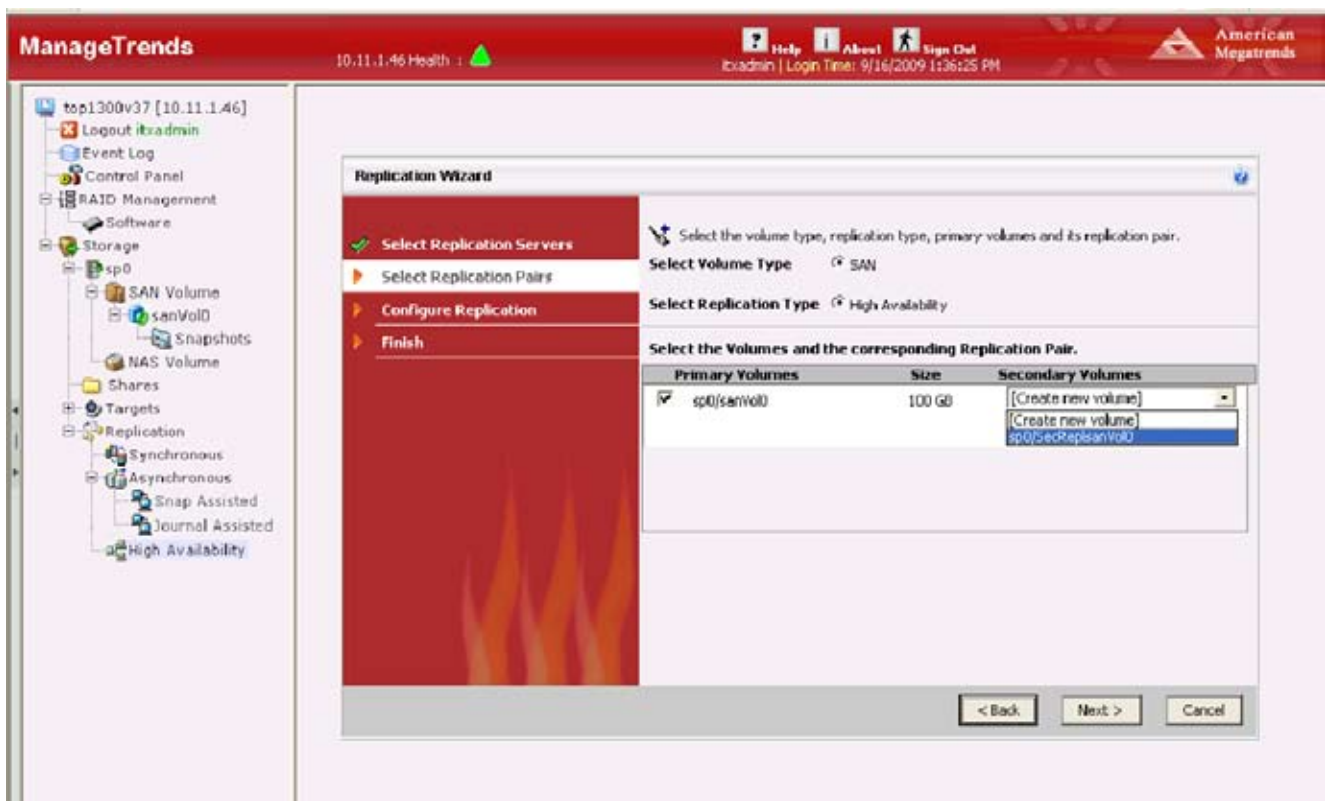


New High Availability Architecture for Improved Data Continuity and System Availability in StorTrends® iTX 2.7 v1030.3.x



© Copyright 1998-2009 American Megatrends, Inc.

All rights reserved.

American Megatrends, Inc.

5555 Oakbrook Parkway, Building 200

Norcross, GA 30093

TRADEMARK AND COPYRIGHT ACKNOWLEDGMENTS

This publication contains proprietary information that is protected by copyright. No part of this publication can be reproduced, transcribed, stored in a retrieval system, translated into any language or computer language, or transmitted in any form whatsoever without the prior written consent of the publisher, American Megatrends, Inc. Trademarks and trade names may be used in this document to refer to either the entities claiming the marks and names or their products. American Megatrends, Inc. disclaims any proprietary interest in trademarks and trade names other than its own.

