

Applying IEEE 802.1w (RSTP) to Improve Service Availability

James T. Yu, Ph.D.

School of CTI, DePaul University

jyu@cs.depaul.edu

This paper presents a new approach to improving the availability of redundant servers, which include a primary server and a backup server[1]. The current fault tolerant schemes for redundant servers cover failure cases for the servers themselves, but do not provide a cost-effective solution to network failures[2]. Examples of the current approach to addressing network failures are described as follows:

1. **Single Ethernet Switch:** This scheme simply ignores network failures. If the Ethernet switch goes down, it is a *total service outage*. Although this is not a fault tolerant scheme, it is probably the most common practice used today.
2. **Dual Ethernet Switches:** This scheme has two Ethernet switches where each server connects to only one Ethernet switch. If there is a network failure, the backup server takes over. This scheme creates a large failure group where a single failure could cause the whole group of nodes out of service.
3. **Redundant Link Connections:** This scheme also uses two Ethernet switches, and each server has dual links to both switches as illustrated in Figure 1.

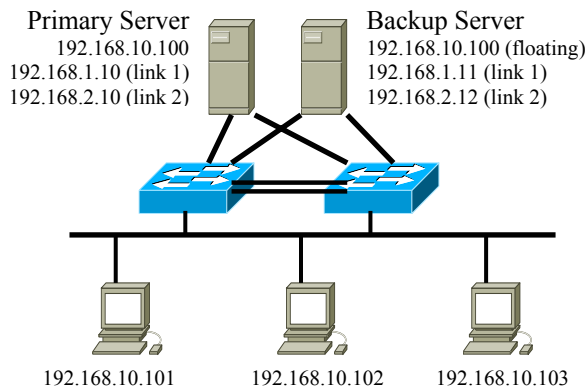


Figure 1. Current Fault Tolerant Architecture

The redundant link architecture in Figure 1 offers the best service availability of three. However, this solution requires a complex IP address scheme and advanced management software running on the server. For example, each server needs to run a *heartbeat* program to constantly audit the link integrity to each Ethernet switch. Although this solution offers high availability,

the equipment cost and on-going support cost are relatively high.

The approach presented in this paper is based on the new IEEE 802.1w standard [3] which is a major improvement over IEEE 802.1D [4][5], also known as the Spanning Tree Algorithm and Protocol (STP). The need for STP is to address the issue of loop topology on a switched Ethernet network. STP transforms a loop topology to a loop-free topology (i.e., a tree topology) by blocking certain ports so that broadcast messages would not loop around the network (known as broadcast storm). This technique provides a fault tolerant scheme for Ethernet networks. However, STP suffers two major issues: long failover time (45-60 seconds) and total network outage during recalculation of STP. As a result, STP is rarely used in practice. The new IEEE 802.1w standard is also based on the Spanning Tree Algorithm; but it does not need to recalculate the Spanning Tree Algorithm when topology changes occur. It has local *intelligence* so that it knows how to move a port from the **blocking** state to the **forwarding** state while maintaining a tree topology. It has been demonstrated that 802.1w could achieve a failover time in 1-3 seconds. It is even possible to achieve sub-second failover time for physical link failures (e.g., loss of signal). Because of the fast failover time, IEEE 802.1w is also called Rapid Spanning Tree Algorithm and Protocol (RSTP).

We propose to apply the new RSTP standard by sending the server traffic to both Ethernet switches. With RSTP, the port on one of the switches is automatically blocked. As a result, the receiver does not see any duplicate frames. In the case of a network failure (link failure, port failure, or switch failure), the blocked port is changed to the **forwarding** state immediately and the traffic could resume in 1-3 seconds. In the experimental phase of this study, we use an Ethernet *hub* to split the server traffic to both Ethernet switches. The network diagram is illustrated in Figure 2. An alternative is to use a RJ-45 Y-connector to achieve the same function. We understand that a hub introduces another single point of failure, and it may not be appropriate for commercial use which is discussed later in this paper.

