

# **SANRAD White Paper: Continuous Data Access – SANRAD Enterprise Level High Availability WP 009-01**

SANRAD  
US Tel: +1-800-461-2616  
International Tel: +972-3-767-4800

**Copyright SANRAD 2004**

All rights reserved. The copyright and all intellectual property rights in this article belong to SANRAD. It is strictly forbidden to copy, duplicate or otherwise use this article or any part thereof in any way shape or form without the prior written consent of SANRAD.

---

## Table of Contents

INTRODUCTION.....	3
THE 99.999 CHALLENGE FACING NETWORK ADMINISTRATORS.....	3
ISCSI BACKGROUND.....	4
ISCSI PROVIDES MULTIPLE DATA PATHS FOR A SERVER .....	6
Automatic Multi-path Failover with the iSCSI Initiator .....	6
SANRAD Demonstration of Server Multi-path Failover .....	7
THE V-SWITCH IP TAKE-OVER FOR ENHANCED HIGH-AVAILABILITY .....	8
LOCAL AND REMOTE SYNCHRONOUS DATA MIRRORING AND FAILOVER .....	9
Local Synchronous Data Mirroring within a LAN.....	9
SUMMARY.....	11

## Table of Figures

FIGURE 1.	“LOCAL’ DISK DRIVES 2 & 3 ACTUALLY ALLOCATED TO SERVER BY V-SWITCH .....	4
FIGURE 2.	AVERAGE SERVER COMMAND STACK .....	5
FIGURE 3.	ISCSI INITIATOR WITH STANDARD AND MULTI-PATH SESSION.....	6
FIGURE 4.	SANRAD DEMONSTRATION OF SERVER MULTI-PATH FAILOVER.....	7
FIGURE 5:	EXAMPLE OF V-SWITCH IP TAKE-OVER AND HIGH AVAILABILITY .....	8
FIGURE 6.	EXAMPLE OF V-SWITCH DATA MIRRORING TO TWO FC SYSTEMS WITH AUTO-FAILOVER.....	10

## Introduction

Building, managing and optimising company networks and their associated server applications, file systems and storage resources are the challenges of every network and system administrator – going back to the 1960's. The search continues for products that increase management capabilities and optimize resources yet are cost-effective and easy to deploy.

SANRAD's V-Switch gives administrators a storage management and optimisation tool that provides scalability, flexibility, reliability and high performance in a single, easy-to-use, cost-effective platform specifically designed for enterprise class applications. SANRAD's V-Switch meets the requirements of mission critical environments with its performance and robust high availability features.

## The 99.999 Challenge Facing Network Administrators

*Dealing with the new challenge of "Never Down" data and storage solutions.*

Driven by the Internet, government regulations, business continuance policies and a global economy, 24 hr data availability is becoming the norm. Failover, component redundancy, multiple-path connections, and data mirroring are becoming common requirements in the business-computing environment. Zero down-time is the goal but up until now, not always technically or fiscally achievable for most businesses.

These challenges are even more acute when you consider that data is expanding annually at geometric rates while resources and budgets lag far behind. Achieving more with ever decreasing resources is not an impossible task, but it does require a shift in direction. Network and system administrators are eager to overcome their challenges and are more than willing to try a new solution if it will truly solve their problems.

Storage Area Networks (SANs) can deliver data storage solutions that provide 99.999% data storage availability but, originally, SANs were built on FC technology. However, for the majority of businesses, using FC to build a SAN is too expensive to deploy; too rigid in its components' geographical constraints and too complex to configure according to real-time network needs. Industry leaders like Microsoft, Cisco, Hewlett-Packard and IBM realized that a new network architecture was needed and developed a way to route SCSI over TCP/IP or "iSCSI". iSCSI paved the way for a new type of SAN, an "IP-SAN", based on TCP/IP over Ethernet. iSCSI makes everyday SCSI a routable network fabric protocol. iSCSI enables any SCSI command or data to become routable over any sized network, including the internet, wireless LANs, line-of-sight infrared connections or even satellite relays. With the right products, any business can now build a SAN that is simple, cost effective, highly scaleable and fast enough to exceed their performance requirements.

