

Achieving Superior Resiliency and Fast Convergence with Foundry's MRP

Introduction

Service providers are looking for new ways to produce incremental revenue by providing Carrier-Grade Ethernet services, business VPNs, and Internet connectivity. The challenges in bringing these services to market are minimizing upfront investment and maintaining a high level of service in the infrastructure.

Using a Layer 2 architecture at the very edge of a metro network provides a simple, low-cost solution. Building ring networks is a proven method to cost-effectively maximize the fiber layout so as to reach a large number of sites within a target geography, while still providing inherent redundancy. However, such ring topologies raise the question of the appropriate solution to use that meets scalability, availability, and efficient link utilization objectives. While traditional Spanning Tree Protocol (STP) provides a loop free environment in arbitrary topologies, it can take up to 45 seconds to converge in the event of failure within a network. Rapid Spanning Tree Protocol (RSTP, IEEE 802.1w) provides improvements, lowering convergence time in the order of a second but that is still not acceptable when offering Carrier-Grade Ethernet services.

Foundry's Layer 2 metro solution includes two proven protocols designed to offer an alternative to Spanning Tree based designs:

- Foundry's Metro Ring Protocol (MRP)¹ is specifically designed for metro rings as an alternative to Spanning Tree and offers very fast fault detection, isolation, and fail-over in just a few milliseconds. Foundry's MRP offers protection similar to the one promised by Resilient Packet Ring (RPR) but requires no special Ethernet PHYs. As a result, MRP can be used to meet the same goals at a dramatically different price point.
- Virtual Switch Redundancy Protocol (VSRP) is a "sister" technology to MRP, with the same objective in mind—an alternative to STP with sub-second fail over, but specifically designed for mesh topologies. Combining these two protocols can provide an extremely robust network design.

This document explains the operation and benefits of MRP. VSRP is explained in a separate white paper.

Benefits of MRP and VSRP

Recently, a protocol called Ethernet Ring Protection (G.8032) has been developed by ITU to address similar goals. Compared to G.8032, Foundry's MRP and VSRP protocols offer the following benefits:

- G.8032 does not support overlapping rings, i.e. rings that have shared links. In large ring networks, it is much more optimal to build rings with shared links. In contrast, Foundry's MRP-II supports overlapping rings.
- G.8032 works specifically for ring environments and does not work in mesh designs. Typical networks have a mix of mesh and ring-based topologies. In contrast, Foundry's VSRP supports mesh topologies. Additionally, VSRP and MRP have been integrated to provide efficient failover.

¹ This document uses the abbreviation MRP to refer to Foundry's Metro Ring Protocol. The reference to MRP must not be confused with Multiple Registration Protocol that is specified in IEEE 802.1ak. The two protocols are totally independent of one another and have no commonality whatsoever.

- G.8032 requires specialized hardware to process protocol messages, particularly when operating at very high speeds. In contrast, the MRP protocol can be implemented with hardware-based protocol forwarding on any class of Ethernet devices, thereby capitalizing on the economies of scale offered by mass-market Ethernet.
- Both MRP and VSRP have been proven across very large Layer 2 networks for several years across a variety of Foundry products