FOR ADDITIONAL INFORMATION

Call American Megatrends at 1-800-246-8600 for additional information. You can also visit us online at ami.com.

LIMITATIONS OF LIABILITY

In no event shall American Megatrends be held liable for any loss, expenses, or damages of any kind whatsoever, whether direct, indirect, incidental, or consequential, arising from the design or use of this product or the support materials provided with the product.

LIMITED WARRANTY

No warranties are made, either express or implied, with regard to the contents of this work, its merchantability, or fitness for a particular use. American Megatrends assumes no responsibility for errors and omissions or for the uses made of the material contained herein or reader decisions based on such use.

DISCLAIMER: Although efforts have been made to assure the accuracy of the information contained here, American Megatrends expressly disclaims liability for any error in this information, and for damages, whether direct, indirect, special, exemplary, consequential or otherwise, that may result from such error, including but not limited to the loss of profits resulting from the use or misuse of the information contained herein (even if American Megatrends has been advised of the possibility of such damages). Any questions or comments regarding this document or its contents should be addressed to American Megatrends at the address shown on the back cover of this document.

American Megatrends provides this publication "as is" without warranty of any kind, either expressed or implied, including, but not limited to, the implied warranties of merchantability or fitness for a specific purpose. Some states do not allow disclaimer of express or implied warranties or the limitation or exclusion of liability for indirect, special, exemplary, incidental or consequential damages in certain transactions; therefore, this statement may not apply to you. Also, you may have other rights that vary from jurisdiction to jurisdiction. This publication could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. American Megatrends may make improvements and/or revisions in the product(s) and/or the program(s) described in this publication at any time.

Original Release: 09/18/2009

Table of Contents

Introduction	4
Business Continuance	4
High Availability Architectures	4
1) <i>Dual Redundant Storage Servers</i>	4
2) <i>Separate Controllers Sharing Dual Ported JBODs</i>	5
3) <i>Highly Available Mirrored Storage Servers</i>	6
High Availability Implementations in StorTrends iTX	6
StorTrends iTX HA Implementation Using Multipathing	6
StorTrends iTX HA Implementation Using IP Virtualization	7
Configuration and Management of HA Nodes with StorTrends iTX	8
StorTrends iTX High Availability in Various Environments	9
StorTrends iTX HA with Heterogeneous Servers	9
StorTrends HA with Clustered Application Servers	9
StorTrends HA in a Virtual Server Environment	10
StorTrends HA and Disaster Recovery	11
Conclusion	12

Introduction

One of the greatest challenges facing the storage industry today is coping with the rapid pace at which the underlying technology continues to evolve. Until just a few years ago, storage “silos” that mainly utilized Direct Attached Storage (DAS) architecture were the storage technology of choice, but from today’s perspective, DAS is practically “ancient” technology. In short, we now live in an age of storage virtualization and consolidation. Consequently, the storage architecture that proliferates today is often centered on Application Server Virtualization and Virtualized Network Storage servers, be they Storage Area Networks (SAN) or Network Attached Storage (NAS).

In the days of DAS, when the storage entity was serving one physical server, the demands on the storage system were not incredibly severe. By comparison, in the typical enterprise-class storage environment of today, a storage server caters storage to multiple virtual and physical application servers and is therefore subject to much more rigid requirements in terms of its reliability and performance.

The acronym “RAS”, meaning “Reliability, Availability, and Serviceability” is a general indication of how a storage server performs on this level, and the storage servers of today are expected to deliver a peak RAS score to cope with the increased demand placed on storage. No matter what, a networked storage server is expected to deliver continuous data access; it is simply unacceptable for single component failures to bring down the storage unit. This document, therefore, will discuss the central attribute of RAS, Availability, and how the new and improved High Availability (HA) architecture in the latest version of StorTrends iTX 2.7.1030 (version 3.x) from American Megatrends (AMI) can be employed to deliver outstanding Availability and keep RAS scores at their peak.

Business Continuation

In the world of data storage, the term “Availability” describes the ability to keep data online and available to the client systems without interruption. To be considered as an available storage system, one or more component failures within the storage system should not have the capacity render it “down”, i.e., completely unavailable. Therefore, the redundancy of components becomes important, making it very common for multiple power supplies, fans, and swappable hard disks to be featured in a storage system so failures of Field Replaceable Units (FRUs) can be easily managed. In this way, when one such component fails, the failed component is quickly swapped out and the storage server continues to run.

