

High Availability and Failover: Ensuring Availability for IP-SAN Solutions



© Copyright 1998-2008 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: 05/11/2007

First Revision: 10/09/2007

Second revision: 04/09/2008

Table of Contents

Introduction	4
High Availability	4
No Single Point of Failure	4
<i>Redundancy</i>	5
<i>Redundancy in Detail:</i>	5
<i>Dual Box vs. Dual Controller Redundancy</i>	5
<i>Hot Swapping</i>	5
<i>Hot Spares</i>	5
Multipathing	5
<i>Hardware Load Balancing</i>	6
<i>Software Load Balancing</i>	6
Failover	6
Failback	6
<i>Failback Join</i>	6
Conclusion	6
<i>The use of HA and Failover/Failback in IP-SANs</i>	7
<i>How SMBs can Take Advantage of the Benefits of High Availability and Failover/Failback</i>	7
<i>High Availability and Failover/Failback with StorTrends®</i>	7
<i>High Availability in StorTrends iTX</i>	7
<i>Network Teaming in StorTrends iTX</i>	8
<i>To enable Network Teaming in StorTrends iTX:</i>	8
<i>Load Balancing with StorTrends iTX:</i>	8
<i>Failover / Failback in StorTrends® iTX</i>	9
<i>Clustering in StorTrends® iTX</i>	9
<i>Why AMI?</i>	9

Introduction

Organizations today are more aware than ever of how their business is exposed and can be immediately impacted by the interruption of access to critical applications and data. However, traditional data access solutions, tasked by many large organizations to provide near-continuous uptime and access to data (also known as **High Availability**) often exceed ever-diminishing budgets due to their complexity and sophistication, and can leave small and medium-sized businesses (SMBs) in particular exposed to catastrophic outages and data loss.

This document will serve to outline High Availability in detail, focusing on its implementation and operation, and review the concepts of points of failure, multipathing, and load balancing, all integral components of the foundation of high availability. This paper will also discuss failover and failback mechanisms, and their relationship to a high availability environment. The discussion will include an analysis of how high availability, failover, and failback can be implemented in storage area networks (SANs) to provide continuous operation in the face of a wide variety of threats and potential disasters. To conclude, an outline of how high availability can be made more accessible to SMB users without terrific increases in cost and expense will be provided.

High Availability

High Availability (HA) is a term used to describe the system design protocol and associated implementation that ensures a certain absolute degree of operational continuity during a given measurement period. It is measured in percentages approaching 100%, which represents the ideal, namely total availability and complete access to all data and resources at all times. Typically numbers that represent close to "five nines" availability (99.999% available) are the most sought-after.

To determine the lack of availability of a system in terms of a unit of time, first the unavailability (U) of a system is calculated by subtracting A from one:

$$U = 1 - A$$

For example, the unavailability of a "five nines" system would be calculated as follows:

$$U = 1 - 0.99999 = 0.00001$$

Since a year has an average of 8765.52 hours, a system with "five nines" availability will average 0.08766 hours or 5.26 minutes of unscheduled down time per year.

Availability (365.25 x 24)	Downtime Per Year
99.9999%	32 seconds
99.999%	5 minutes, 15 seconds
99.99%	52 minutes, 36 seconds
99.95%	4 Hours, 23 minutes
99.9%	8 Hours, 46 minutes
99.5%	1 day, 19 hours, 48 minutes
99%	3 days, 15 hours, 40 minutes

Table 1: Breakdown of Availability Calculations in Terms of Downtime Per Year¹

A reminder must be made with regards to the fact that there is a difference between availability and uptime. In short, a system can be "up" or functioning, but unavailable, such as in the case of a lost network connection. In a case such as this, the server is still operating, but no way exists to reach it, read from it, or write to it. As a matter of protection against this kind of unavailability or inability to be reached, the only way to resolved it is through redundancy, in eliminating the larger potential points of failure. In contrast to the situation described above, how does a solution provide near complete availability? Two elements in particular underpin this phenomenon. The key elements of a high availability implementation are:

- No single point of failure
- Multipathing (MPIO) and Load Balancing support

No Single Point of Failure

This meaning of this term is rather straightforward, and describes the critical need for ensuring the availability of a system is by eliminating the isolated points whose failure could bring the system to its knees. Some of the typical system components in a storage system that are seen as potential points of failure are power supplies, cooling fans, hard drives, and NIC cards / network

1 Gary Audin, "Reality Check On Five-Nines," May 19, 2002: http://www.bcr.com/management/networking_intelligence/reality_five_nines_20020519301.htm

connections. These items must therefore be duplicated or reinforced in such a way that if one item fails, operation can be continued because an identical component is in place to take over at the moment of failure.