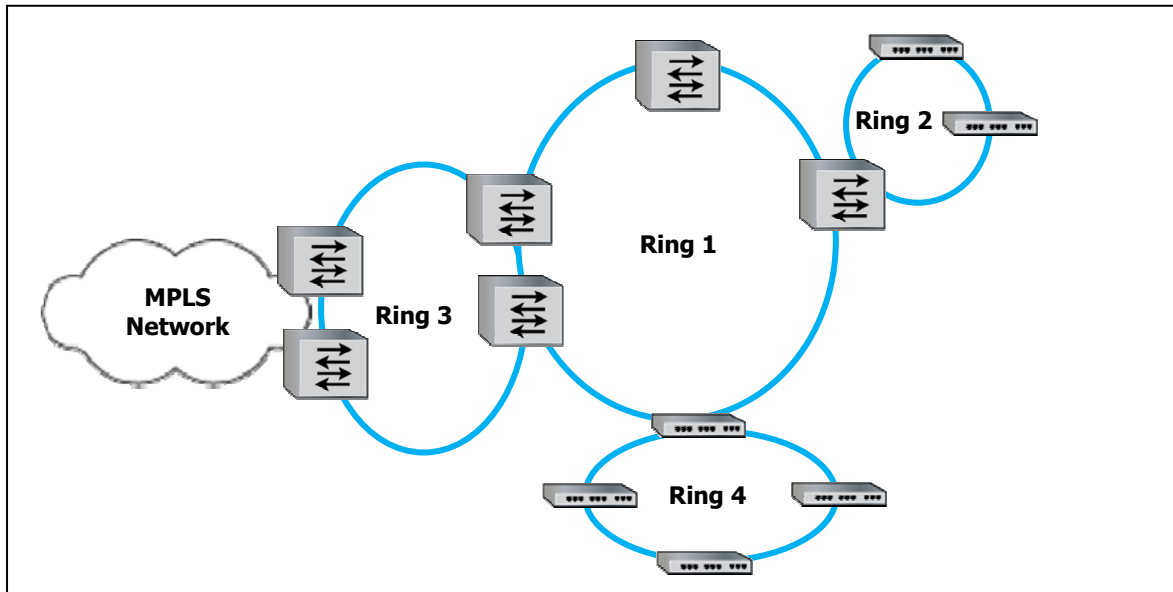


Figure 1: A typical metro network with multiple rings

Figure 1 above shows a traditional metro ring topology, which may utilize existing SONET infrastructure, Gigabit Ethernet, and 10-Gigabit Ethernet in overlapping or non-overlapping rings.

Ring topologies

There are 3 basic ring topologies that cover all metro ring designs:

- Single ring – A ring with no direct connection to other rings
- Non-overlapping rings – Rings that connect to a single switch but do not share any links among them
- Overlapping (or Shared) rings – Rings that overlap with shared links, i.e. links that impact the fate of multiple rings

These 3 types of rings are pictorially represented in Figure 2. Foundry's MRP is flexible to handle all three of these ring topologies.

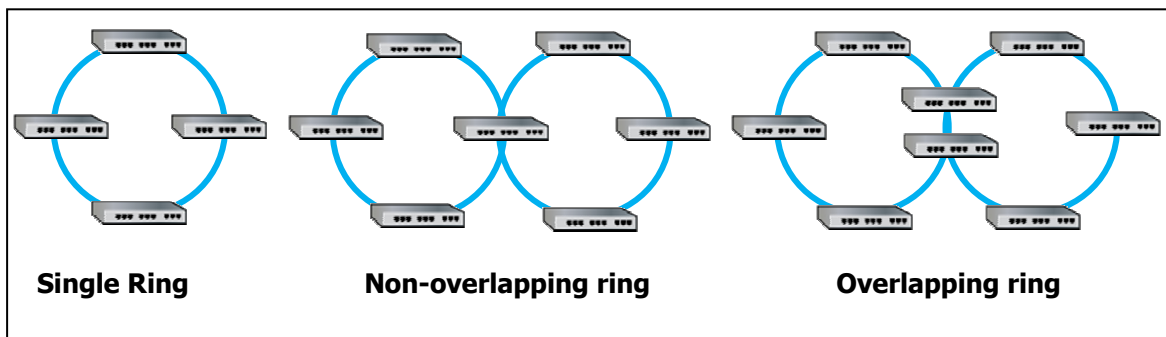


Figure 2: Types of rings

Basics of MRP Technology

The primary goal of MRP is to provide a loop-free topology. In order to accomplish this, each ring is configured with one switch as the Master of the ring, along with its ring interfaces. MRP creates a loop-free ring topology by declaring one of the Master's ring ports to be "Blocking" state. When a port is in "Blocking" state, it receives only MRP protocol packets and discards any other data packets. In the event of any link or node failure in the ring, MRP changes the Master's "Blocking" port to "Forwarding" state, thereby allowing traffic to avoid the failed link.

