

## WHITE PAPER | Uninterrupted Access to Cluster Shared Volumes (CSVs) Synchronously Mirrored Across Metropolitan Hot Sites

*Failover Cluster support in Windows Server 2008 R2 with Hyper-V provides a powerful mechanism to minimize the effects of planned and unplanned server downtime. It coordinates live migrations and failover of workloads between servers through a Cluster Shared Volume (CSV). The health of the cluster depends on maintaining continuous access to the CSV and the shared disk on which it resides.*

In this paper you will learn how DataCore Software solves a longstanding stumbling block to clustered systems spread across metropolitan sites by providing uninterrupted access to the CSV despite the many technical and environmental conditions that conspire to disrupt it.

### **CSV shared disk exposes cluster to single point of failure & disruption**

A glaring shortcoming with many cluster configurations comes from their dependence on a single, shared disk subsystem to house their CSV. Despite standard precautions like RAID sets and internally redundant controllers, a single “bullet proof” storage subsystem cannot be solely counted on during:

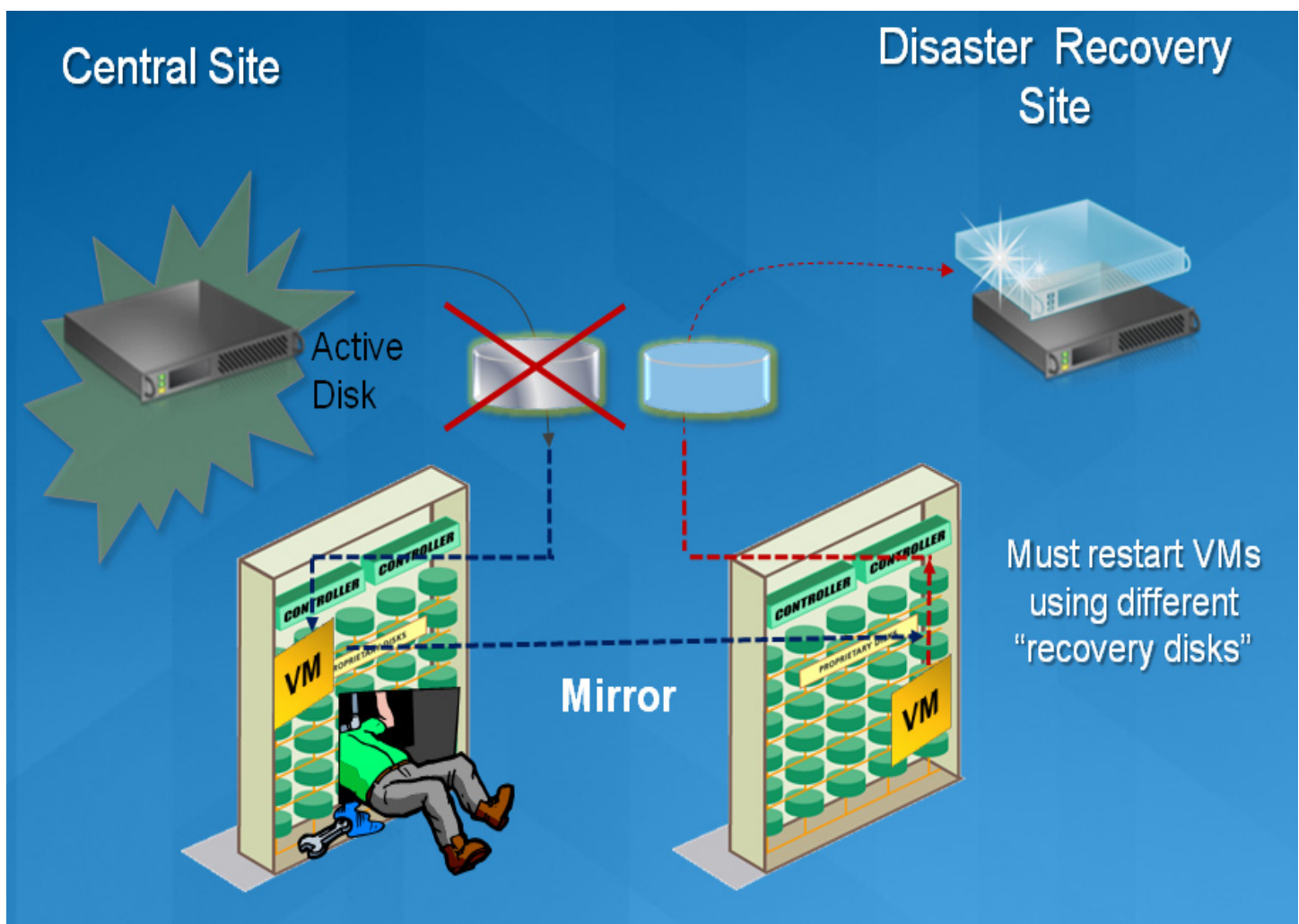
- a. Planned maintenance (major firmware upgrades, major hardware repairs)
- b. Facility problems (power failure, loss of air conditioning, water leak)
- c. Technician error (unintentional power down)
- d. Site-wide disaster