This white-paper will review why IP-SANs are equal to or more robust than tradition FC-SANs and how SANRAD's V-Switch can be deployed in an IP-SAN to provide an architecture for enterprise environments requiring 99.999% data availability.

## iSCSI Background

Before examining how iSCSI can enable high availability for an IP-SAN, let's review how an iSCSI storage target (volume) interfaces with the operating systems and file system on the host. This is especially important to understand since with iSCSI the host can conceivably be in Florida while the data and the storage targets are in California. Where NAS and its associated CIFS and NFS work at the file level and are limited in their application support, an IP-SAN and iSCSI work under the file system at the block level. This allows iSCSI to support virtually any application, including databases, backup and restore applications and Exchange. In addition, because iSCSI works at the block level, iSCSI packets are five to ten times more efficient than NFS or CIFS. iSCSI also performs well within a LAN, a topography where NFS and CIFS are far too slow.

With traditional DAS or FC-SAN attached disk drives or tape, a file system makes read and write requests via the server to the storage devices using a set of standard SCSI commands. The OS and file system have 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 OS. The iSCSI volume can be partitioned, named and formatted like a normal disk drive – independent of the OS and file system.

The following MS Windows screen shows two new disk drives, Disk 2 and Disk 3, that are actually allocated to the server from a SANRAD IP-SAN using our V-Switch. These two disk drives (17 GB each) have been partitioned and partially formatted and are ready to store data or even run applications. To any operating system like Windows, 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.