A system with no single point of failure typically incorporates three characteristics: *Redundancy*, *Hot Swapping* and *Hot Spares*. Each of these concepts is described below in more detail.

Redundancy

Redundancy removes the single point of failure by making available an identical appliance or one or more of its components to take its place. Whether this is a complete duplicate server appliance, a duplicate, identical controller, fan, hard drive, or even a duplicate network connection, the idea is the same. The two components are interchangeable, and the one not in use can quickly be brought into use should the active component or communications link fail or otherwise be rendered useless.

Redundancy in Detail:

Because no single point of failure exists, a solution that incorporates these countermeasures provides protection against disk failures, path failures and node failures. In the event of node failure, failover action is expected to be automatic and seamless. When the failed node returns, it must be detected by the stack and synchronized quickly to bring back redundancy to the setup. Synchronization should therefore be efficient and ensure that only out-of-sync data is resynchronized.

Dual Box vs. Dual Controller Redundancy

In the storage industry, different manufacturers provide redundancy in different ways. Some have applied redundancy at the controller level, designing appliances with two controllers inside a single box. Other manufacturers have opted against this approach, arguing that a single box can fail at other points beside the controller, and instead opted for redundancy of the entire server box itself. This latter method of providing high availability, as an active/active or active/passive configuration of two storage boxes, has some benefits over dual controller redundancy as a mechanism for availability, as follows:

- *Lower exposure when a failure occurs.* A RAID-5 drive failure creates a window of vulnerability in a system's redundancy in a single-box solution, whereas a two-box solution has both RAID-5 and box-to-box redundancy.
- *Protect against failure of non-hot-swappable components.* If, for example, a backplane fails, box-to-box redundancy allows the system to continue functioning.
- *Faster rebuild.* Since the resynchronization mechanism between the two boxes tracks the sectors that have gone out of sync, rebuilding is faster than a RAID-5 rebuild.
- *Multi-site / remote capability:* It is possible to have the two mirrors at different physical locations to protect against site disasters.
- *Simple box upgrades and maintenance:* It is possible in this arrangement to remove a box for maintenance and upgrades and reinsert it into the system without having to compromise on availability.

Hot Swapping

Hot swapping builds on the idea of redundancy, by giving a system's users the ability to replace a system's failed components on-the-fly without taking the system down. In the case of a hard disk, for example, a faulty drive can be removed, and a new drive inserted while the system is in operation. Beyond this, ideal implementations of hot swapping should recognize the new component and begin utilizing it with a minimum of time or effort on the part of the user.

Hot Spares

Hot spares also build on the idea of redundancy; in fact this concept is quite similar to that of hot swapping. A hot spare is a complete system (in this case a storage server appliance) that is connected and powered up (i.e., "hot") as part of a larger storage network. If and when a key component (hard drive, power supply, NIC card, etc.) fails, the hot spare is then brought into operation. Until it is brought into use via a failover, hot spare equipment is kept powered on, but typically not actively functioning in the system. In general, in order to enable the hot spare to begin functioning immediately at the time of failover, data that is being written to the primary is also "mirrored" to the secondary. In this way, when the hot spare is brought online, the system continues to function with no real impact on operations or performance.

Multipathing

Multipathing solutions use redundant physical path components, such as adapters, cables, switches and interfaces, to create logical "paths" between the server and the storage device. In the event that one or more of these components fails, causing the path to fail, multipathing logic uses an alternate I/O path so that applications can still access their data.

In an iSCSI IP-SAN device, for example, a hard disk is connected to two redundant controllers within the same system. In the event that one controller fails, the system can automatically and transparently route I/O throughput to the second controller, and data writes and reads continued uninterrupted.

It is important to note that in multipathing, the redundant physical paths can be leveraged to improve the performance of the overall system, and enable load balancing, another critical component of High Availability solutions, and the next topic in our discussion.

Load Balancing

Load Balancing is the component of a HA solution that ensures that servers are not overwhelmed to the point of being unable to function properly. It does so by spreading out or distributing (“balancing”) I/O traffic over redundant switches or controllers, to achieve the maximum, steady throughput.