Other strategies abound to keep the system available. Implementing a RAID subsystem negotiates disk failures, and network interface failures may be dealt with by using NIC teaming. Even so, a very important question remains unanswered: what about the failure of the controller hardware of the motherboard inside the storage server itself? Like other simpler and easily replaceable components of the storage system, some redundancy or fault tolerance must also be implemented. In fact, controller redundancy is quickly becoming the key challenge of any highly available storage subsystem.

High Availability Architectures

In terms of system architecture, there are three different approaches or implementations that are generally accepted within the storage industry to negotiate the challenge of imparting resilience to failures of storage controllers. Let’s take a look at each below:

1) Dual Redundant Storage Servers

In a dual redundant architecture, two controller units are housed inside different hot-swappable canisters. These controllers connect to the same backplane / mid-plane to share the disk subsystem. In the event of a controller failure, the surviving controller takes over the role of the failed unit. Since both the controllers share the same disks, this takeover does not affect reliability or availability since the data written from the failed controller is readily available in the shared storage. This architecture has the benefit of providing the smallest space footprint and is also quite energy efficient.



Figure 1: Dual-redundant storage server

One drawback, however, of this solution is that it requires special hardware and is therefore significantly more expensive than comparable solutions. Another key disadvantage of this architecture is the shared backplane, which is still a point of weakness in the architecture. It can be argued, however, that the backplane contains only passive or “semi-active” components and therefore the probability of its failure is very low. Despite this, the fact remains that the backplane may still fail and in such a situation the entire storage subsystem will come to its knees, with an adverse effect on data availability.

A second drawback that puts this architecture in disfavor is the tight housing of the controller canisters in close proximity. These controllers, though two distinct units, are subjected to the same environmental stresses and threats. A simple disturbance such as the leaking of a fire sprinkler can completely disable the entire storage system. The same threat exists for the shared hard drives as well. Because each drive in the storage unit is subjected to identical shock and vibration conditions, any excessive vibration in the storage rack could temporarily disable functionality of the hard disks, causing the storage server to compromise its availability.

2) Separate Controllers Sharing Dual Ported JBODs

This architecture is fundamentally similar to the one described above, the main difference being that the two controllers are not housed inside the same device chassis. These controllers are essentially processing units with no internal storage in the chassis; instead, they share hard disks housed inside JBOD (“just a bunch of disks”) units using a transport fabric, typically Fibre Channel (FC) or Serial Attached SCSI (SAS), to access the shared storage.

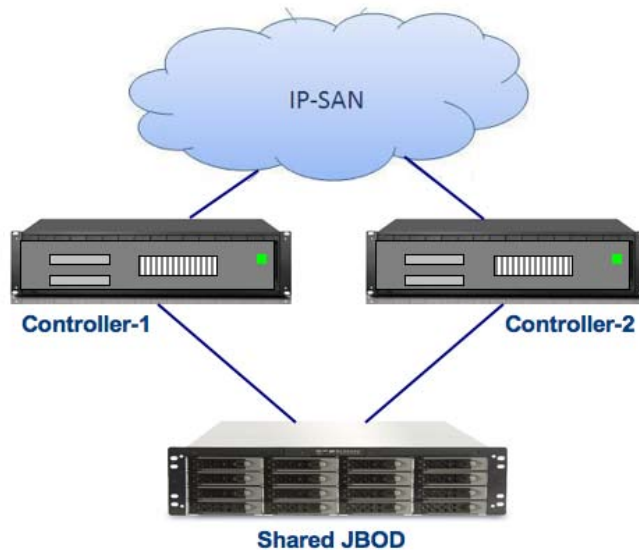


Figure 2: Redundant controllers with dual-ported shared JBOD

The advantage here is that the controller units utilize commercial, off the shelf (COTS) hardware and do not require special purpose-built hardware. Otherwise, this architecture is also plagued with similar disadvantages to the Dual Redundant Servers described earlier. If a power outage, fire, or similar catastrophe occurs, both controllers are vulnerable as they are located in the same physical location; the only difference is that the controllers do not reside in the same chassis and avoid the single point of

failure in the shared backplane.