Once configured for MRP, the Master switch monitors the health of the ring by sending a specific protocol packet called **Ring Hello Packet (RHP)** on its forwarding interface at a specific interval. This interval is called the "hello time" and is configurable in 100 ms increments. The interface of a node on which RHP packets are forwarded is called the primary interface and the interface on which RHP packets are received is termed the secondary interface. The non-Master switches in the ring forward received RHP packets in hardware to its primary interface. If the ring is healthy, the Master will receive the RHP on its secondary interface. Only the secondary interface of the Master node is in "Blocking" state in a healthy ring.

Fault Detection and Failover

When a fault such as link failure occurs in the ring, the node adjacent to the failed link detects and sends a special type of RHP packet called "**Alarm RHP**" to report a fault in a network. On receiving the Alarm RHP packet, the MRP Master transitions its secondary interface from Blocking to Forwarding, thereby immediately healing the ring. Subsequently, it sends three Topology Change Notification (TCN) messages to ring members, alerting the nodes to flush their MAC address tables. The sequence of operations for a link failure is depicted in Figure 3.

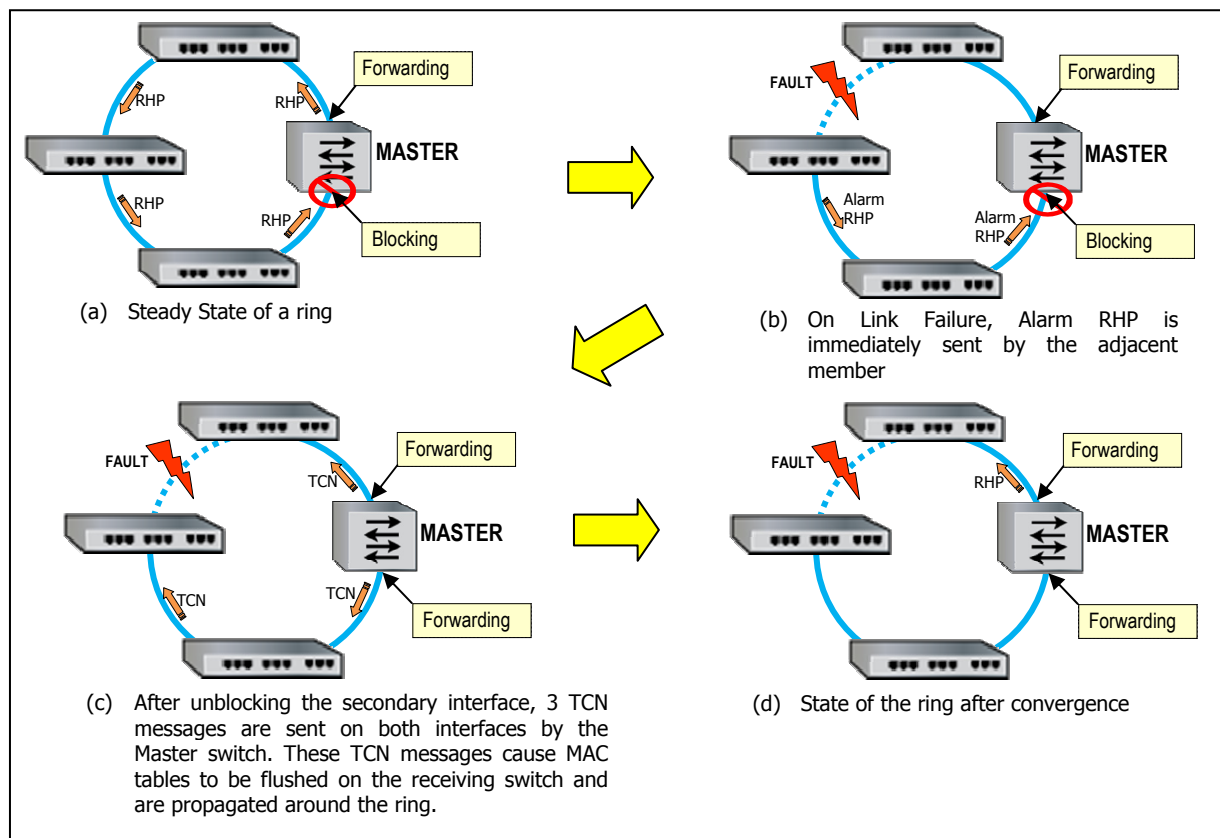


Figure 3: Sequence of operations in MRP during link failure

MRP is also resilient to detect node failures such as a failed node or even a rare event such as a data plane failure on one of its member nodes. In such a scenario, the RHP packets sent by the Master node would no longer be received on the Master's secondary interface. If the Master does not receive the RHP packet on its secondary interface, it waits for a certain period of time called the "dead timer". Upon expiry of the dead timer, the secondary interface on the Master