## Synchronous replication alone doesn't prevent downtime

Some IT organizations seeking protection against a single point of failure go to great expense to synchronously mirror the CSV between a primary online disk at one location and an offline mirrored replica at a different location, while spreading servers across sites. When one site's storage is taken out-of-service, they use special scripts to remap the offline replica to the clustered servers. While this precaution protects against data loss, it does not prevent the cluster from going down and taking with it numerous applications (Figure 1).

Simply put, the process of cutting over to the offline replica disrupts cluster operations causing workloads to suffer downtime from the moment access to the primary copy of the CSV is lost. Only after the offline replica is brought online and mapped to the surviving servers can the environment be restarted. For this reason, the procedure is considered a Disaster Recovery alternative.

Figure 1. Application Downtime

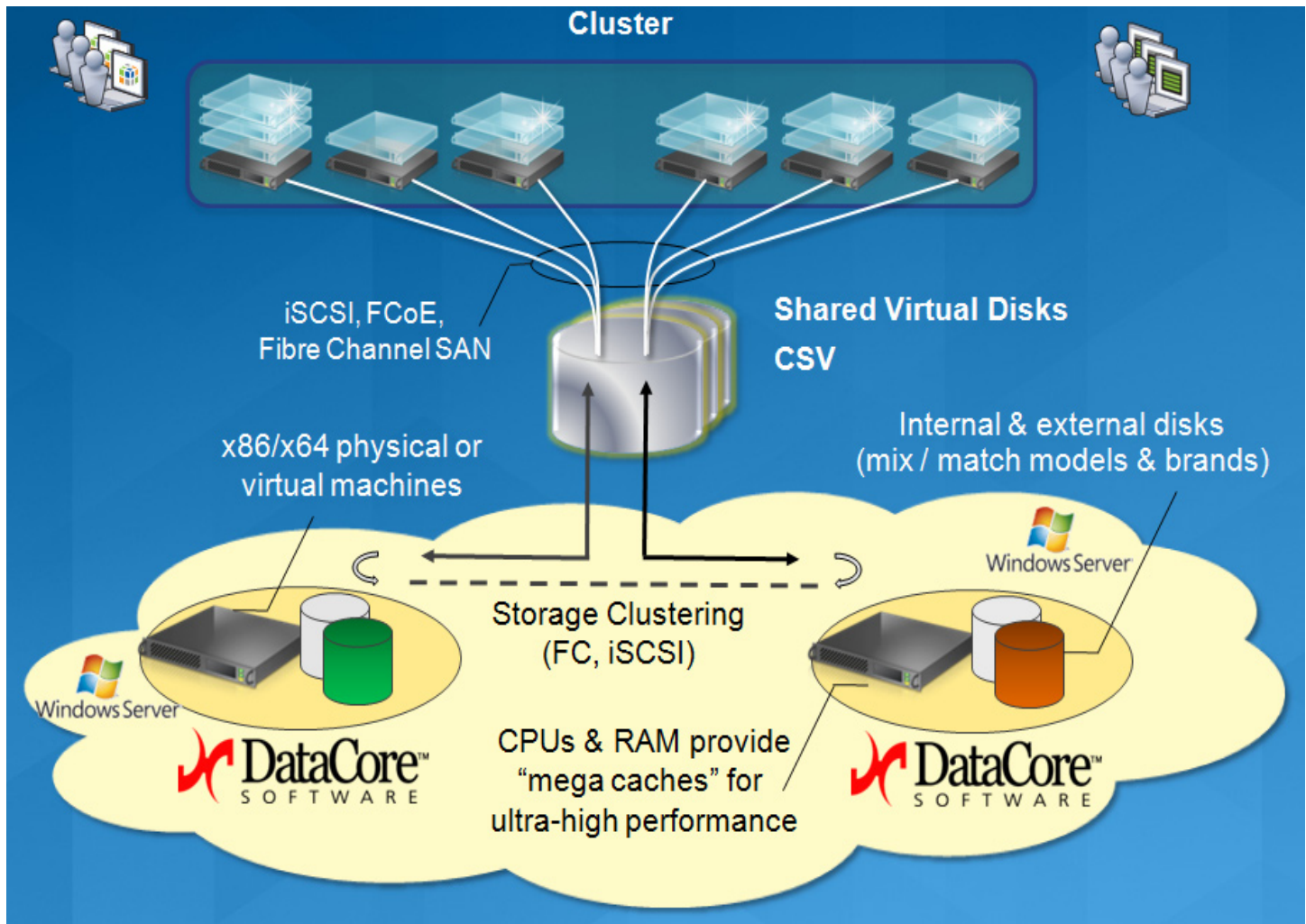


Switching workloads non-disruptively to another site when the main site's storage device goes down proves to be a far more elusive, business continuity objective.

## Non-stop Access to CSV from DataCore Software

DataCore Software plays a unique role in Microsoft Failover Clusters configurations by providing uninterrupted access to CSVs that are synchronously mirrored between two sites.

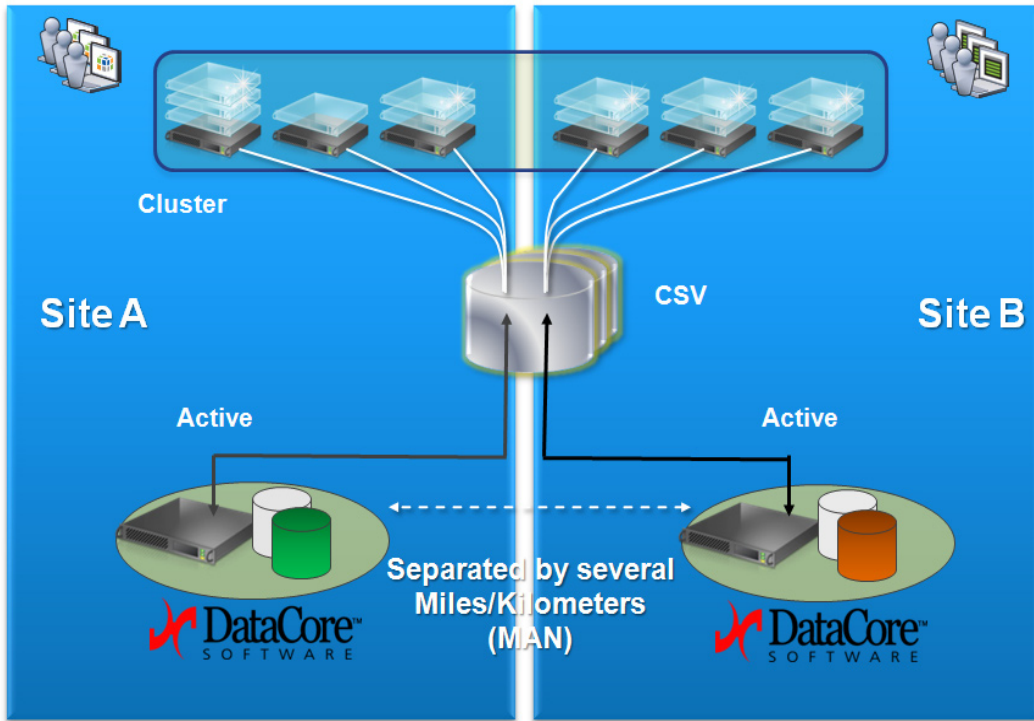
**Figure 2. Conceptual view of CSV Shared disk provided by DataCore Software**



Like other replication technologies, DataCore storage virtualization software synchronously mirrors CSV updates to separate physical disks stored at each location, but with the following important distinctions:

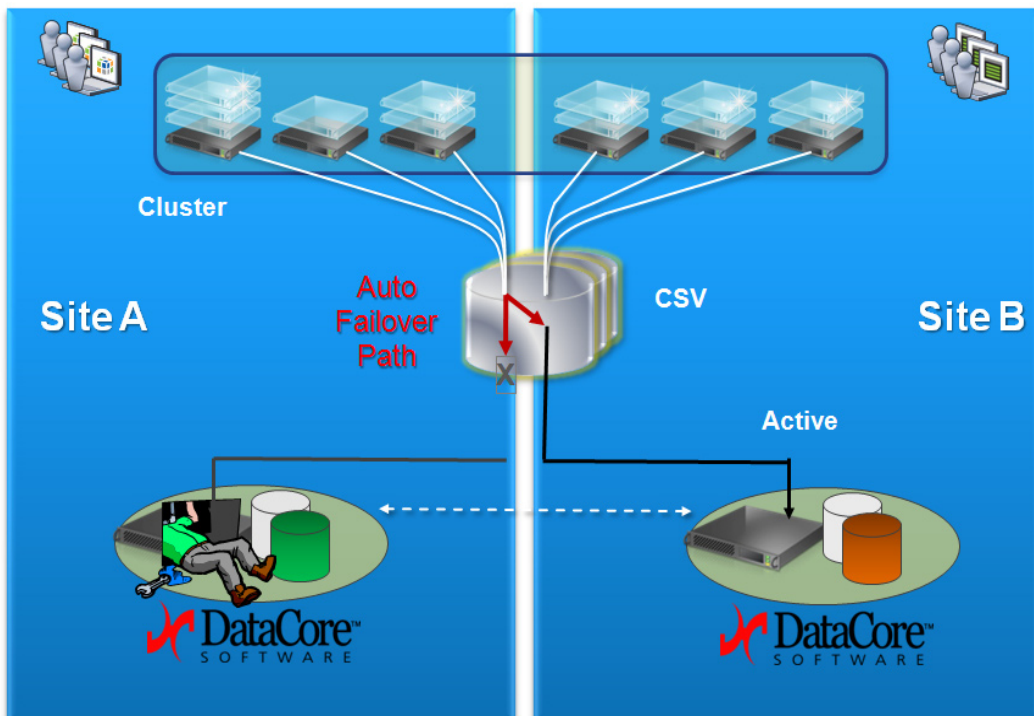
1. The multi-site mirrors behave as a single virtual shared disk from the cluster's perspective. Each cluster node designates a preferred path and an alternate path to the virtual disk on which the CSV is stored using a standard Multi-path I/O (MPIO) driver. The preferred path routes I/O requests to the mirror image closest to the node (at the same site), while the alternate path directs requests to the mirror image at the other site.

**Figure 3. Single, shared disk view of CSV spread across metropolitan Hot-Hot sites**



2. When the disk or disk subsystem housing the CSV at one site is taken out-of-service, the clustered node's MPIIO driver will merely receive an error on the preferred path and automatically fail-over to the alternate path. The mirror image of the virtual disks at the other site transparently fields the I/O request without hesitation. Neither the cluster nor its applications are aware of the automatic redirection. There is no downtime waiting for an offline copy to go online and be remapped at the other location.