3) Highly Available Mirrored Storage Servers

In this model, two independent storage servers are coupled inside the same SAN fabric and data is mirrored between them. Any data write that happens to the disks of one unit is mirrored online to the other unit.

From the hardware perspective, this configuration will also often take advantage of COTS units, a definite positive mark in its favor. The disadvantage of this architecture, however, is its requirement for twice the number of hard disks, which makes it much more expensive in terms of its hardware cost and adds to the space footprint and electrical power requirements.

From the reliability perspective, the upsides of this implementation are also nothing short of phenomenal. Duplication of the disk subsystem, which is independently protected by RAID, enhances the reliability by orders of magnitude. Moreover, the two units are typically housed in different buildings, isolating them from identical environmental and stress characteristics as experienced with the first two solutions. Basically, there is no shared component and as such they do not suffer from a shared risk.

One common myth or misperception about this architecture is the overhead of mirroring, and the decreased performance that could result from the need to mirror each individual data write to two separate storage servers. On the contrary, when designed properly, this implementation can in fact greatly enhance I/O performance. In a SAN environment, where multiple application servers share the same storage subsystem, the I/O load to the storage appliance is random. By contrast, a properly architected mirrored system can deliver phenomenal I/O performance, because the number of disk spindles in such implementations is doubled.

High Availability Implementations in StorTrends iTX

StorTrends iTX, the data storage software that comes pre-installed on all StorTrends storage appliances from American Megatrends, employs the mirrored server approach for its HA implementation. It is important to note here that in general there are two different ways this architecture may be realized: the first method uses a multipathing module on the client or initiator side, and the other uses IP virtualization or impersonation on the server side. Older versions of StorTrends iTX 2.7 (prior to version 1030.3.x) take the multipathing approach, while the most recent versions of the software (version 1030.3.x and beyond) take advantage of IP virtualization to implement HA.

StorTrends iTX HA Implementation Using Multipathing

The figure below illustrates the HA implementation found in StorTrends iTX, using client side multipathing. The concept illustrated here is in fact very simple, as the target side takes care of the mirroring and the client side is responsible for the failover mechanism.



Figure 3: StorTrends HA with DSM

In more detail, the StorTrends iTX HA architecture is structured so that the two storage servers create the HA pair of volumes. The client logs in to both the server units using different IP addresses as shown in the illustration above. From the client side, in order for the HA model to function properly, all of the I/O writes to the volume are first only sent to a single storage server. This server is thus designated as the primary unit for this volume. It then becomes the responsibility of the primary storage server to mirror every I/O write of the volume to its mirrored unit.

Next, the primary server then replicates this I/O online to the secondary unit before it completes the write command. If at some point during this replication process the secondary server becomes unavailable due to link or unit failure, the primary server will assume the responsibility to re-mirror or re-synchronize the secondary volume upon remedy of the situation. Alternatively, if the primary server becomes unreachable from the client due to link or system failure, then the multipathing module of the client retries the failed I/O from the alternate server.

Earlier implementations of StorTrends iTX (prior to iTX 2.7.1030 ver.3.x) are based on the Microsoft MPIO multipathing framework. In order for HA to operate smoothly for its customers using Microsoft Windows® servers, AMI provides a special module called the Disk Specific Module (DSM) to its customers. This module runs on the client system to assist in path and server selection during server failover.

Note that in HA environments where one storage server is mirroring data to another, there is a chance that each may think the other is “dead” or otherwise unresponsive due to a link failure between the units. This in turn could cause an unfavorable situation where each of the storage servers attempts to serve I/Os to the client at the same time. This phenomenon is called “Split-Brain Syndrome” and a more elaborate treatment of this subject is beyond the scope of this document. For the purposes of this discussion, it is sufficient to say here that the DSM module featured in StorTrends iTX works as the tie-breaker in these situations to eliminate the possibility of this condition.

To summarize, this architecture fits very well in the Microsoft Windows® server environment. The drawback here is that it is not platform or OS agnostic and separate implementations are to be done for each specific environment.