switch transitions from "Blocked" to "Pre-forwarding" state. In "Pre-forwarding" state, the blocked interface listens only for RHP packets. The pre-forwarding delay is configurable as a multiple of the hello time, and must be at least twice the hello time. The pre-forwarding state is necessary to prevent loops and ring oscillations during failure/recovery.

On completion of the pre-forwarding delay time, the secondary interface on the Master switch transitions from "Pre-forwarding" to "Forwarding" state. As in the case of link failures, the Master Switch subsequently sends 3 TCN messages on all the ring interfaces, triggering a flush of the forwarding tables on MRP member nodes. All MRP nodes in the ring will now learn the new topology.

Support for Overlapping Rings with MRP-II

A common practice in building out large metro networks is to utilize the fiber plant efficiently by allowing overlapping links between different rings. Foundry's MRP-II implementation allows rings to overlap and use a shared interface among those rings. In such a scenario, every switch maintains state for each of the rings it is a part of, in order to keep track of the appropriate interface that needs to forward the RHPs.

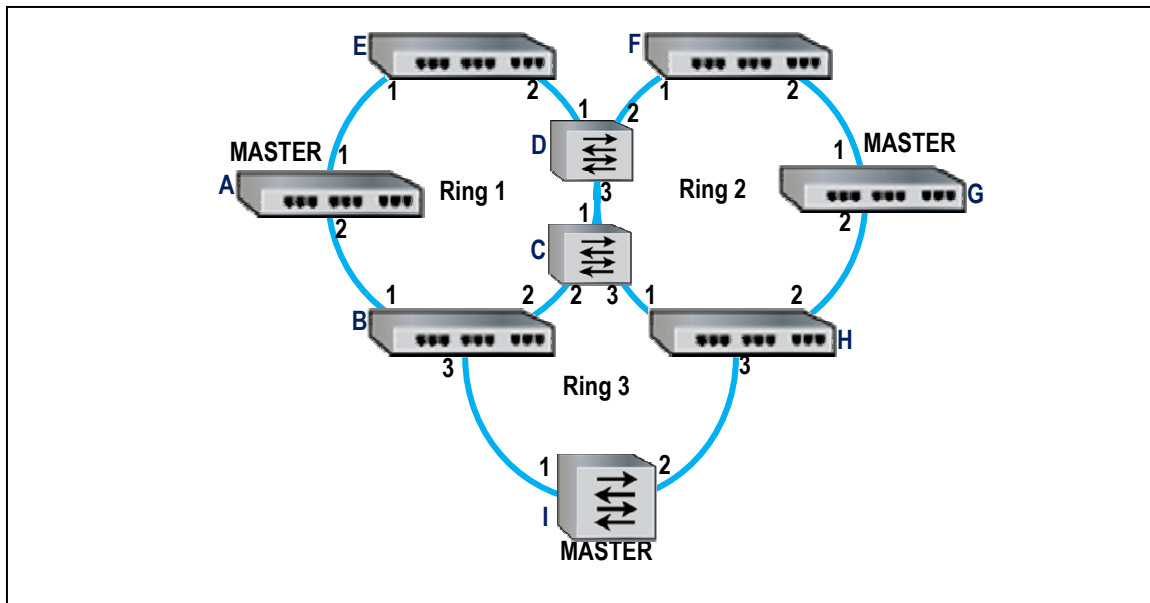


Figure 4: An example of a network with overlapping rings

In order to support overlapping rings, MRP-II uses the concepts of ring priority and shared interfaces. **Ring priority** is based on the ID of the ring itself, with higher numerical values denoting a lower priority. For example, in Figure 4, Ring 1 has the highest priority and Ring 3 has the lowest priority. A shared interface is an interface that is connected to multiple rings. For example, interface 2 on switch B is a shared interface since it is part of both ring 1 and ring 3. Similarly, interface 3 on switch C is a shared interface because it is part of both ring 2 and ring 3.