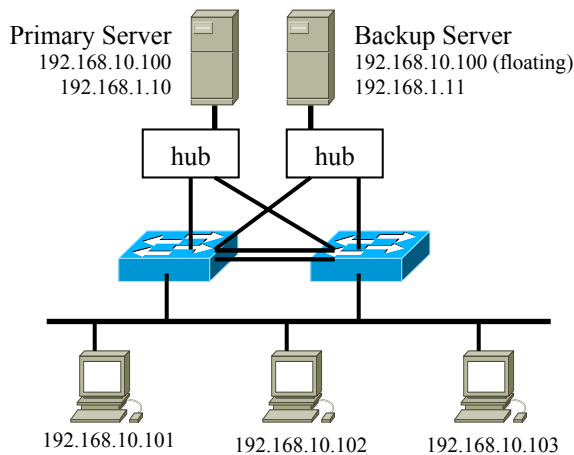


Figure 2. Fault Tolerant Architecture base on RSTP

As illustrated in Figure 2, the new fault tolerant scheme is summarized as follows:

1. The two Ethernet switches are directly connected with a pair of links. We apply the Link Aggregation technology (IEEE 802.1ad) to create a single logical link for RSTP. This logical link is given the weight so that the ports are always in the **forwarding** state.
2. Each server uses a hub to split the traffic to both Ethernet switches. Because of the loop topology created by the hubs, RSTP comes to the picture and blocks certain ports. The blocked ports are called **alternate** ports as defined in RSTP.
3. If there is a network failure (link, port, or switch), an **alternate** port changes its role to become a **designated** port as specified in RSTP. When a port becomes a designated port, it is changed to the **forwarding** state immediately. As a result, the server traffic continues to and from the clients with little interruption.

The advantages of this RSTP-based fault tolerant scheme are summarized as

- This failover due to network failures is transparent to the servers. If there is a network failure (link, port, or switch), the hub diverts the traffic to the **alternate** port. Servers do not need to do anything, so there is no need for costly management software to monitor the network.
- This is a layer-2 solution, which is transparent to upper layer functions. For example, the IP layer maintains the same *socket* during failover, so there is no interruption to any services. The solution illustrated in Figure 1 requires a new *binding* of the IP address to the backup Ethernet interface. As a

result, all active applications need to reestablish new sockets to resume the service.

- The proposed solution is considered the most cost-effective solution to address network failures because of the low cost of hubs or Y-cables. If the prototype is proved to be effective for commercial use, we expect to see dual ports Network Interface Cards (NIC) with built-in hub functions. The NIC needs only one Ethernet address and can be cost-effectively manufactured.
- Because of the fast failover time, RSTP could work seamlessly with upper layer protocols to provide a complete fault tolerant solution. For example, we can run RSTP on the switches and VRRP (Virtual Redundant Routing Protocol) on the servers. VRRP (a layer-3 protocol) takes 3-5 seconds to recover from a typical failure scenario[5], and RSTP takes only 1-3 seconds. The faster failover scheme at the lower layer prevents any cyclic effect. Note that this scheme is impossible for STP which takes 45-60 seconds.

The proposed solution does have limitations and constraints, and we need further studies to address them. One issue is the use of a *shared* media technology at the server side. As a result, the link between the server and the switch runs in a *half duplex* mode, and it effectively reduces the performance by 50% for applications that need to send and receive data simultaneously at a high speed. Another issue is on the Ethernet cable specifications. We do not expect issues with 10BaseT and 100BaseT networks. We are, however, not sure about 1000BaseT network where the use of shared media may cause many collisions. To address these two limitations, we plan to study and experiment on the applicability of running bridge software[7] and RSTP on multi-port NIC or multiple NIC on the server.

References

- [1] Alan Wood, "Predicting Client/Server Availability," *IEEE Computer*, April 1985, pp. 41-48.
- [2] A. K. Somani and N. H. Vaidya, "Understanding Fault Tolerance and Reliability," *IEEE Computer*, April 1997, pp. 45-50.
- [3] IEEE 802.1w, <http://standards.ieee.org>
- [4] D. Venkatesuli and T. A. Gonsalves, "Efficient Fault-Tolerant Reliable Broadcast in an Extended LAN," Proceedings of the 22nd IEEE Conference on Local Computer Networks 1997.
- [5] IEEE 802.1D, <http://standards.ieee.org>
- [6] IETF RFC 2338, Virtual Router Redundancy Protocol, April 1998. VRRP is also known as Hot Standby Routing Protocol (HSRP) in the Cisco environment.
- [7] <http://www.linux.org/docs/ldp/howto/BRIDGE-STP-HOWTO/>