A load balancing service or solution is often referred to as a “director,” reflecting the load balancer’s role in managing connections between clients and servers. Load balancing can be achieved in several different ways, and both hardware and software solutions exist. The deciding factors for choosing one over the other typically depends on the overall availability requirement, desired features, complexity, and cost.

Hardware Load Balancing

Hardware load balancers route I/O traffic to various nodes in a storage network. They are extremely reliable and guarantee high availability, but the downside of hardware load balancing is higher cost for adoption and implementation. Because of this, hardware load balancing is encountered less frequently than the software-based variety.

Software Load Balancing

The majority of load balancing is done through software-based approaches, and is often combined with additional high availability functionality in a software stack. The major benefit of software load balancing is its lower cost when compared to hardware load balancing, along with its flexibility in terms of its variable configuration in accordance with user requirements. The major drawback of software load balancing is that in many cases it needs to be hosted on a dedicated hardware unit in order to function properly in its role as “director”.

Failover

Failover is a process by which a passive, secondary system is made active in the event of an irrecoverable failure of the primary system. If the system is using snapshots for data protection, during a failover operation, the secondary system will rollback to the latest available snapshot or snap group across all volumes. After a failover, the passive box will become the primary, active unit, until the point where a failover is performed to reverse the roles back to their original state. In the case of an IP-SAN volume, the iSCSI initiator can connect to the volumes in the consistency group and start performing I/Os. NAS volumes will be mounted and shares will be made available after a failover from the newly active / primary system.

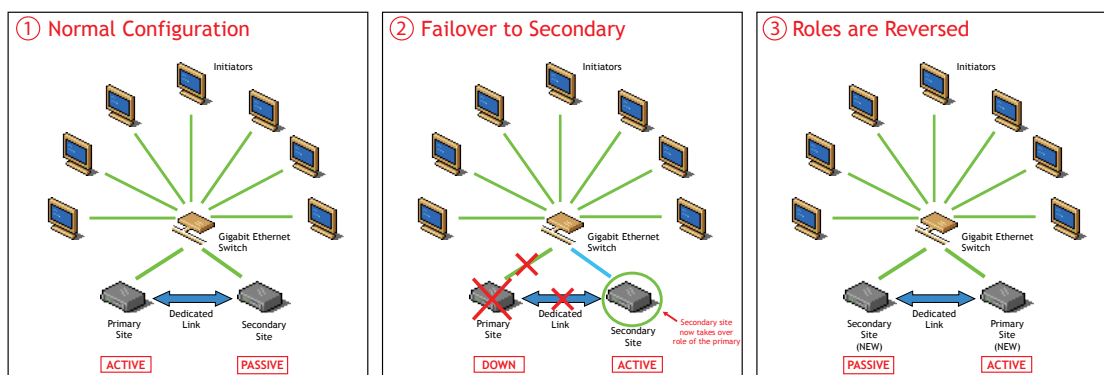


Figure 1: A Typical Failover Operation

Failback

As mentioned briefly above, failback is a process that is typically initiated after a failover, in which the current passive system regains its original active, primary role from the current active (former secondary) system. Often the administrator prefers to return the new passive server to its original primary / active role after a failover due to administrative or logistical reasons. Failback is essentially an identical, reverse operation of failover, and all the steps and safeguards present in failover are also done during a failback. In particular, if snapshots are in use, failback will only be initiated after all snapshots in the active system are replicated to the passive system. Upon completion of failback, the original primary system is once again active, and the original secondary system returns to its passive role.