The key characteristic of MRP-II is that a node sends RHP packets *only* to all interfaces on higher priority rings in the same VLAN. This behavior ensures that RHP packets from higher priority rings never traverse a lower priority ring. Figure 5 below demonstrates the flow of RHP packets for the network shown in Figure 4.

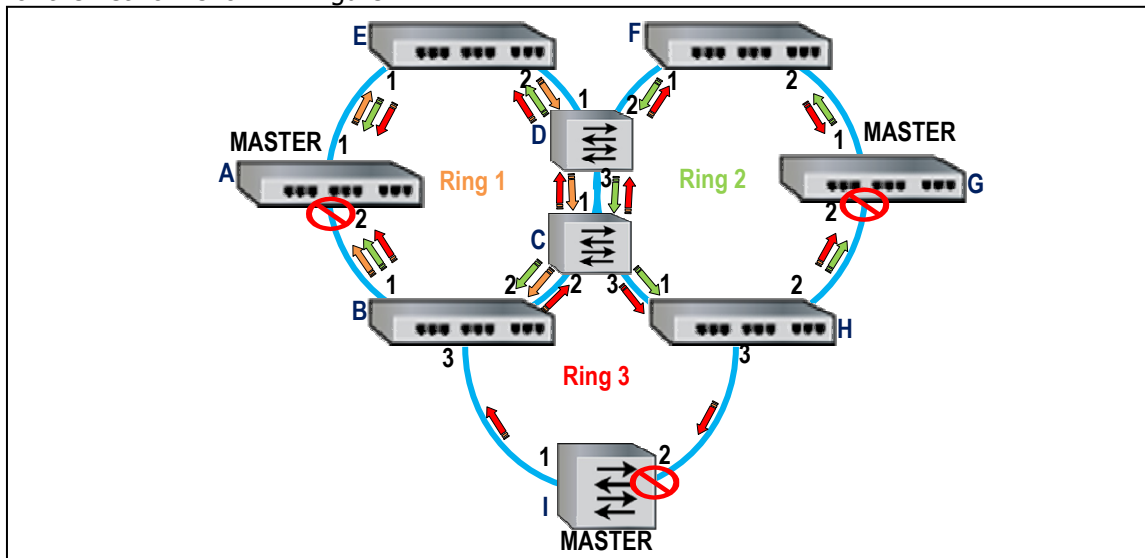


Figure 5: Forwarding of RHP packets in the 3 overlapping rings

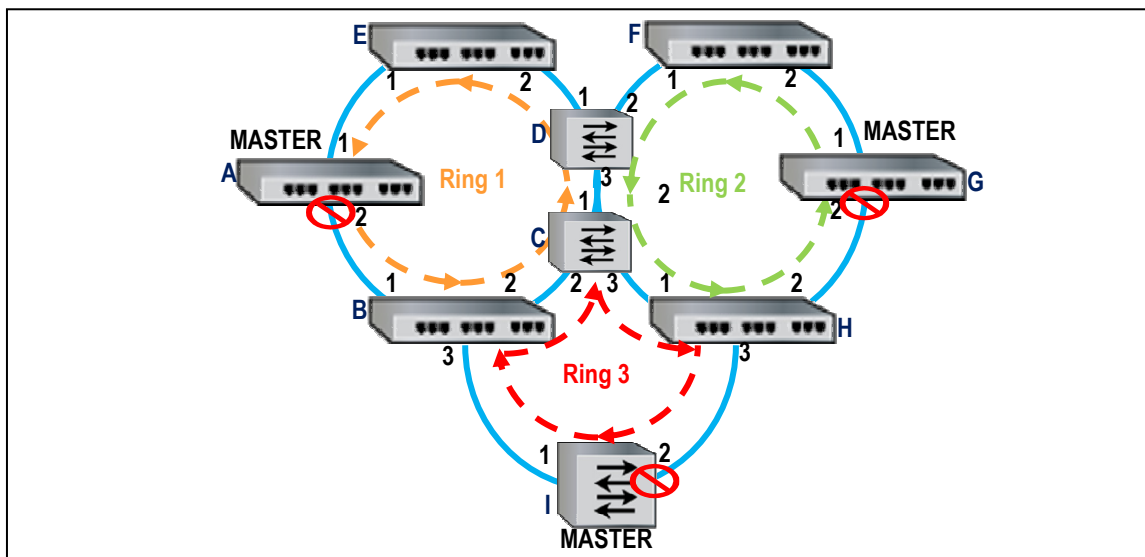


Figure 6: Flow of RHP packets to complete the ring

As can be seen from Figure 5:

- RHP packets from Ring 2 & Ring 3 flow on Ring 1 because Ring 1 is a higher priority ring.
- RHP packets from Ring 1 do not flow on Ring 2.
- RHP packets from Ring 2 flow on Ring 1 but do not flow on Ring 3.
- On the Master nodes, RHP packets are not forwarded out a blocked interface. Further, RHP packets received on a blocked interface for which that node is not the Master are also ignored. Thus, RHP packets for Ring 2 (Green) and Ring 3 (Red) received by node A on interface 1 are not forwarded to interface 2. Similarly, RHP packets for Ring 2 (Green) and Ring 3 (Red) received by node A on interface 2 are ignored.

Consequently, the actual flow of RHP packets to complete each of the rings is shown in Figure 6.

Detecting Failures in Overlapping Rings

When a failure occurs on a link connected to a regular (non-shared) interface, the same procedure is used as in the case of MRP. On the other hand, if the failure happens on a shared interface, then the ordered, cascaded RHP flow already taking place, from lower priority rings over higher priority rings, will ensure a loop free topology. This is demonstrated in Figure 7 and Figure 8. Switch A transitions interface 2 into the Preforwarding state when it gets an Alarm RHP for Ring 1. Alternatively, if switch A stops receiving RHP packets for Ring 1, interface 2 on switch A transitions from Blocking -> Preforwarding state. In the Preforwarding state, there is no data forwarding on interface 2 of switch A. However, RHP packets from Ring 2 will traverse ring 1 completely then arrive at switch G (Master for Ring 2). Thus, interface 2 on Switch G will continue to be blocked. After the Preforwarding interval, switch A transitions interface 2 into the forwarding state for actual data forwarding. Hence, throughout the re-convergence process, transient forwarding loops are avoided.

