

# Technical Bulletin

## Designing High Availability Switched LANs



ZNYX's Fast Ethernet adapters with embedded RAINlink technology can be used in conjunction with Fast Ethernet switches to build a high availability LAN that provides continuous service in the event of any LAN component failure. The purpose of this bulletin is to provide network managers with design guidelines for construction of a high availability LAN and an understanding of failover scenarios. The design described in this paper can be applied to many applications including

- Client-server
- Peer-to-peer
- Load sharing and failover clusters
- Web farms
- Messaging and voice mail
- Virtual Private Networks (VPNs) and firewalls
- General purpose AIN platforms

The diagram below shows the topology of the high availability LAN. Each node on the LAN contains a ZNYX RAINlink-enabled Fast Ethernet adapter. The switches shown in the diagram can be any Fast Ethernet switch. The switches must contain an interconnect between SWITCH A and B. When performance is not an issue, any speed link can be used as an inter-switch link. Practically speaking, the link should be a 100 Megabit trunk or redundant Gigabit links to guarantee throughput for multiple 100 Megabit connections and redundancy.

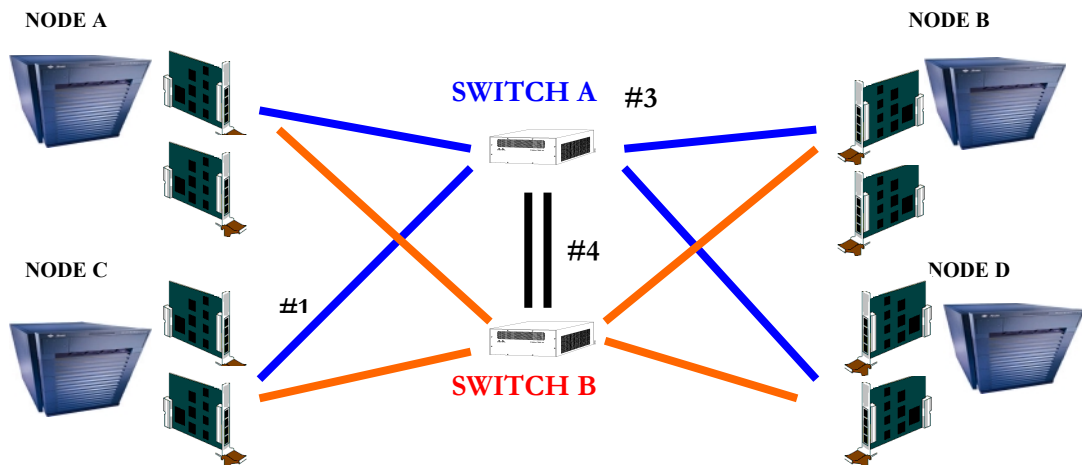


Figure 1. High Availability LAN

In the above diagram, each node on the LAN communicates with its peers using the primary links attached to either one of the switches. If any failure occurs, RAINlink will switch from the primary to the secondary link.

## **Fast Deterministic Failover**

RAINlink's link failure detection is interrupt driven. The time between a link failure and completion of a failover transition can vary somewhat on operating systems and processors but generally happens within a few milliseconds; typically less than 10ms on a 300 MHz system. RAINlink initializes the secondary link with the MAC address of the primary link. Auto-Negotiation is completed to put the secondary link in a hot standby state. This hot standby procedure insures that the fail over transition occurs quickly. Neighborhood ARP caches are preserved. A few packets may be lost but end users will be unaware that a failure occurred; the upper layer protocols will automatically re-send the lost packets.

## **Primary and Secondary Links**

When RAINlink drivers bring up the links, the first active link becomes the primary link. The second link to become active becomes the secondary link. Thus SWITCH A and SWITCH B may both be part of the primary communication path. Some applications could benefit from a static notion of primary and secondary links. However, selecting the first active link as the primary link reduces configuration complexity and adds an amount of determinism during boot and power up sequences. Moreover, this approach is more manageable as the high availability LAN grows. Redundant links from any node can be plugged into the switches without having to consider and configure primary links, cables, and switches. This approach to High Availability LAN configuration can be characterized as "plug and play".

## **Fault Scenarios**

These are the most common system failure scenarios.