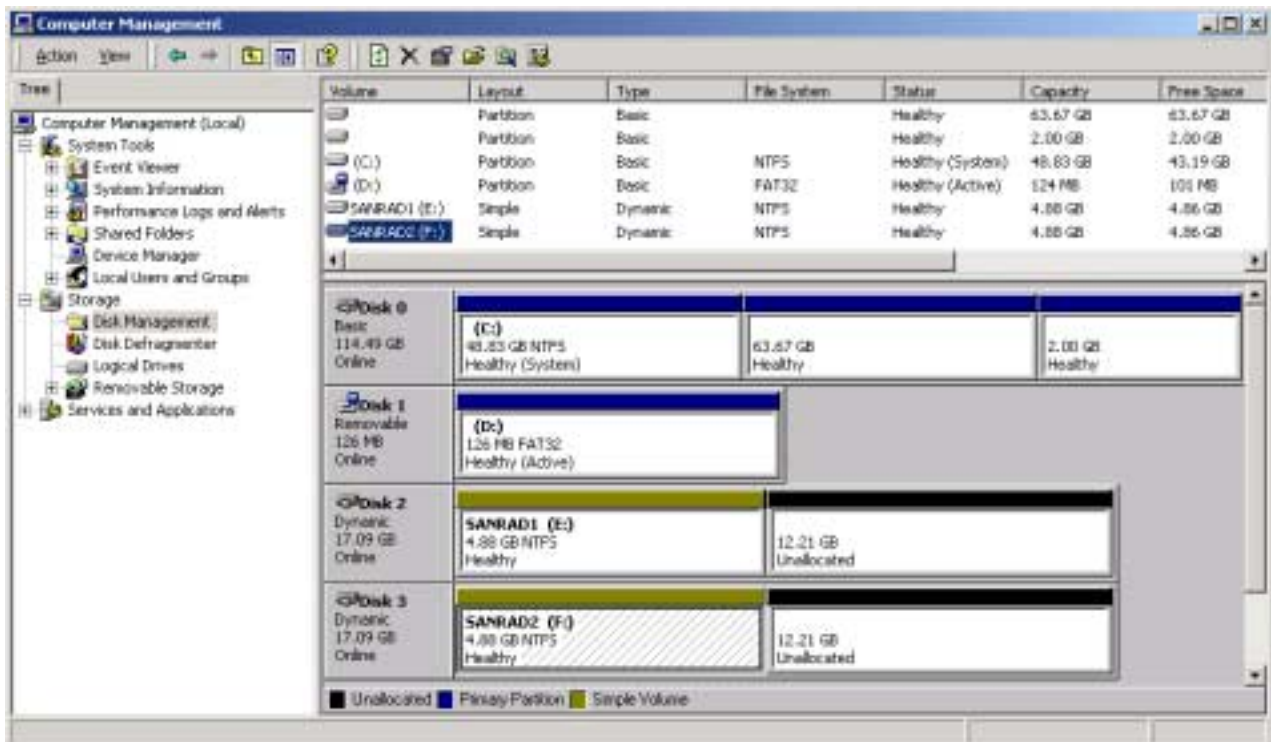


Figure 1. “Local’ Disk Drives 2 & 3 Actually Allocated to Server by V-Switch

An iSCSI initiator driver is native on Windows, Novell, Linux, MAC OSX, HPUX, AIX and many other popular operating systems. Standards for the iSCSI initiator are governed by the IETF (Internet Engineering Task Force). The iSCSI initiator driver responds to file system SCSI commands that are targeted to the disk drives the initiator represents. The iSCSI initiator driver encapsulates these SCSI commands and data into iSCSI packets that are, in turn, encapsulated into TCP/IP packets. See Figure 2, page 6. 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 decapsulated 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 driver on the server where they are decapsulated and delivered to the SCSI layer and then the file system. The SANRAD V-Switch supports storage systems and connections of the following types: SCSI, FC, FC Fabric, or any IDE, SATA, or ATA system with a SCSI or FC controller.

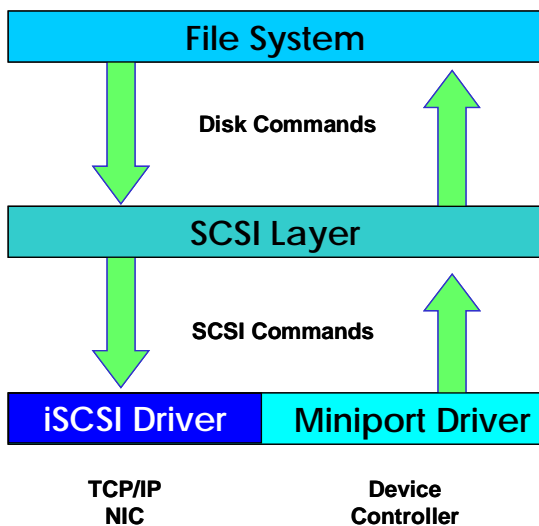


Figure 2. Average Server Command Stack

**Note:** **Performance:** The overhead of the actual iSCSI encapsulation process and networking is generally 3 – 5% that of the disk drive I/O so using iSCSI vs. internal disk drives will generally yield the same performance. In some application servers, a dedicated 10/100 Ethernet link (100Mb) is adequate since most servers don't generate more than 50Mbs of sustained iSCSI / storage traffic. A Gb Ethernet link (1000Mb) is recommended to handle I/O spikes if spikes result in I/O bursts greater than 50Mbs assuming a 50% overhead.

Performance gains can be achieved best by using a dedicated Gbit LAN and storage systems that have better performance than your average internal server drives.

# iSCSI Provides Multiple Data Paths for a Server

## 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, like those provided by the V-Switch. The two main layers within the iSCSI initiator are the session and connection layers.

The first layer is the “session” 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 to 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 Figure 3 there are two sessions running which means that the server iSCSI initiator is using two unique iSCSI connected targets or disk drives /volumes.