**Figure 4. Transparent switchover to CSV mirror at Hot site**



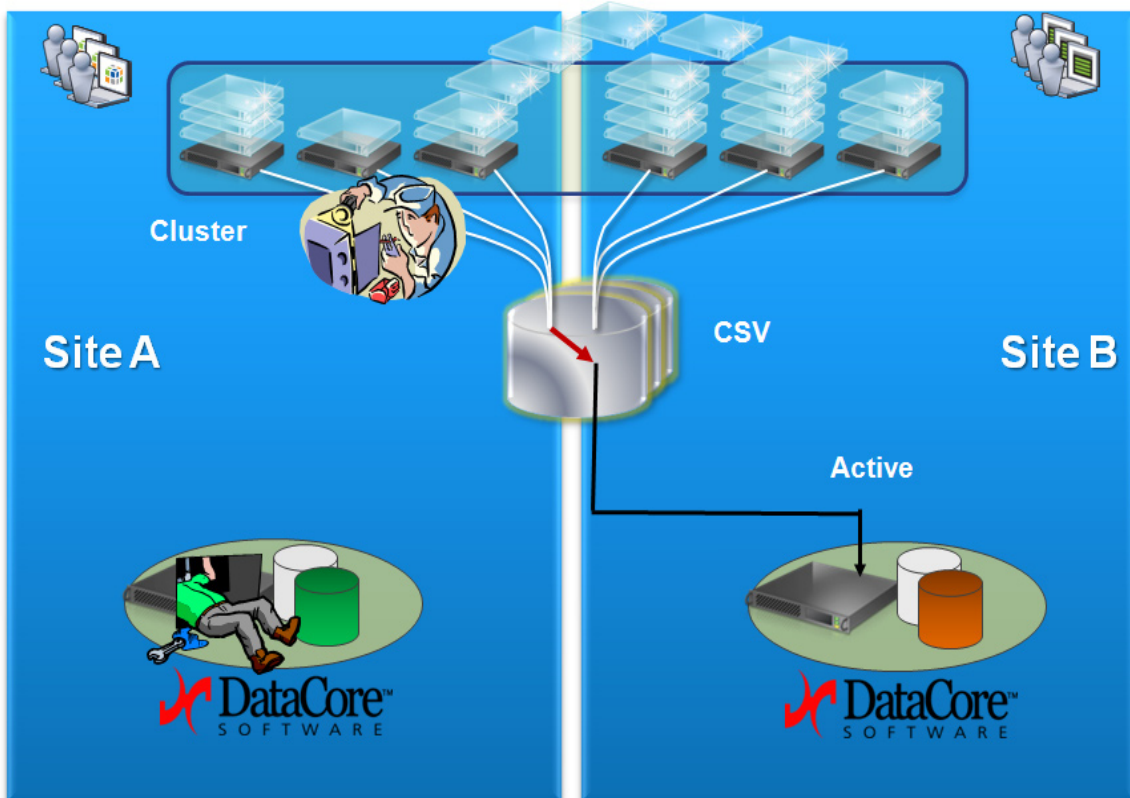
3. There are no host scripts, or custom coding changes to take advantage of the shared virtual disk split across sites because the mirrored virtual disk appears as a well behaved, single multi-ported, shared device.
4. All synchronous mirroring is performed outside the disk subsystems allowing different models and brands to be used at either end.
5. The synchronous mirroring software and associated virtualization stack does not compete for host cycles when run on its own physical nodes and takes advantage of large caches to speed up the response.
6. The storage virtualization/mirroring software may be run in any one of 3 places:
  - a. On a dedicated node at each site acting as a front-end to that site's storage pool
  - b. On a virtual machine (VM) alongside other VM workloads in a physical server at each site.
  - c. Layered on top of Hyper-V accessible by multiple VMs on the same physical server at each site.

### Non-stop Access to CSV from DataCore Software

Combining Windows Failover Cluster capabilities with DataCore multi-site shared virtual disk redundancy prevents a site-wide outage or disaster from disrupting mission critical applications. It uniquely enables hot-hot sites within metropolitan distances.

Rather than declaring a “disaster” when one site’s cluster servers and storage subsystems are taken offline, the cluster fails over applications to the alternate site servers. Those servers’ will simply continue to execute using the same virtual shared disks. Only now, the I/O requests for the CSV and the virtual shared disks are fielded by the surviving storage subsystems at the alternate site.

**Figure 5. Cluster switchover to Hot site using mirrored CSV**



## Restoring Original Conditions

How the original conditions are restored following a site-wide outage depends on whether the outage was planned or unplanned. It also depends on whether equipment was damaged or not. To better understand the restoration behavior, it's useful to see how DataCore software handles an I/O to the CSV stored on a virtual disk. For any mirrored virtual disk, one DataCore node owns the primary copy and another holds the secondary copy. They maintain these copies in lock step by synchronously mirroring updates in real-time from the primary copy to the secondary copy.

