

SANRAD White Paper: **Designing iSCSI IP-SANs for Mission Critical Applications**

Using IP-SAN Multi-paths

Active / Active IP-SAN Clustering

Network Level Data Mirroring

Remote Replication and DR

Designing iSCSI IP-SANs for Mission Critical Applications

Over the last several years, there has been a gradual shift away from server-based storage to centralized network-based storage. Until recently, the two choices for network-based storage were either network attached storage (NAS) for sharing storage resources using CIFS and NFS or fibre-channel storage area networks (FC-SAN) for sharing block-based storage resources using fibre-channel protocol. Fibre-channel is fast and highly reliable, perfect for revenue critical applications, but proved too expensive and complex for all application servers within the data center. NAS, on the other hand, was much simpler and not as expensive as FC-SANs, but proved to be difficult to scale and not reliable enough for all applications. The limitations of NAS and FC-SAN left the majority of servers still using internal storage. There was a need for another form of network storage that combined the simplicity and flexibility of IP with the scalability, reliability and block-based properties of an FC-SAN. The IETF committee and industry giants worked together to create the new iSCSI standard storage protocol. iSCSI is reliable and block-based like FC-SANs, but it's based on IP so it's simple and cost effective to deploy like NAS.

This white-paper provides information for businesses that are considering an iSCSI IP-SAN to provide centralized storage architecture for large numbers of mission critical application servers. Topics reviewed include:

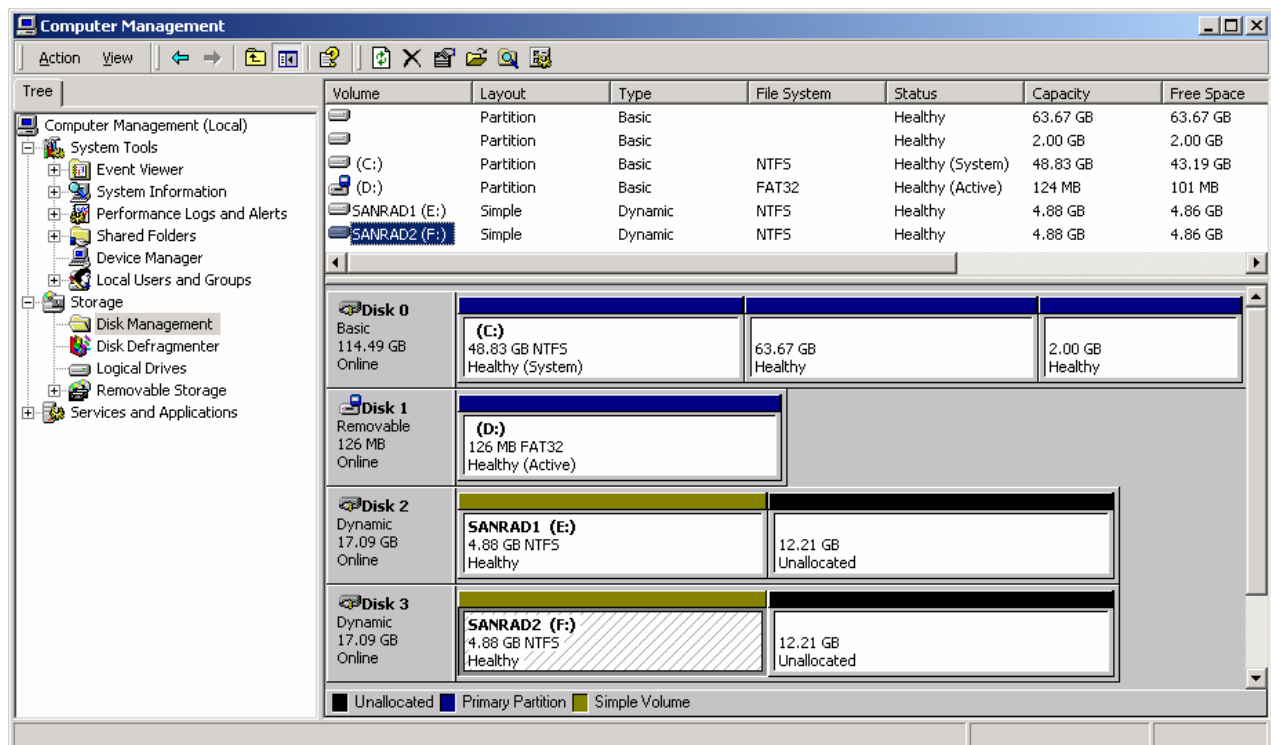
- **100% redundant connection paths between servers and the iSCSI IP-SAN**
- **iSCSI IP-SAN component clustering with automatic failover using IP-Takeover**
- **Real-time network based data mirroring and multi-pathing**
- **Creating a remote hot, warm or cold DR site**