StorTrends iTX HA Implementation Using IP Virtualization

StorTrends storage appliances using StorTrends iTX 2.7.1030 version 3.x take advantage of a new, improved IP virtualization mechanism to implement High Availability. The advantage with IP virtualization is that there is no dependency on the client side, unlike multipathing, and therefore no special module that has to be loaded on client systems. This makes the implementation both OS and platform agnostic. It also allows StorTrends to deliver a robust HA solution across the board, even for deployments that use application server virtualization or application server clustering.

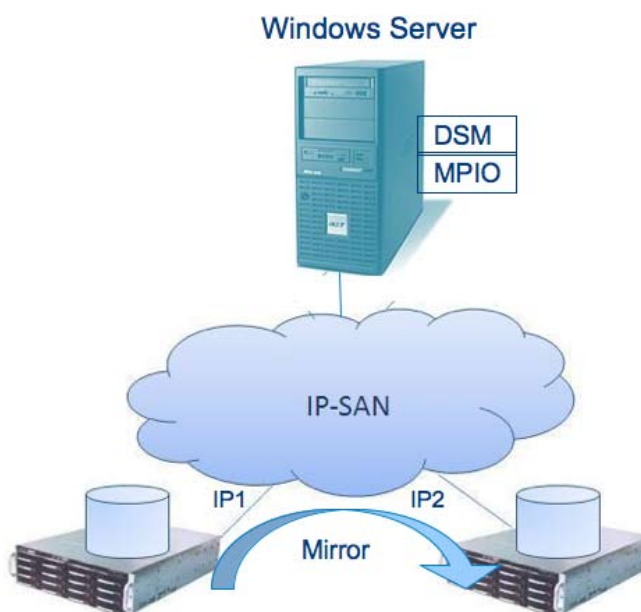


Figure 4: HA with IP Virtualization and Heartbeat Connection between Storage Appliances

As in the multipathing model described earlier, the primary StorTrends appliance assumes the responsibility of mirroring data to the secondary unit, including re-syncing as and when needed. However, any storage node failure is handled in a very different

manner than its predecessor. First, the client server logs in to the volumes in the primary storage server using a virtual IP address. In this model, if the primary storage server fails, the secondary server immediately recognizes this failure, thanks to the “heartbeat” connection that exists between the primary and secondary storage systems. Upon detection of the failure, the secondary server “impersonates” or takes over the virtual IP address. When the client server next retries the failed I/O using the same virtual IP connection, the I/O will land on the surviving storage server. StorTrends iTX employs a special technique using Address Resolution Protocol (ARP) resolution with the switches in the transport fabric to address the potential for split-brain syndrome.

Configuration and Management of HA Nodes with StorTrends iTX

Configuring and managing High Availability in StorTrends iTX is intuitive and simple. StorTrends iTX makes it easy by giving users the option of configuring via ManageTrends™, which is the Web-based management interface for StorTrends iTX, or by using Command Line Interface (CLI) commands. Note that HA configuration for SAN users requires the use of two StorTrends nodes setup for Active/Active High Availability configuration; the HA Configuration Wizard in ManageTrends helps to quickly setup the HA pair. Once the mirrored volume configuration is setup, Resource Groups are created to control the I/O flow from the clients.