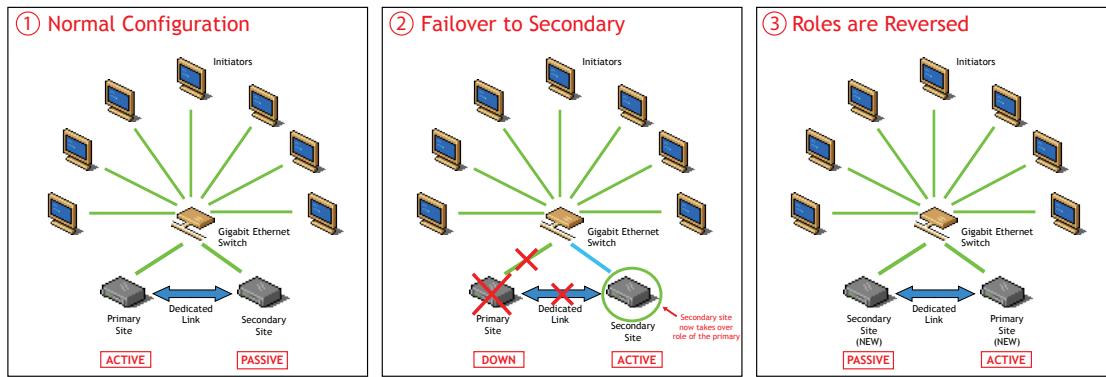


Figure 2: A Typical Failback Operation

Failback Join

Failback Join is an elaboration of the failover process, unique to storage systems that utilize snapshots to enhance data protection. In a failback join, the old active system is joined, or synchronized with the new active system after a failover by migrating snapshots from one to the other. The process is quite simple: in essence, the old active system simply becomes passive and starts receiving snapshots from the new active system in order to match the contents of their data. In such scenarios, the most recent matching snap group is identified between the old active system and the new active system, and Snapshot-Assisted Replication is initiated from that point onwards from the new active system to the old one, in order to synchronize their contents.

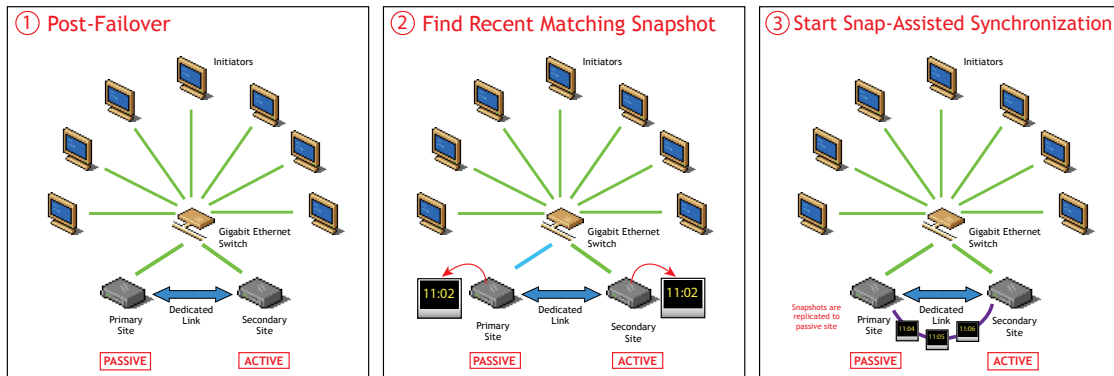


Figure 3: Failback Join Configuration

In conclusion, this document outlined the concepts and components that comprise a High Availability solution. It provided an overview of how availability is traditionally measured, and discussed the interplay between the components of HA. A detailed discussion was made of how the concepts of No Single Point of Failure, Load Balancing, Multipathing, and Failover / Failback are all integral parts of an effective HA configuration. A survey was also made of how these various concepts combine and leverage each other's strengths to ensure that a system remains in operation in the face of component mishaps.

Next, to finalize this discussion, a brief analysis of the specifics of how High Availability is addressed by IP-SAN appliances will be given, along with notes on how SMB users can take advantage of these enterprise-level features in their own IP-SAN devices to protect themselves from component failure.

The use of HA and Failover/Failback in IP-SANs

In IP-SAN storage appliances, synchronous replication is undeniably the choice for highly critical operations, because it offers the benefit of zero RPO (Recovery Point Objective). However, when the RTO (Recovery Time Objective) becomes equally important, an active-active mirroring configuration is the preferred approach, because of its ability to instantly recover from disaster.

In a HA configuration, with its the primary and secondary storage servers *both* field I/Os from the initiator. However, they typically service different logical block addresses (LBAs) in each box. The distinction between 'primary' and 'secondary' servers becomes somewhat blurred in this configuration – for different zones, both servers may act as the primary, mirroring data to their peer in a synchronous fashion. An agent on the initiator takes care of shipping the right I/Os to the right server, thereby saving on extra hops.

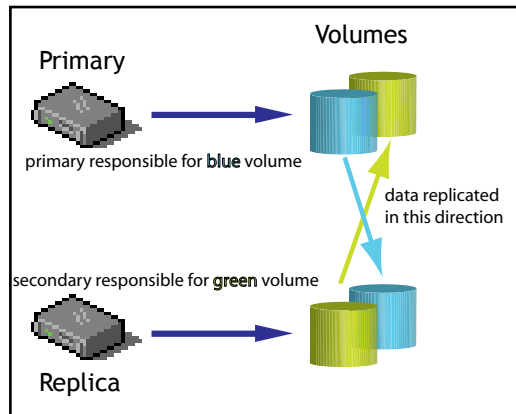


Figure 4: High-availability Configuration with Synchronous Mirroring

When a disaster occurs on one of the machines subsequently, the machine that is still up takes over the primacy for zones that were owned by the failed machine in a seamless manner. The agent running on the client side initiates this process, when an I/O failure is detected.