The second layer is the “connection” layer. The connection layer is the TCP/IP connection between the server and the iSCSI storage target which, in our case, is the V-Switch. The session layer can maintain several connections. In common applications there is only a primary iSCSI TCP/IP connection. But for 99.999 applications there is a primary iSCSI TCP/IP connection and an alternant connection. See Figure 3. All iSCSI traffic between the server and storage system travels over the primary connection. In most IP-SAN deployments, this is a Gb Ethernet link and is more than fast enough to handle average server traffic. With some application servers, a 10/100 Ethernet link (100Mb) is adequate since most servers don’t generate more than 50Mbs of sustained iSCSI / storage traffic. A Gb Ethernet link (1000Mb) is recommended to handle I/O spikes if spikes result in I/O bursts greater than 50Mbs assuming a 50% overhead.

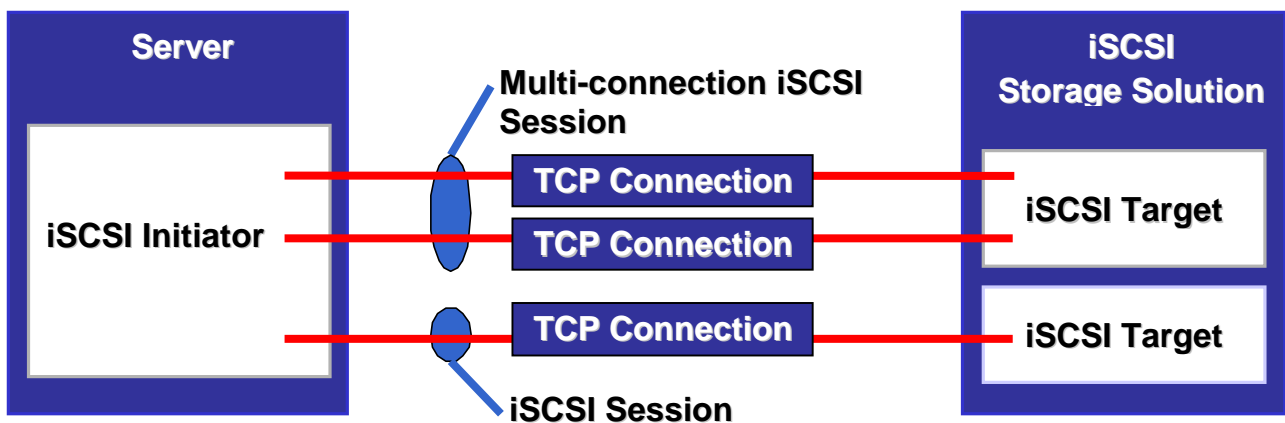


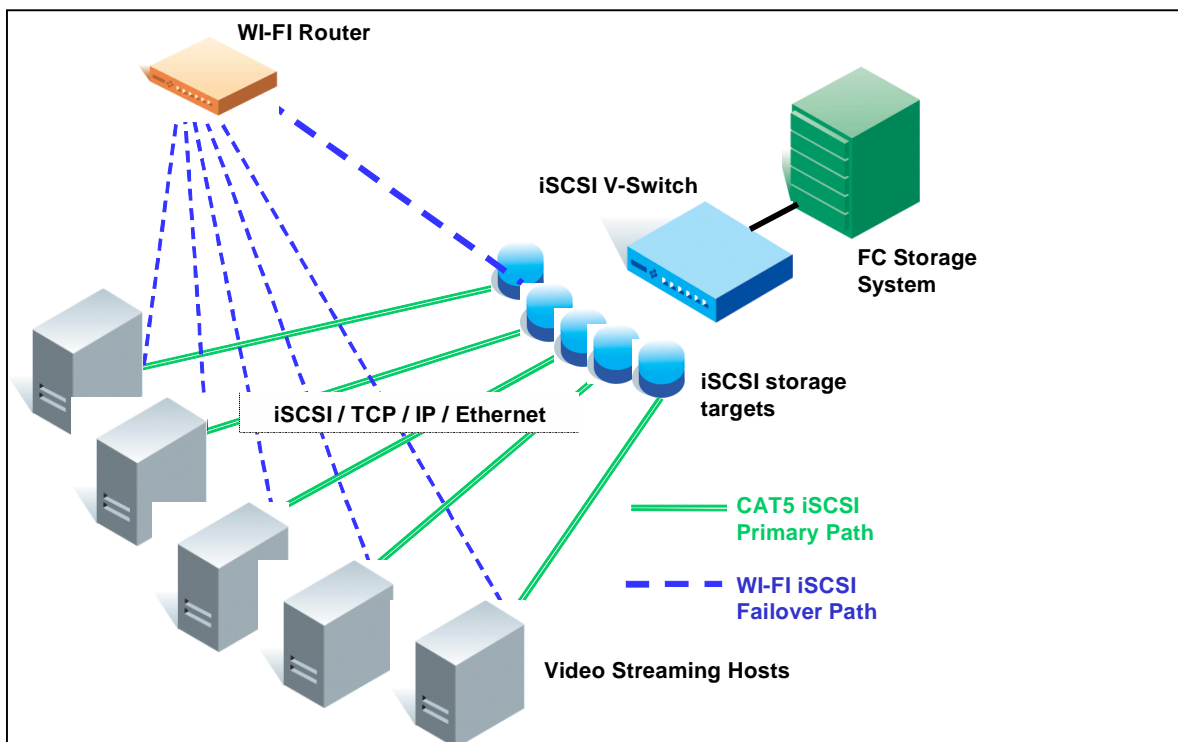
Figure 3. iSCSI Initiator with standard and multi-path session

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 alternate TCP/IP connection. Because the SCSI commands are numbered, the iSCSI storage target is able to arrange the commands received across multiple connections.

## SANRAD Demonstration of Server Multi-path Failover

To demonstrate how iSCSI failover functions, SANRAD participated in a third party demonstration where 5 videos were streamed from 5 iSCSI storage targets on the V-Switch to 5 Microsoft Windows 2003 hosts. See Figure 4, page 10. Each host used the native Microsoft-supplied iSCSI initiator driver 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 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 SANRAD V-Switch (iSCSI to FCP-SCSI), then to a core-edge FC fabric, then to a virtual port, and finally mapped to an FC open-9 LUN on our enterprise class storage system. This forced failover test was extremely positive and fast enough to keep all 5 videos streaming.”*