### ***FAILURE #1 – PRIMARY LINK FAILURE***

If a link failure occurs at #1 (or any equivalent point on another node) RAINlink detects the failure and switches to the secondary link. The traffic is routed through the secondary link to SWITCH B. SWITCH B then forwards the packets to all ports including the link between the two switches. SWITCH A sees that NODE C has changed from the original port to the inter-switch link. In most switches, this is an instantaneous change (see Spanning Tree Issues below). The primary link of other nodes still communicates through their primary links. The secondary link of NODE C receives this traffic.

RAINlink supports a Watchdog Timeout mechanism that can detect failures that are not hard link failures. This includes malfunctioning adapter ports or switch ports that send error or malformed packets. Good packets are required to reset the Watchdog Timeout timer.

The Watchdog Timeout must be used with some sort of a heartbeat generator such as a ping generator or application.

#### ***FAILURE #2 – ADAPTER HARDWARE FAILURE***

An adapter failure scenario can be dealt with by using redundant adapters. RAINlink can transparently failover from one link to another whether those links are on the same adapter or different adapters. The switches will perform in the same manner described in FAILURE #1.

#### ***FAILURE #3 – SWITCH FAILURE***

If SWITCH A fails, RAINlink will failover from the primary links attached to SWITCH A to the secondary links attached to SWITCH B. The adapters that had their primary links attached to SWITCH B will not failover. If the switch is powered down or reset RAINlink will immediately sense that the links are down and all nodes will start communicating over the secondary links through SWITCH B.

If the switch fails in such a way that it keeps the links up but does not forward traffic, the Watchdog Timeout feature of RAINlink will switch to the secondary links after the user specified timeout interval (default is 20 seconds, it can be set at millisecond intervals).

#### ***FAILURE #4 – INTER-SWITCH LINK FAILURE***

The inter-switch link consists of a multi-channel trunk. The multi-channel trunk implementations provide high aggregate throughput capability, load balancing, and active redundancy. If a member of the trunk fails, then the other members assume the load. The trunk implementation eliminates the inter-switch link as a single point of failure.

## **Spanning Tree Issues**

The function of 802.1d Spanning Tree Protocol (STP) is to remove redundant paths on a bridged or switched network. If STP is running between the two switches the failover scenario is gated by the STP timeout. The switches will not pass traffic from a failed node until after the STP timeout has occurred. This is usually about eight seconds.

In the topology described above it is not necessary to run STP. Because RAINlink only uses one link at a time, there is never a redundant path that needs to be removed from the tree topology. Most switches, running without STP, will recognize the secondary link is active and adjust their internal tables automatically. The primary switch will see the MAC address of the failed node coming in through the inter-switch link and update its address table instantaneously.

To achieve fast deterministic failover it is best to run without STP on simple topologies such as the one illustrated above.

## Switch Assumptions

It is assumed that the Ethernet switch used in this scenario supports a virtually instantaneous table adjustment when a MAC address is moved from one port to another. This appears to be true for most of the newer switches when STP is not activated.

## RAINlink Enabled Drivers

At the time of this writing the following operating systems are supported (or will be in the near future) with RAINlink-enabled drivers:

- Microsoft Windows NT
- Sun Solaris
- Linux
- Lynx LynxOS
- Integrated Systems pSOS+
- WindRiver VxWorks

Please visit [www.znyx.com](http://www.znyx.com) for the latest information on RAINlink-enabled drivers.



ZNYX Corporation  
48501 Warm Springs Blvd., Suite 107  
Fremont, CA 94539 USA  
(510) 249-0800  
Fax (510) 656-2460  
[www.znyx.com](http://www.znyx.com)

Document # DU0872-02

© 1999 ZNYX Corporation. All rights reserved worldwide. All information in this document is subject to change without prior notice. ZNYX, RAIN, and RAINlink are trademarks or registered trademarks of ZNYX Corporation in the United States and/or other countries. All other marks, trademarks or service marks are the property of their respective owners.

ZNYX may have patents, pending patent applications, trademarks, copyrights or other intellectual property rights covered in the subject matter of this document. By furnishing this document, ZNYX does not license nor waive its license to those intellectual property rights except as expressly provided in a written license agreement from ZNYX. Information in this document is subject to change without prior notice.