How SMBs can Take Advantage of the Benefits of High Availability and Failover/Failback

As stated earlier in this paper, until recently HA solutions have typically been targeted at the enterprise-level user. However, falling prices of hardware combined with creative software implementations have made these solutions more accessible to SMB users. By sticking to COTS (commercial, off-the-shelf) hardware and letting the system software handle HA and failover/failback management, rather than adopting a more costly hardware-based solution, High Availability comes into the reach of smaller organizations with more modest budgets.

High Availability and Failover/Failback with StorTrends®

Although the goal of High Availability is to recover critical applications, the reality is that generally speaking, Operating Systems are stable and Critical Applications are reliable. Regardless, the scope of a solution must have the ability to encompass each IT infrastructure element on which these applications depend.

Organizations both large and small are looking for an alternative way to achieve high levels of availability to data without the costly expense of additional servers and software licenses needed for Operating System/Application protection. StorTrends® appliances from American Megatrends solve this problem by enabling high availability of data through continuous data accessibility at much lower cost than ever before. StorTrends does this by capitalizing on a software-centric approach to HA and failover/failback management, implementing hardware approaches to only where it is absolutely necessary, for example, to address redundancy issues. This type of cost-conscious approach is something that should be attractive to nearly all organizations, no matter what their size.

The StorTrends® iTX solution enables:

- * Zero downtime of access to data, even in the unlikely event of complete failure of the appliance
- * An RTO and RPO of zero
- * Application transparency at the I/O level during failover to the secondary appliance
- * The elimination of the strict requirement for Application failover capabilities in most High Availability implementations

High Availability in StorTrends iTX

StorTrends Storage appliances from American Megatrends ensure no single point of failure through redundancy, hot swapping, and hot spares. Power supplies, cooling fans, and hard drives are all hot swappable components in the 3U StorTrends storage appliances. Network connectivity is made redundant through the use of two NIC cards.

Some of the other High Availability characteristics of StorTrends Storage Solutions with iTX storage software are:

- Active-Passive and Active/Active modes of operation
- Only for iSCSI volumes
- Automatic Failover/Fail-back when link or power failure happens
- HA works with iTX DSM and notifies the state of its nodes
- Works alongside SAR – same volume can be involved in SAR & HA
- Failover Mode for two iTX nodes
- Volume level *Load Balancing* across multiple paths of same node.
- Handles “Split-Brain Syndrome” by permitting Failover only if the secondary is totally in-sync at the time of primary failure

- Target boxes notify the DSM whenever they go out of sync and are synchronized.
- Load Balancing & Failover Settings through the registry

High Availability in StorTrends is linked closely to the system's advanced replication technology. A simple Replication Wizard is used to enable HA, and the steps involved are quite user-friendly. To enable HA, all a user needs to do is first select the primary box where the volume to be replicated is located. Next, High Availability is selected as the secondary server option. It is important to remember here that for HA the remote server must be a separate box, and must have the exact same specifications as the primary StorTrends server. The Replication Wizard will then configure the two appliances as a High Availability pair. Finally, StorTrends iTX will display a High Availability Management Page for managing the HA pair once configuration of the appliances is complete.



Figure 5: StorTrends iTX High Availability Management Page

After a HA pair has been enabled through StorTrends iTX, some of the features that are viewable and/or manageable functions in High Availability are:

- **Availability Display:** Table shows all existing HA pairs, along with Local Volume, Remote Volume, remote Host, Role, Link Status, and Synchronization Percent
- **Local / Remote Pool**
- **Local / Remote Volume**
- **Volume Size**
- **Replication Role:** Primary or Remote
- **Connection Mode:** Auto or Manual Synchronization
- **Resynchronization Priority:** Allows user to customize resynchronization as a percentage value. 25% percent is the default setting. Setting the resync priority closer to 100% makes the volumes more current, however it uses more system resources (CPU and memory) when the setting is higher.
- **Remote Host:** This field displays the IP address of the Remote Host. In the event that this changes (such as when using DHCP), the new IP address can be updated by typing it into this field. In the case of a local replication pair (involving 2 local volumes in the same box), this column will show "local".
- **Join:** Join allows the user to join a split replication pair. If a replication pair has been split, the connection between the two storage appliances can be reestablished.
- **Split:** Split allows a user to temporarily disconnect a replication pair. In a split, the connection is broken between the local and remote storage appliances. This function is useful when performing maintenance on the network or one of the storage appliances.
- **Delete:** The deletion of a replication pair is a straightforward procedure. After deletion, the volumes will be considered separate volumes, but the volumes themselves will not be deleted.
- **Update:** By clicking Update, the Resync Priority and Remote Host fields will be updated with any changes.

Network Teaming in StorTrends iTX

Network teaming in StorTrends iTX is a robust technology that capitalizes on the two NICs that are present in each appliance for the purpose of redundancy. This arrangement allows for both failover and for multipathing (MPIO) between Multiple NICs.

To enable Network Teaming in StorTrends iTX:

Setting up network teaming in StorTrends iTX is a simple affair, yet once completed, contributes greatly to the ability of the appliance to keep the network link alive. Prior to configuring network teaming, both NICs must be connected to the same subnet. However, it is important to note that whether NIC-1 or NIC-2 is selected as the primary NIC, the IP address of the network team will be that previously assigned to NIC-1.

In StorTrends iTX, the following teaming modes are supported:

- *Balanced Round Robin*: This option transmits packets in sequential order from the first available slave-NIC in the team through the last, and provides load balancing and fault tolerance.
- *Dynamic Link Aggregation* (IEEE 802.3 AD, requires switch configuration): This mode supports the creation of aggregation groups that share the same speed and duplex settings. Note that it requires a switch that supports IEEE 802.3ad dynamic link aggregation.
- *Balance TLB* (Transmit Load Balance): This mode (adaptive transmit load balancing) is channel bonding that does not require any special switch support. The outgoing traffic is distributing according to the current load (computed relative to the speed) on each slave-NIC in the team.

Load Balancing with StorTrends iTX:

Load balancing is also a key component of the HA capability of StorTrends iTX. The following list covers the different load balancing values that can be set in the StorTrends registry:

- *LB_FAILOVER_ONLY*. For any I/O failures it switches the path/box.
- *LB_READ_REMOTE*. All the Writes will be redirected to current primary and Reads from the current secondary. (*This is a special mode for use in Data Integrity QA tests*)
- *LB_WRITE_CURR_PRI*. All the Writes will be load balanced with all the paths available to the current primary.
- *LB_READ_CURR_PRI*. All the Reads will be load balanced with all the paths available to the current primary.
- *LB_BOTH_CURR_PRI*. All the Reads / Writes will be load balanced with all the paths available to the current primary.
- *LB_ALL_AVAILABLE*. All the Reads will be load balanced with all the paths available from both boxes and Writes will be load balanced with all the path available to current primary.

Failover / Failback in StorTrends® iTX

- *Failover*: Failover allows for manual switching of the roles of the primary and remote volumes. This is useful in the event of a failed primary storage appliance. Failover can only be performed from the remote appliance.
- *Failback*: Failback allows for manual switching of the roles of the primary and remote volumes to their original state after a failover. Failback can only be performed on the original primary storage appliance (or its replacement, in the event of a catastrophic failure).
- *Failback Join*: After returning the primary and remote volumes to their original state in a failover, join allows the user to join the split replication pair, and reestablish the lost connection between the two storage appliances.

Clustering in StorTrends® iTX

StorTrends uses a Microsoft® Cluster shared-nothing model to correctly identify and react to disaster. In this configuration, it is possible to use the StorTrends iTX synchronous replication stack to provide high availability for a wide variety of applications. One of the added benefits of using the shared-nothing model is that there is no Digital Lock Manager (DLM) managing access to resources, so the performance drains of additional traffic between nodes and serialized access to hardware that drain system performance when a DLM is in use are avoided.

Why AMI?

AMI offers a wide array of disaster recovery and high availability solutions for your business needs. We provide services that range from storage needs analysis to the design and implementation of a custom disaster recovery solution. We can help your business plan for when things are at their worst while reduce costs and complexity of your storage environment. For more information on AMI StorTrends solutions, visit www.StorTrends.com, email to sales@ami.com, or call (800) U.Buy.AMI.

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.

© 2008 American Megatrends, Inc.

All Rights Reserved



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