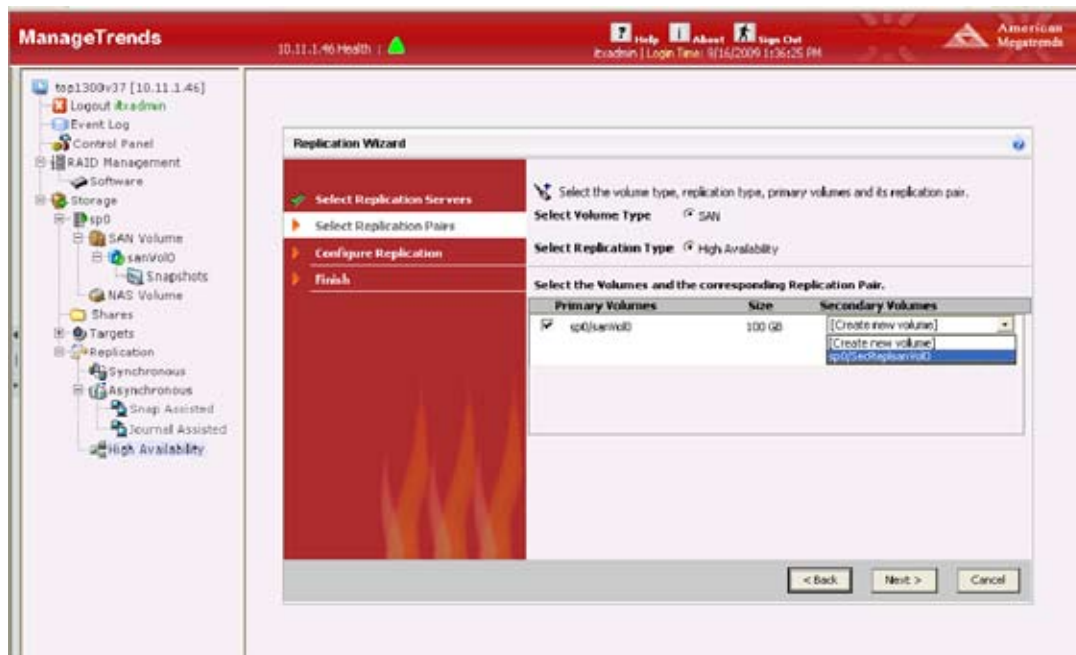


Figure 5: HA Wizard in the ManageTrends GUI of StorTrends iTX

The Resource Group (RG) is essentially a conceptual or virtual entity that defines how the clients communicate to a set of volumes using virtual IP addressing. Each RG is owned by a StorTrends node, which is designated as the primary node for that particular RG. While setting up a RG, virtual IP addresses have to be assigned and the participating volumes must be selected. An RG can support up to eight virtual IP addresses, through which an initiator can login to the volumes contained in the Resource Group. Note that each volume in a Consistency Group has to be contained inside a RG.

The newest version of StorTrends iTX supports up to two Resource Groups, while any of the nodes may own a RG. The preferred method for setting up an Active/Active configuration is to create two Resource Groups, each of which is owned by a different node of the StorTrends HA pair. This way, both the nodes actively participate in serving I/Os to the clients, thereby maximizing system performance. StorTrends allows various operations with the Resource Groups through the ManageTrends management interface; volumes may be migrated online within the RGs for load balancing, and the ownership of a Resource Group can also be controlled online.

As mentioned earlier, node failures are detected by the heartbeat module, and the secondary storage appliance takes over the virtual IP address to which the client is serving I/Os. This process is also known as automatic failover. Seen from the perspective of Resource Groups, in automatic failover the surviving node (secondary) takes on the ownership of the affected RG in the absence of the primary. I/O activities continue seamlessly, and clients may experience nothing more than a brief pause, after which I/Os will continue normally. Upon restoration of the failed node, the volumes of the RG are synchronized automatically and then the RG ownership is handed back to the original owner.

StorTrends iTX High Availability in Various Environments

StorTrends iTX HA with Heterogeneous Servers

The HA implementation in StorTrends iTX 2.7.1030 version 3.x is executed in a “client transparent” manner, which enables the HA configuration to be used in a SAN environment with various application servers running under multiple disparate operating systems. As shown below, each of the client servers may run its own independent OS and still be able to connect to the StorTrends HA pair.