iSCSI Background

Before examining how iSCSI can enable high availability for an iSCSI IP-SAN, let's review how an iSCSI storage target (volume) interfaces with the operating system and file system on the host. This is especially important to understand since the host can conceivably be in Florida with data and the storage targets in California. Where NAS works at the file level and is limited to specific applications, an iSCSI IP-SAN works under the file system, at the block level. This allows iSCSI to support virtually any application, including databases, backup and restore applications and server clusters. In addition, because iSCSI works at the block level, iSCSI packets are five to ten times more efficient than NFS or CIFS.

The file system makes read and write requests to the storage devices using a set of standard SCSI commands. The file system has 100% control over these storage devices. iSCSI, like standard SCSI, is a block-based storage protocol layered underneath the file system. This means that an iSCSI volume appears as an additional disk drive when mounted by the file system. The iSCSI volume can be partitioned, named and formatted like a normal disk drive.

The following MS Windows screen shows two new disk drives, Disk 2 and Disk 3, that are actually allocated to the server from an IP-SAN using the storage services V-Switch. These two disk drives have been partitioned and formatted and are ready to store data or run applications. To any operating system, these iSCSI delivered volumes are “disk drives” and can run any application or store any data you are today storing on the internal disk drives.



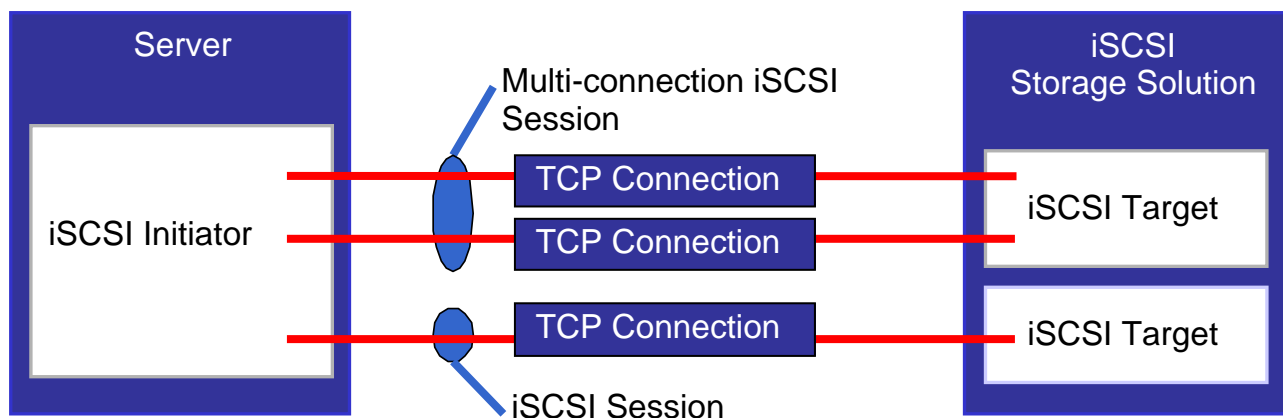
Disk drives 2 & 3 allocated to server by V-Switch

An iSCSI initiator (software driver) is native on Windows, Netware, Linux, MAC OSX, HP-UX, AIX and many other popular operating systems. The initiator is used to mount and use the iSCSI storage. Standards for the iSCSI initiator are governed by the IETF (Internet Engineering Task Force). The iSCSI initiator responds to file system SCSI commands that are targeted to the disk drives the iSCSI initiator represents. The iSCSI initiator encapsulates these SCSI commands and data into iSCSI packets that are, in turn, encapsulated into TCP/IP packets. The TCP/IP packets are then routed very quickly over the Ethernet network where they are delivered to an iSCSI storage target. The iSCSI storage target can be located on the same network within the building or halfway around the world. This iSCSI target has all the same attributes as a standard SCSI storage system. Once the iSCSI packet arrives at the iSCSI storage system representing the targets, the SCSI commands and data are un-encapsulated from the iSCSI/TCP/IP packet and are executed on the storage system. Once executed, the results are encapsulated back into iSCSI/TCP/IP and returned to the iSCSI initiator on the server where they are de-capsulated and delivered to the SCSI layer and then the file system.