**Figure 4. SANRAD Demonstration of Server Multi-path Failover**

## The V-Switch IP Take-Over for Enhanced High-Availability

In addition to the server layer, there are two more main requirements for providing high availability. They are to provide multiple paths through the SAN and to replicate not only the data but the access point to data. The first is to provide multiple paths to the storage devices. IP take-over is key to enabling high availability and failover paths with the V-Switch. In the event a V-Switch is temporarily off-line, the second V-Switch attached to the same storage and the same host network can 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 “passive” 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 or partners. When a site or V-Switch goes off-line, iSCSI will terminate the host connections with the problematic site but maintain the iSCSI session within the host while waiting for the IP addresses for storage to be re-exposed. The partner V-Switch will now expose the IP addresses from the down site or V-Switch. The iSCSI initiator will discover the re-exposed IP addresses and create a new connection thus enabling the hosts to proceed with communication through a new V-Switch to the storage systems. 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 or site are repaired. The IP addresses will be automatically restored to the returning switch and the original connections will be re-established, thus providing continuous data availability. See Figure 5.

IP take-over can be used with a single data center to provide multiple failover paths to storage systems. It can be used across multiple data centers to automatically provide a new route to storage resources in the event of a sectional site failure. As mentioned in the previous section, when used in conjunction with V-Switch mirroring and FC or third party FC tunneling techniques, it can provide an off-site disaster recovery facility which re-connects hosts with mirrored partners within seconds of primary site failure.