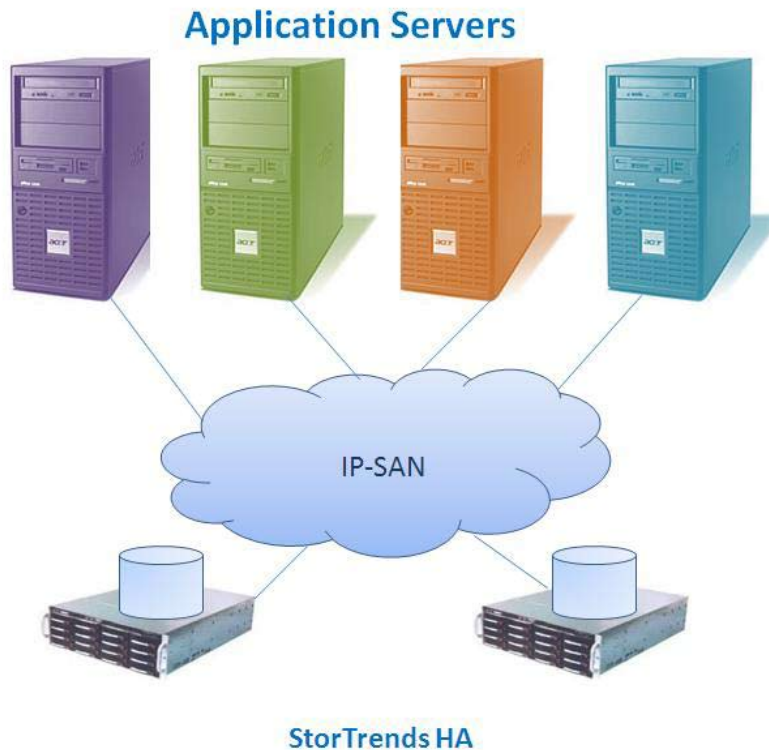


Figure 6: StorTrends HA in a heterogeneous environment

In the event of failure of the primary storage appliance, the secondary server will transparently fail over to the secondary so that all of the clients can continue their operation unhindered. Once the failure is remedied and the failed server is brought back online, the surviving StorTrends server will re-mirror the new server and restore full redundancy. At this point, the original primary server will take over the I/O load of the volume(s) that it previously owned; this is known as automatic failback.

Note that in the normal mode of operation, each storage server can assume the primary role for different volumes. In such situations, bidirectional replication actually takes place, and the pair actively participates in I/O, so that they are not necessarily working as a typical Active/Standby pair. In storage terminology, this is referred to as an Active/Active mode of operation and offers very high performance, since both storage servers play an active role in serving the storage needs.

StorTrends HA with Clustered Application Servers

The HA implementation in the most recent version of StorTrends iTX is not only designed to offer high availability for individual application servers, but also can be used in a clustered application server environment. The figure on the next page depicts this type of configuration, where two clustered Windows® servers connect to an HA pair of StorTrends storage appliances.



Figure 7: StorTrends HA with clustered application servers

The StorTrends HA pair is completely aware of the clustered clients connected to it and utilizes SCSI Reserve/Release (both SCSI-II and persistent), which is necessary to support a clustered server configuration. This enables a very robust and highly available SAN environment where any node failure (application or storage server) will not pull down the operation.

The HA configuration in StorTrends iTX can support Windows® Storage Server 2003 as well as Windows® Storage Server 2008 clustered servers. Moreover, all flavors of clustered application servers including Microsoft Exchange® and SQL servers are fully supported by the latest StorTrends High Availability implementation.

StorTrends HA in a Virtual Server Environment

Finally, StorTrends iTX HA configuration may be used to provision storage in a virtualized application server environment. The following figure illustrates a HA configuration with virtualized application servers using ESX™ from VMware:



Figure 8: StorTrends iTX HA in a VMware Virtual Server Environment

Application server virtualization is increasing in popularity at a rapid pace. In application server virtualization, a single physical server may be configured to host multiple virtual application servers, VMware ESX™ virtualization being one of the most popular methods of doing so. VMware ESX comes with many advanced server virtualization features like VMotion™ for live virtual server migration and Dynamic Resource Scheduling (DRS) to implement a very fault resilient application server environment. As shown above, these virtual servers may share the StorTrends HA pair to extend the resilience to the underlying storage subsystem. StorTrends iTX is certified by VMware to be fully compatible with VMware ESX for the StorTrends 3200 series appliances, and will also work seamlessly in other popular virtual server environments such as Microsoft Hyper-V, Citrix XenServer™, Virtual Iron and more.

StorTrends HA and Disaster Recovery

The HA implementation in StorTrends iTX can play a key role in the Disaster Recovery (DR) strategy of an organization looking to ensure the continuity and security of vital data in the face of ongoing threats to its integrity. One very effective approach is to configure a StorTrends storage appliance as a highly available storage server in the primary data center site, and use asynchronous replication with WAN to backup the SAN volumes to a remote site. The Asynchronous Replication module in StorTrends iTX features powerful WAN acceleration and de-duplication features in its WAN Data Services (WDS) component to help users get the most out of their WAN bandwidth utilization.

During normal operation, the HA node that owns the Resource Group replicates the contained volumes asynchronously according to a selected schedule. In the event of a HA node failure, the peer node takes on the ownership of the primary I/O, as well as the responsibility of asynchronously replicating the volumes to the remote side. Asynchronous replication requires the creation and scheduling of snapshots on the HA nodes. In the event of failure of one of the HA nodes, the other node will take over the job of replicating the data, from the last snapshot replicated to the remote side prior to the node failure.

