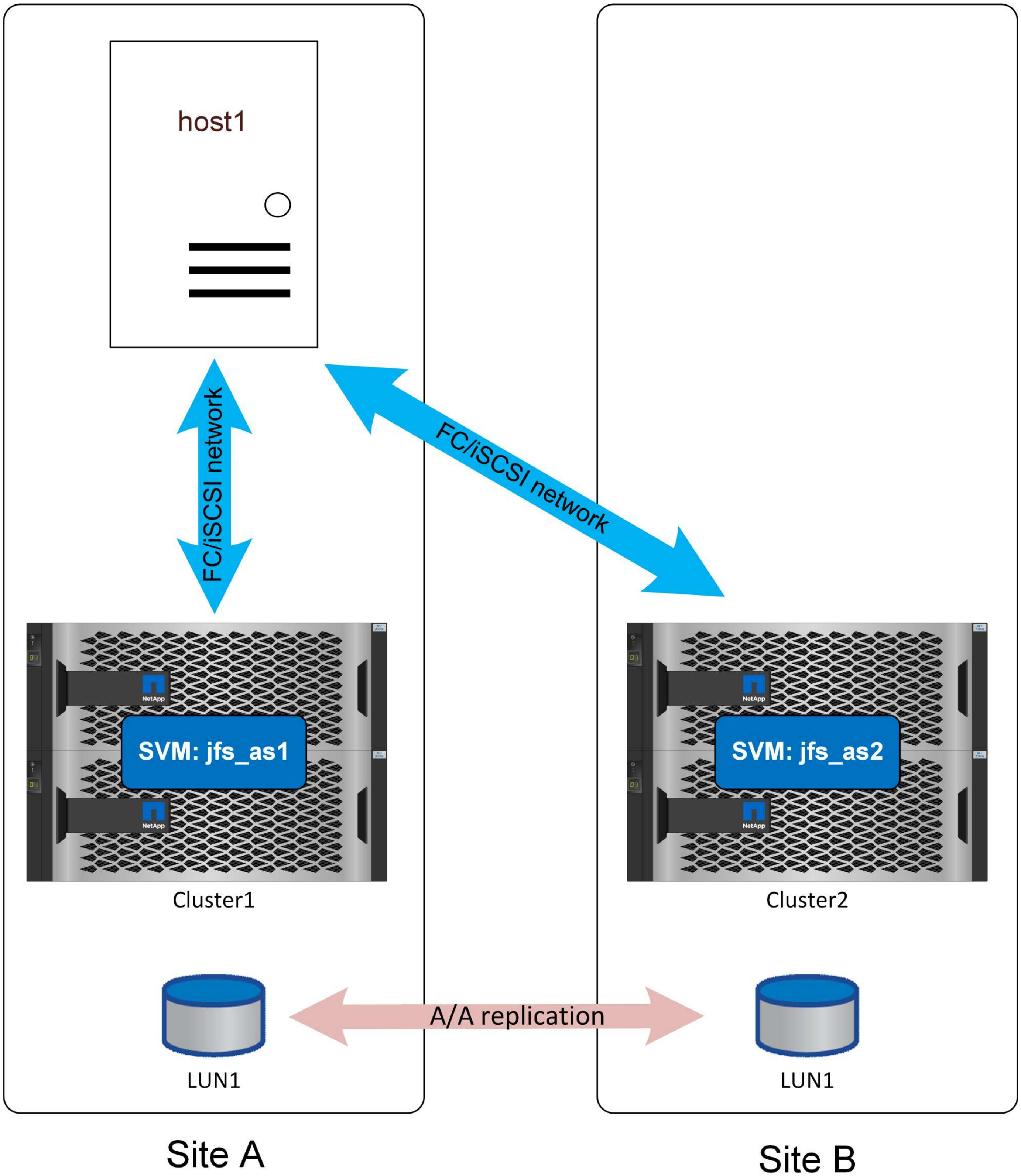
Overview

SQL Server can be configured to work with SnapMirror active sync in several ways. The right answer depends on the available network connectivity, RPO requirements, and availability requirements.

Standalone instance of SQL Server

The best practices for file layout and server configuration are the same as recommended in SQL Server on ONTAP documentation.

With a standalone setup, SQL Server could be running only at one site. Presumably uniform access would be used.



With uniform access, a storage failure at either site would not interrupt database operations. A complete site failure on the site that included the database server would, of course, result in an outage.

Some customers could configure an OS running at the remote site with a preconfigured SQL Server setup, updated with an equivalent build version as that of the production instance. Failover would require activating that standalone instance of SQL Server at the alternate site, discovering the LUNS, and and starting the database. The complete process can be automated with the Windows Powershell cmdlet as no operations are required from the storage side.