100% Redundant Connection Paths between Servers and iSCSI IP-SAN

Automatic Multi-path Failover with the iSCSI Initiator. The iSCSI initiator comprises layers that are key to providing multiple data paths between servers and iSCSI storage targets. The two main layers within the iSCSI initiator are the **session** and **connection** layer.

The session layer is an upper layer and is responsible for maintaining the communication to the SCSI layer within the server. It also ensures proper order of SCSI commands and data to and from the server file system and the iSCSI storage target. SCSI commands are numbered in sequence as they are sent from the server. The iSCSI storage target arranges the SCSI commands according to their order, ensuring that commands are not lost, taken out of order or duplicated. Within every server there is usually only one iSCSI initiator but there can be more than one session established and running within a single initiator. For example, if there are two iSCSI storage targets being used by the server, then there would be one initiator with two sessions running. In below there are two sessions running which means that the server iSCSI initiator is using two unique iSCSI connected targets or disk drives /volumes. One is using a single path and the second is using multiple connections for high availability.



The connection layer manages the TCP/IP connection between the server and the iSCSI storage target which, in our case, is represented by the storage services V-Switch. The session layer can maintain several connections. In common applications there is only a primary active iSCSI TCP/IP connection. But for applications requiring 99.999% high-availability there is a primary iSCSI TCP/IP connection and an alternant connection. All iSCSI traffic between the server and storage system travels over the primary connection. In most IP-SAN deployments, this is a Gb Ethernet.

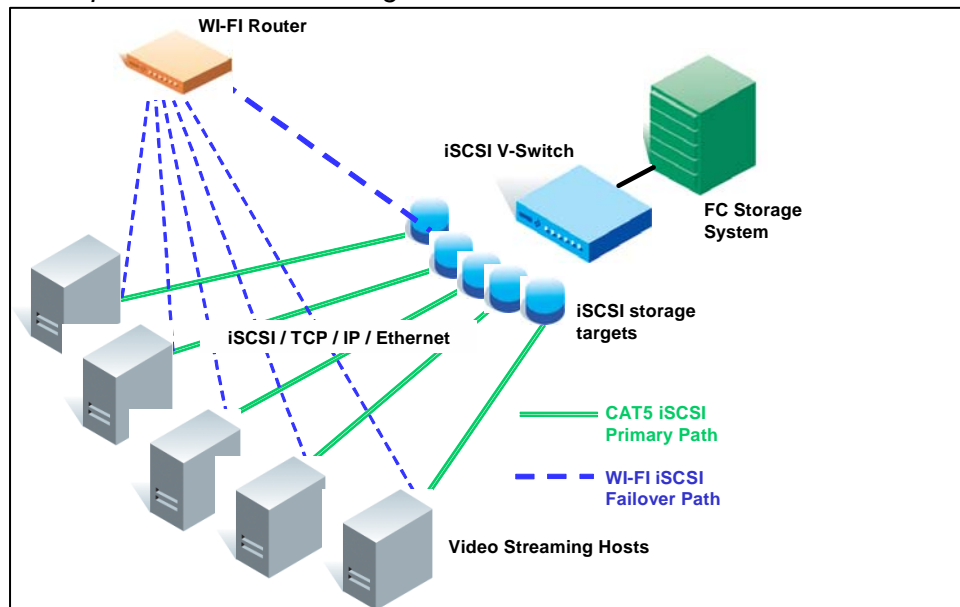
Multi-pathing and path-failover are automatic with iSCSI. Because the iSCSI session is aware of alternate TCP/IP paths to the iSCSI storage target, it will automatically transfer traffic through an alternate TCP/IP connection. So, if one TCP/IP connection between a server and iSCSI storage target fails, the traffic is automatically routed through the alternant TCP/IP connection.