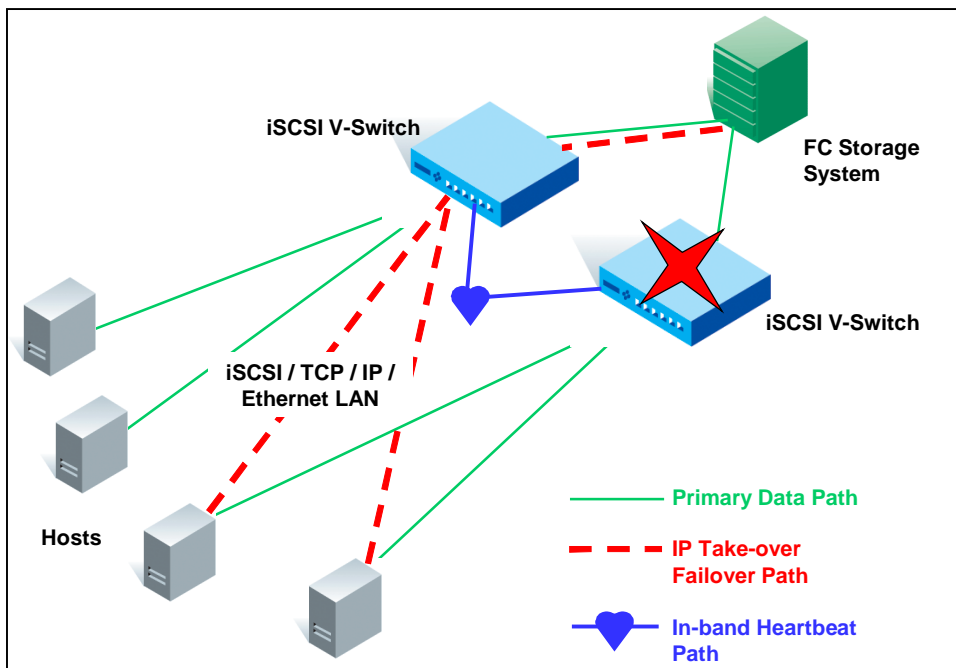


Figure 5: Example of V-Switch IP take-over and High Availability

## Local and Remote Synchronous Data Mirroring and Failover

Mirroring is a RAID technique for creating and maintaining identical data sets on different physical disks within an array. Mirroring protects data and keeps applications operational in the event of disk failure. Without mirroring, a damaged disk drive would halt an application and potentially cause data loss. With RAID, if a disk drive fails within a mirror, it will have an identical partner or partners that will seamlessly take over the I/O requests previously serviced by the down disk drive. Once the failed drive is replaced, the partner can rebuild the new replaced disk drive with the current data until its data is 100% resynchronized to match the partner. At this time the replaced disk drive rejoins its partner or partners and can again support I/O requests. Read commands are generally divided among all mirrored partners. Any new data write commands must be replicated and sent to all the mirrored partners to maintain data continuity across all mirrors. All partners must acknowledge and confirm the write request before the RAID controller returns a write acknowledgment to the file system of the host(s). This insures that all mirrors remain synchronized and the file system of the host(s) remains synchronized with the data on the mirrors. If a mirrored partner does not return a write acknowledgement, then the RAID controller will retry several times to confirm that one of the partners has failed. If there is still no write acknowledgement, the RAID controller will send an alert to the system administrator but it will continue to service file system I/O requests with the surviving partner or partners. Some RAID controllers provide “self healing” where unused disk drives within the RAID system will automatically be selected as the new “spare” drive and be automatically synchronized and enabled as the new replacement disk drive.