Nonuniform access could also be used, but the result would be a database outage if the storage system where the database server was located had failed because the database would have no available paths to storage. This still may be acceptable in some cases. SnapMirror active sync would still be providing RPO=0 data protection, and, in the event of site failure, the surviving copy would be active and ready to resume operations using the same procedure used with uniform access as described above.

A simple and automated failover process can be more more easily configured with the use of a virtualize host. For example, if SQL Server data files are synchronously replicated to secondary storage along with a boot VMDK, then, in the event of a disaster, the complete environment could be activated at the alternate site. An administrator could either manually activate the host at the surviving site, or automate the process through a service such as VMware HA.

SQL Server failover cluster instance

SQL Server failover instances could be also hosted on a Windows failover cluster running on a physical server or virtual server as the guest operating system. This multi-host architecture provides SQL Server instance and storage resiliency. Such deployment is helpful in high-demand environments seeking robust failover processes while maintaining enhanced performance. In a failover cluster setup, when a host or primary storage is affected then SQL Services will be failover to the secondary host, and at the same time, secondary storage will be available to serve IO. No automation script or administrator intervention is required.

Failure scenarios

Planning a complete SnapMirror active sync application architecture requires understanding how SM-as will respond in various planned and unplanned failover scenarios.

For the following examples, assume that site A is configured as the preferred site.

Loss of replication connectivity

If SM-as replication is interrupted, write IO cannot be completed because it would be impossible for a cluster to replicate changes to the opposite site.

Site A (Preferred site)

The result of replication link failure on the preferred site will be an approximate 15 second pause in write IO processing as ONTAP retries replicated write operations before it determines that the replication link is genuinely unreachable. After the 15 seconds elapses, the site A system resumes read and write IO processing. The SAN paths will not change, and the LUNs will remain online.

Site B

Since site B is not the SnapMirror active sync preferred site, its LUN paths will become unavailable after about 15 seconds.

Storage system failure

The result of a storage system failure is nearly identical to the result of losing the replication link. The surviving site should experience a roughly 15 second IO pause. Once that 15 second period elapses, IO will resume on that site as usual.

Loss of the mediator

The mediator service does not directly control storage operations. It functions as an alternate control path between clusters. It exists primarily to automate failover without the risk of a split-brain scenario. In normal operation, each cluster is replicating changes to its partner, and each cluster therefore can verify that the partner cluster is online and serving data. If the replication link failed, replication would cease.

The reason a mediator is required for safe automated failover is because it would otherwise be impossible for a storage cluster to be able to determine whether loss of bidirectional communication was the result of a network outage or actual storage failure.

The mediator provides an alternate path for each cluster to verify the health of its partner. The scenarios are as follows:

- If a cluster can contact its partner directly, replication services are operational. No action required.
- If a preferred site cannot contact its partner directly or via the mediator, it will assume the partner is either actually unavailable or was isolated and has taken its LUN paths offline. The preferred site will then proceed to release the RPO=0 state and continue processing both read and write IO.
- If a non-preferred site cannot contact its partner directly, but can contact it via the mediator, it will take its paths offline and await the return of the replication connection.
- If a non-preferred site cannot contact its partner directly or via an operational mediator, it will assume the partner is either actually unavailable or was isolated and has taken its LUN paths offline. The non-preferred site will then proceed to release the RPO=0 state and continue processing both read and write IO. It will assume the role of the replication source and will become the new preferred site.

If the mediator is wholly unavailable:

- Failure of replication services for any reason, including failure of the nonpreferred site or storage system, will result in the preferred site releasing the RPO=0 state and resuming read and write IO processing. The non-preferred site will take its paths offline.
- Failure of the preferred site will result in an outage because the non-preferred site will be unable to verify that the opposite site is truly offline and therefore it would not be safe for the nonpreferred site to resume services.

Restoring services

After a failure is resolved, such as restoring site-to-site connectivity or powering on a failed system, the SnapMirror active sync endpoints will automatically detect the presence of a faulty replication relationship and bring it back to an RPO=0 state. Once synchronous replication is reestablished, the failed paths will come online again.

In many cases, clustered applications will automatically detect the return of failed paths, and those applications will also come back online. In other cases, a host-level SAN scan may be required, or applications may need to be brought back online manually. It depends on the application and how it is configured, and in general such tasks can be easily automated. ONTAP itself is self-healing and should not require any user intervention to resume RPO=0 storage operations.

Manual failover

Changing the preferred site requires a simple operation. IO will pause for a second or two as authority over replication behavior switches between clusters, but IO is otherwise unaffected.