Because the SCSI commands are numbered, the iSCSI storage target is able to arrange the commands received across multiple connections.

Demonstration of Server Multi-path Failover

To demonstrate how iSCSI failover functions 5 videos were streamed from 5 iSCSI storage targets on the storage services V-Switch to 5 Microsoft Windows 2003 hosts. See figure below. Each host used the native Microsoft-supplied iSCSI initiator software. As reviewed earlier, the iSCSI session layer was responsible for maintaining the video stream to the video player application on the hosts. Each host had two TCP/IP Ethernet connections used by the iSCSI session. The primary connection was a 10/100 Ethernet CAT5 copper connection between the host and storage services V-Switch via a switched LAN. The second connection was a wireless WI-FI connection between the host and V-Switch. The following statement reviews the failover test and results:

“We were able to disconnect the CAT5 cable from the hosts and the iSCSI session automatically routed the traffic over to the WI-FI connection. We were able to do video streaming to 5 hosts running iSCSI (wireless) going to an access point, then to a hub, then to the storage services V-Switch. This forced failover test was extremely positive and fast enough to keep all 5 videos streaming.”

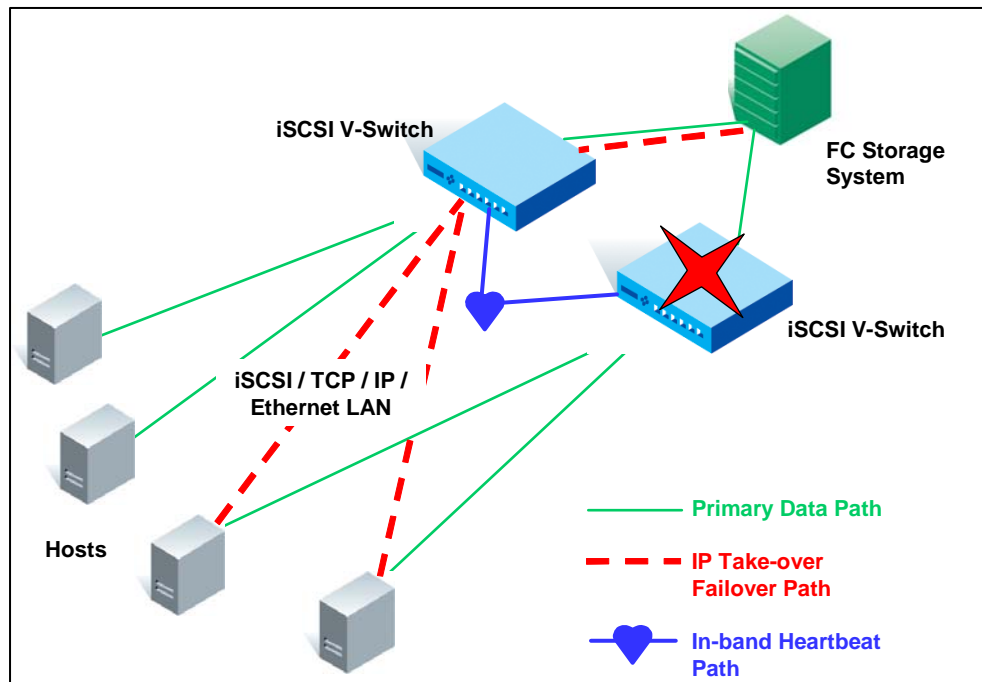


SANRAD Demonstration of Server Multi-path Failover

iSCSI SAN Component Storage Clustering

In addition to server multi-pathing and failover, there are two more main requirements for providing high availability. They are to provide multiple paths through the IP-SAN layer and to replicate not only the data but the access point to the data. To provide multiple paths to the