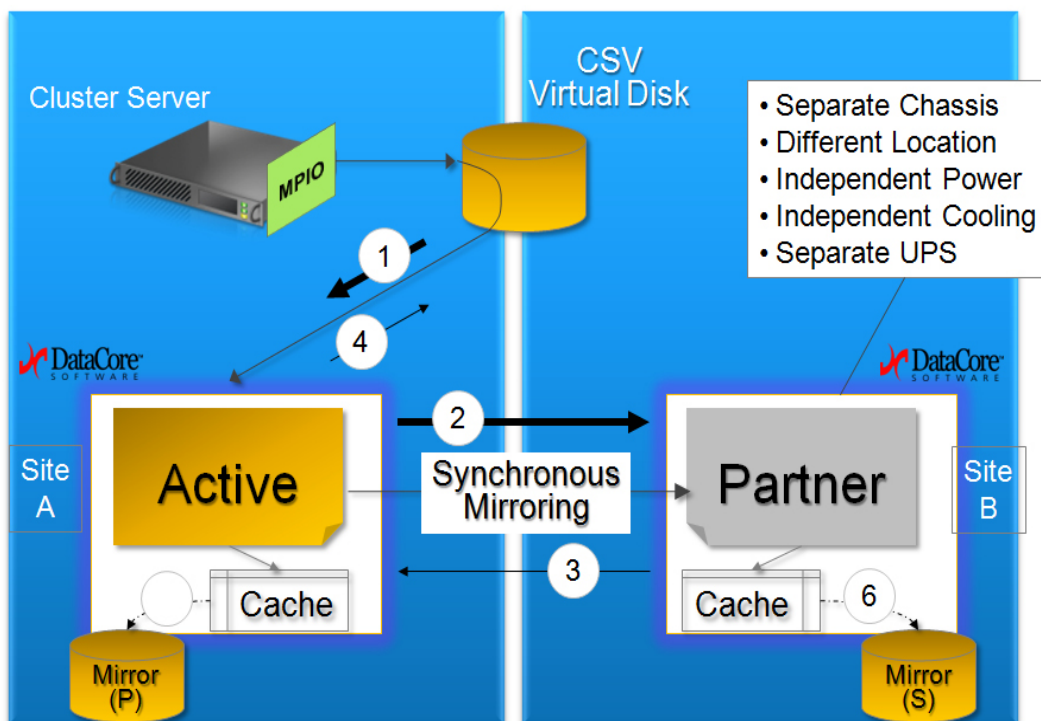
In this Figure 6, Site A owns primary mirror labeled "P" and Site "B" holds the secondary copy labeled "S". The preferred path from the cluster node in Site A to the CSV is assigned to the DataCore node that holds the primary copy of the mirrored set. Under normal operation, all read and write requests issued from Site A cluster servers to the CSV will be serviced by the primary copy local to that site. The secondary copy need only keep up with new updates. Generally, DataCore nodes are configured to control primary copies for some virtual disks and secondary for others, thereby evenly balancing their read workloads.

In step 1, the cluster server writes to the virtual disk on the preferred path, the DataCore node at Site A caches it and immediately retransmits it to the partner node responsible for the secondary copy via step 2. The partner node treats it like any other write request, putting the data into its cache and acknowledging it in step 3.

When Node A receives the acknowledgement from Node B, it signals I/O complete to the cluster server in step 4 having met the conditions for multi-cast stable storage. Sometime later, based on destaging algorithms, each of the nodes stores the cached block onto their respective physical disks as illustrated by steps 5 and 6. But these slower disk writes don't figure directly into the response time from cache, so the applications sense much faster write behavior. An important note: DataCore nodes do not share state to maintain the mirrored copies synchronized as one does in the clustered server design.

From a physical stand point, best practices call for the DataCore nodes to be maintained in a separate chassis at different locations with their respective portion of the disk pool so that each node can benefit from separate power, cooling and uninterruptible power supplies (UPS). The physical separation reduces the possibility that a single mishap will affect both members of the mirrored set. The practical limit for synchronous mirroring today is around 35 kilometers using Fibre Channel connections between nodes over dark fiber.