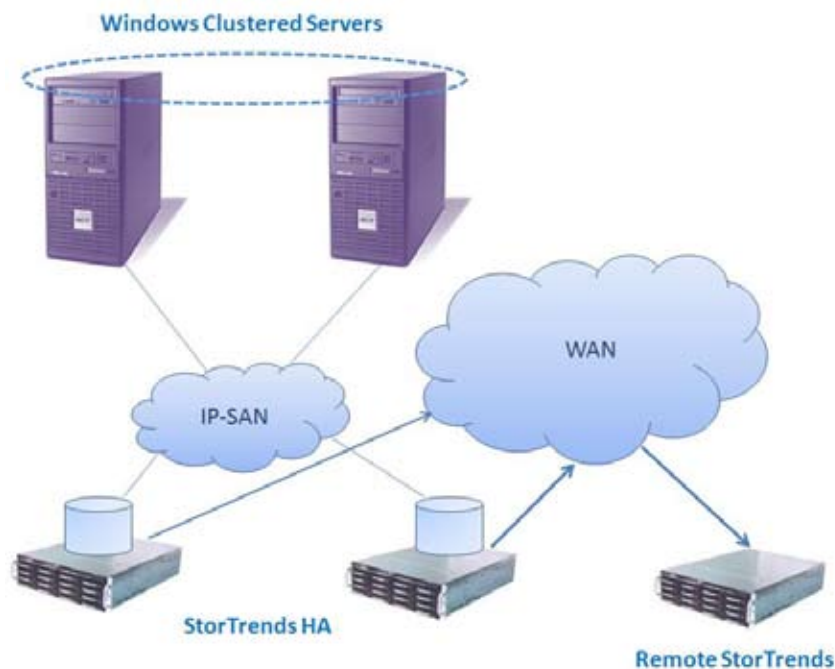


Figure 9: StorTrends HA with Asynchronous Replication using WAN

In the unlikely event that both the nodes in the primary site fail, operations can be resumed by failing over to the remote secondary site. It is important to point out here that currently, failing back to the HA primary site from the secondary is not supported; failback to one of the nodes of the HA pair can only be done after deleting the HA configuration.

Upon successful completion of a failback from the secondary to a node in the primary site, the HA configuration can once more be set up with that node. Note however that this will require a total resynchronization of data from this node. One more important caveat in using an HA pair to replicate data to a secondary site is that the secondary replication site cannot be in a HA configuration.

Conclusion

The new High Availability module of StorTrends iTX offers the utmost in business continuance and data availability for a number of different use cases. For those building a disaster recovery strategy, two HA nodes in a primary data center site can be extended to asynchronously replicate data to a remote site for a complete and end-to-end DR solution. Another attractive possibility is the use of StorTrends iTX for mirrored storage, the major advantage of which is the robust protection it offers from natural disasters. If storage servers are placed in two different locations and disaster strikes in one, data will remain safe in the other. Although a common misperception is that mirrored storage requires more disks and places this architecture at a disadvantage, a deeper look reveals that advantages of mirrored HA far outweigh its downsides. By suitably placing the two storage servers in different locations in an active/active configuration, more spindles will naturally result in higher performance.

Thanks to its elegant and thoughtful design, StorTrends iTX is truly a “single pane of glass” for the simple and easy management and protection of vital data, with an intuitive user interface that belies the power and flexibility of such a capable HA solution. User-friendly wizards and a CLI interface for more fine-grained control are available for the creation and management of HA servers, making the task of the storage administrator less daunting and perhaps even enjoyable.



American Megatrends Inc.
5555 Oakbrook Parkway, Building 200
Norcross GA 30093 | t: 770.246.8600
Sales & Product Information
sales@ami.com | t: 800.828.9264
Technical Support
support@ami.com | t: 770.246.8645
www.ami.com

This publication contains proprietary information that is protected by copyright. No part of this publication can be reproduced, transcribed, stored in a retrieval system, translated into any language or computer language, or transmitted in any form whatsoever without the prior written consent of the publisher, American Megatrends, Inc.

© 2009 American Megatrends, Inc.

All Rights Reserved