storage systems in real-time, we use a technique called IP takeover . In the event an iSCSI storage services V-Switch is temporarily off-line, the second V-Switch attached to the same storage elements and the same host network will automatically take over the IP addresses and data communication for the off-line switch. Both V-Switches are “active” servicing their assigned hosts but they can also provide a failover path for other hosts within the network. This is because both V-Switches maintain the configuration information of other V-Switch within the cluster and monitor the heartbeat of their designated partner. When a site or V-Switch goes off-line, the iSCSI initiator at the host will terminate the iSCSI connections with the offline storage services V-Switch. But it will not terminate the iSCSI session within the host. It will maintain the session while waiting for the IP addresses for the iSCSI storage targets to be re-exposed. IP-SAN V-Switches acting in a cluster send heartbeat packets back and forth to monitor the health of the OP-SAN. When a V-Switch goes off-line there is a heartbeat failure. When heartbeat failure occurs, the surviving V-Switch will automatically expose the IP addresses from the down V-Switch. The iSCSI initiator connection layer will automatically discover the re-exposed IP addresses and create a new connection thus enabling the hosts to proceed with communication through the surviving V-Switch. The V-Switch will continue to service it’s own hosts and the hosts of the off-line V-Switch until the original off-line V-Switch is brought back on-line and the connection paths repaired. The IP addresses will be automatically restored to the returning V-Switch and the original connections will be re-established, thus providing continuous data availability. The entire failover and fail-back process happens online in seconds. All applications servers are unaware of any connection or system outage and continue to function without any downtime. See diagram below.

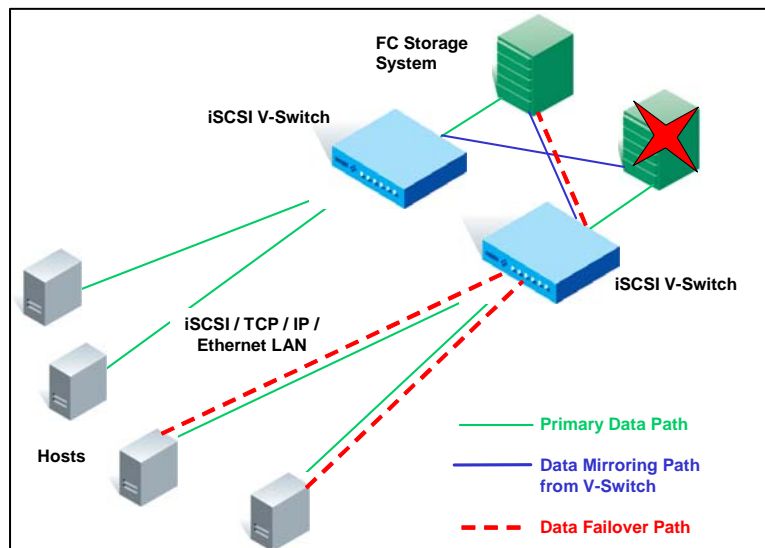


Example of V-Switch IP take-over and High Availability

Real-time Network Based Data Mirroring and Multi-pathing

The storage services V-Switch operates in the network layer, it can create and maintain mirrored partners/ volumes anywhere within the network, indifferent to traditional physical limitations such as enclosures and distance. Local synchronous mirroring can now be performed between two or more storage enclosures. For example, a V-Switch in Building A with an FC-attached storage system can keep the data files on the storage system synchronized with an FC-attached storage system in building B, and another FC-attached storage system in building C. The V-Switch can maintain all three as partners within a mirror.

Like a RAID controller, if one of the mirrored partners goes off-line or experiences a failure, the V-Switch will automatically remove the failed partner from operation but will continue to service the application I/O requests with the remaining mirrored partner. See diagram below.