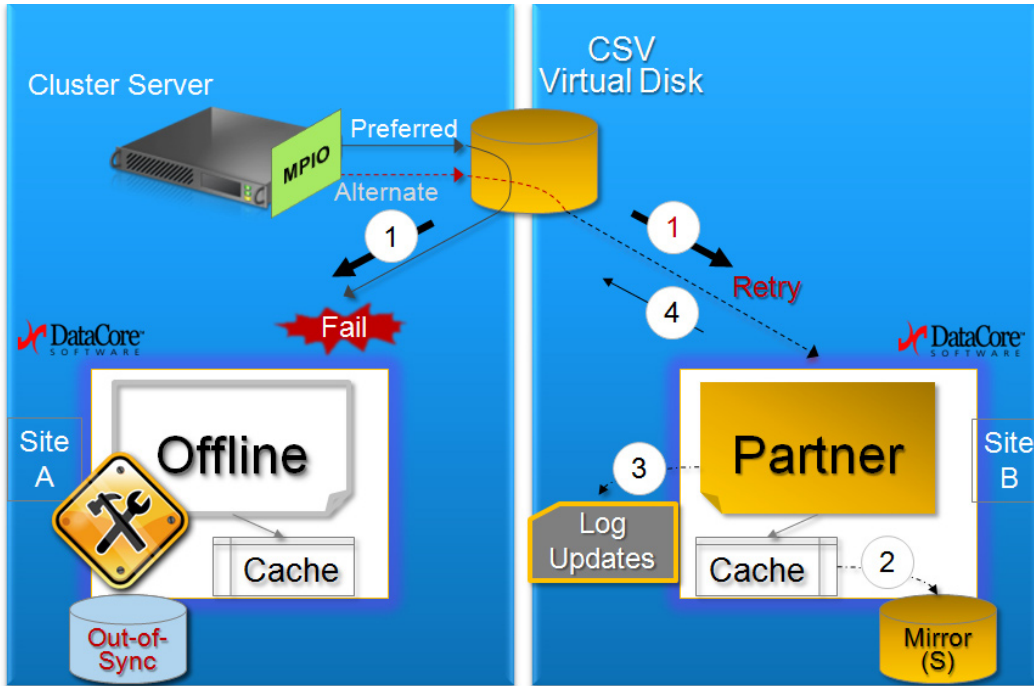
**Figure 6. Synchronous mirroring between DataCore storage virtualization nodes**



## Restoration Following a Planned Site-Wide Outage

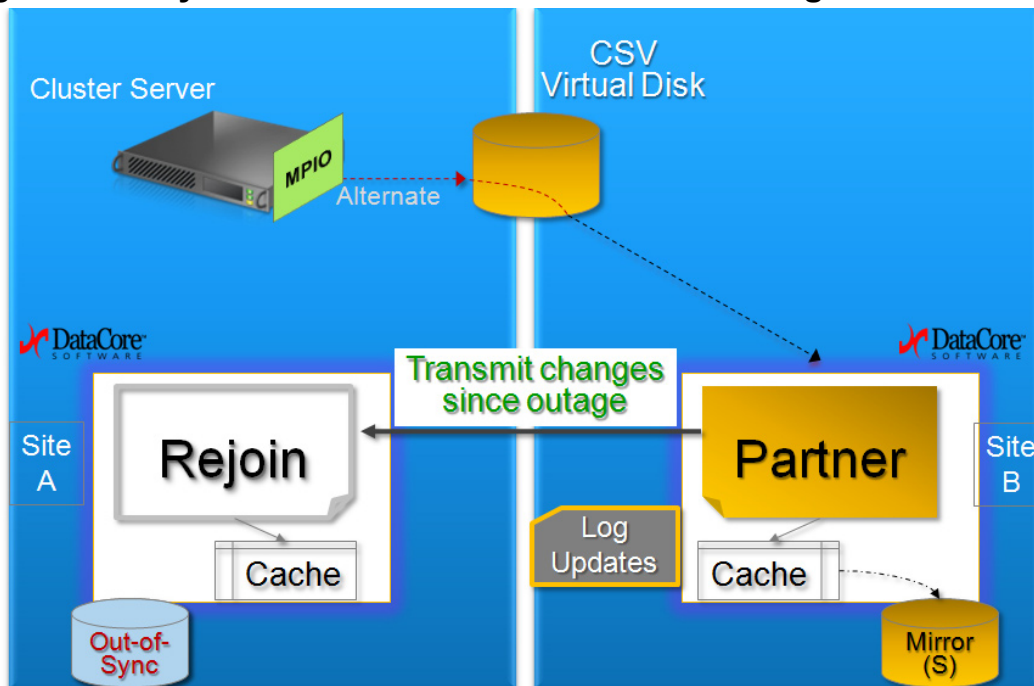
For a planned site-wide outage, say for air conditioning repairs, or in anticipation of a forecasted storm, the site's mirror disks may be gracefully taken out of service while the other site takes over. In this scenario, the DataCore software at the alternate site keeps track of what disk blocks were updated during the planned outage.