### Local Synchronous Data Mirroring within a LAN

The V-Switch can perform mirroring with block-level virtualization capabilities. Mirroring is performed by the V-Switch in a similar fashion as a RAID controller. The major difference is that RAID controllers mirror storage devices within a single enclosure. However, because the V-Switch is 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 FC 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 Figure 6

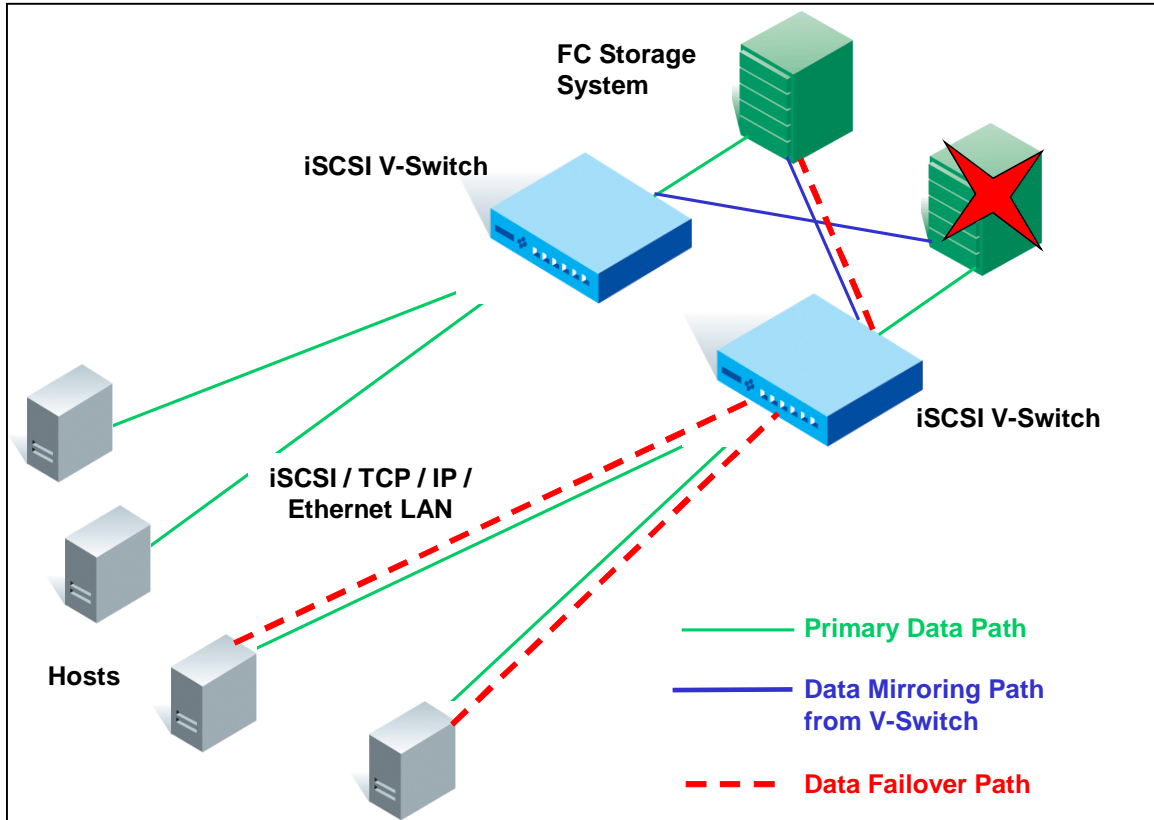


Figure 6. Example of V-Switch Data Mirroring to two FC systems with auto-failover

## Summary

SANRAD IP / iSCSI based Storage Area Networks (SANs) can deliver data storage solutions that provide 99.999% data availability. By combining iSCSI session/connection automatic failover capabilities, SANRAD IP Take-over and SANRAD storage system mirroring, users can use iSCSI V-Switches to replicate their data and provide multiple redundant data paths with automatic failover and fail-back capabilities to ensure that their servers have access to data 24 x 7 x 365.