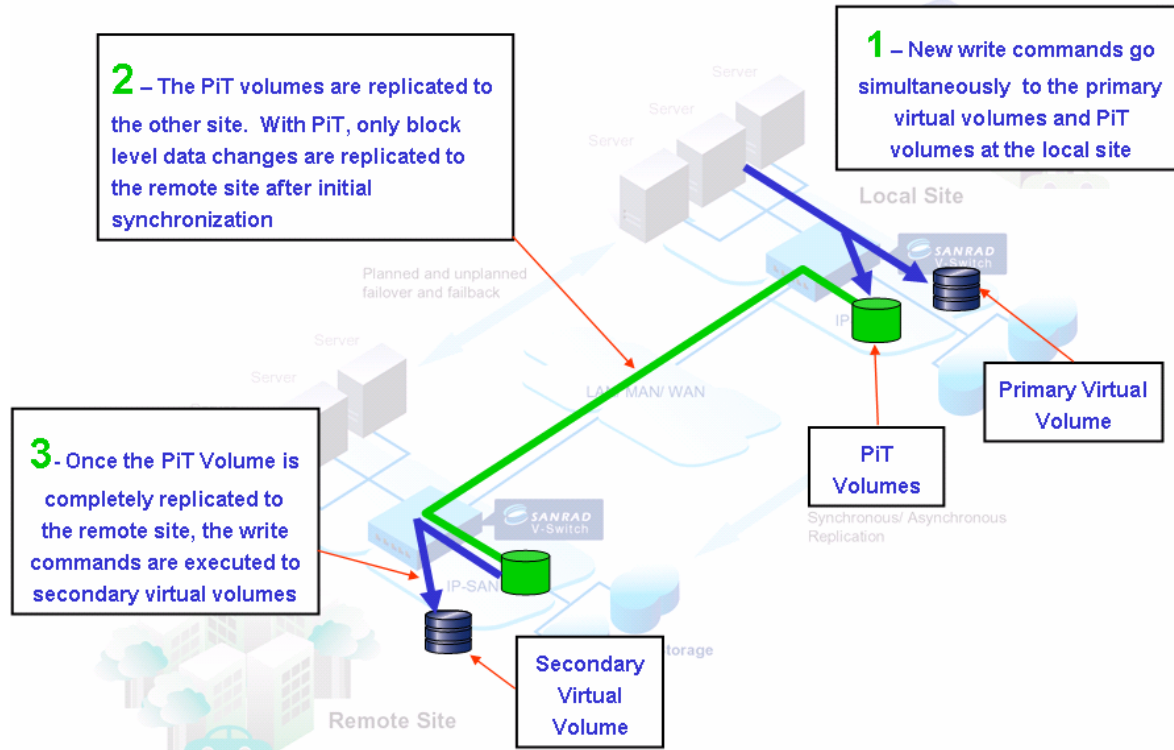


Example of V-Switch Data Mirroring to two FC systems with auto-failover

Creating a Remote Hot, Warm or Cold DR Site

Storage services V-Switches can also replicate data across long distances using iSCSI over IP networks (LAN, MAN, WAN). As mentioned earlier, V-Switches can synchronously mirror data between volumes on storage systems. This mirroring capability also exists between V-Switches. Data can be replicated or mirrored in one to one (see diagram below), many to one, or many to many configurations. This mirroring can be done in a synchronous (real-time) manner if you have a fast network or done in an asynchronous manner when IP links between sites are slow. To keep replication as efficient as possible the IT professional can specify which volumes need to be replicated. During replication, the write commands are cached locally into consistency groups on to non-volatile local media (disk drives). These groups of write commands are compiled into PIT files which are replicated to other storage services V-Switches.

Diagram of IP-SAN using remote data replication in a one to one configuration across IP LAN, MAN or WAN. Replication can be one way or both ways (back to back)



Using the V-Switch to replication is different than host-based solutions. Host-based solutions are platform-dependent, degrade server performance and require an agent on each host. iSCSI SAN level replication is platform agnostic and requires no host agents and will not degrade server performance.. The storage services V-Switch is also storage agnostic, eliminating vendor lock-in and enabling the continued use of all storage investment. For example, you can use of low-cost SATA disks at the secondary site even when high-cost enterprise-class RAID subsystems are used at the primary site.

Since both sites in the diagram are using V-Switches, replication can be done back and forth where each site is acting as the recovery site for the other. The V-Switches also acts as a mount point for the remote site. So the remote site can be used for recovering the primary site or for site failover in the event the primary site is to remain offline for an extended period of time.

Conclusion

iSCSI IP-SANs can be designed to have the same level of reliability as highly available FC-SANs. There are 4 features that provide this level of high availability within an iSCSI IP-SAN

- 100% redundant connection paths between servers and the iSCSI IP-SAN
- iSCSI IP-SAN component clustering with automatic failover using IP-Takeover
- Real-time network based data mirroring and multi-pathing
- Creating a remote hot, warm or cold DR site

By using these features, businesses can take advantage of iSCSI cost savings and simplicity while delivering the same level of reliability as enterprise class FC-SANs.