**Figure 7. Keeping track of changes when one site is taken-out-service**



When the equipment is put back into service, the DataCore software will copy the latest information from those changed block to the corresponding mirror image until both sites are back in sync. At that point, the cluster can be restored to run across both sites. The time to restore the site is determined by how many distinct blocks changed during the outage. Nevertheless, applying only incremental changes shortens the restoration period significantly.

**Figure 8. Resynchronization of mirrored CSV during site restoration**

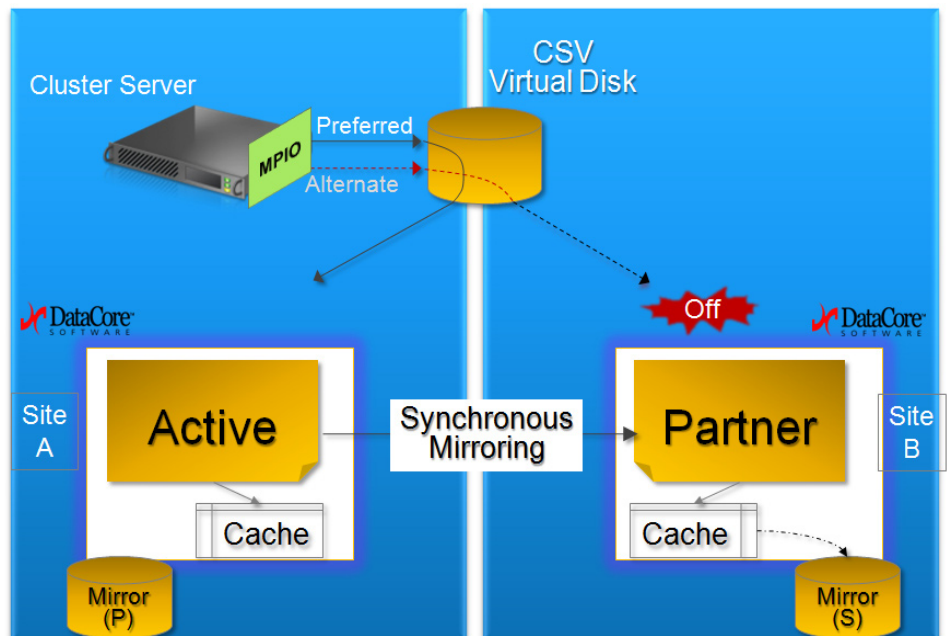


## Restoration Following an Unplanned Site-Wide Outage

Should a more catastrophic event occur where the equipment at one site cannot be gracefully taken out of service, a similar, but a lengthier restoration process occurs. There are two levels of severity. First let's consider the case when the servers and storage virtualization nodes crashed, but the storage devices were not damaged. When the storage virtualization nodes are rebooted, the DataCore software takes inventory of its last known state, particularly of cached disk blocks that were not captured on the disks. The alternate site must then update all those "dirty" blocks along with all blocks known to have changed during the outage before the two sites can fully resynchronize.

In a more severe case, some or all of the physical disks at one site may be damaged and must be completely reinitialized using the latest image from the alternate hot site. Clearly, this will add to the time it takes to restore normal operations. However, in both scenarios, the urgency to do so has been reduced because the alternate hot site is continuing to service applications without any major disruption. And in neither case are any special host scripts required to bring back the original conditions.

**Figure 9. Resume Normal Operations**



### Bottom line- No storage-related downtime

IT organizations derive numerous operational benefits from combining Windows Failover Cluster capabilities with DataCore's device-independent approach to storage virtualization, particularly when the cluster is split between metropolitan sites. The most far reaching is the elimination of single points of failure and disruption that today is responsible for storage-related downtime.

For more information on storage virtualization, please visit:  
[www.datacore.com](http://www.datacore.com)