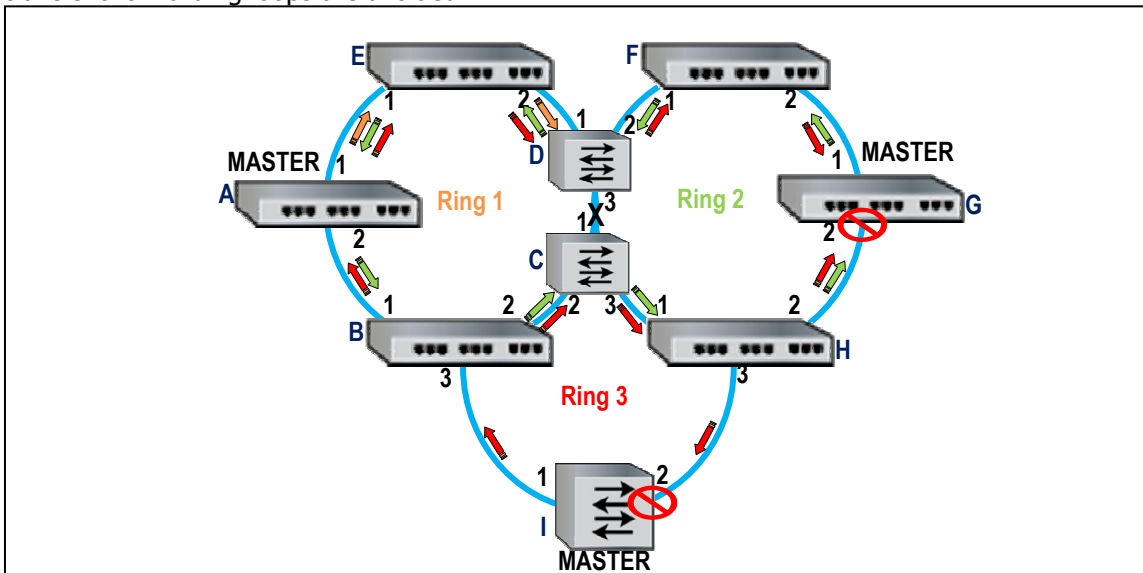


Figure 7: Forwarding of RHP packets when shared interface between Switch C and D fails

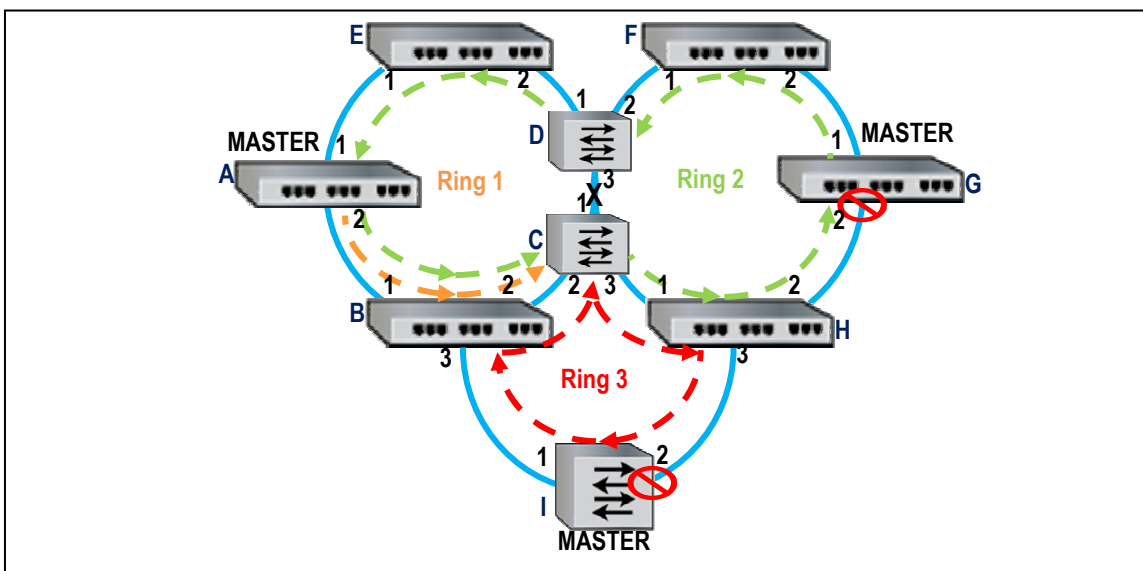


Figure 8: Flow of RHP Packets to complete the various rings when a shared interface fails

Transitions When a Ring is Healed

When a failed link recovers, the nodes adjacent to the recovered link place the interface in "Pre-forwarding" state. In this state, only RHP packets are forwarded. The periodic RHP packets sent by the Master are thus received back by the Master switch on its secondary interface. This notifies the Master switch that the ring has potentially healed. At this stage, the secondary interface on the Master switch transitions to "Blocking" state. Subsequently, 3 TCN messages are sent by the MRP Master on its forwarding interface of that ring to notify ring members of a topology change. The TCN message is used as a signal to flush the MAC tables on the member nodes for the affected VLANs. The Master node also sends a special RHP packet notifying the member nodes to transition to the Forwarding state. Upon receipt of this RHP packet (or upon expiry of the Pre-forwarding time), the nodes adjacent to the recovered link transition that interface from Pre-forwarding to Forwarding.

Efficient Ring Utilization

Foundry's routers and switches offer two mechanisms to further improve the efficiency of the ring. When the ring's Master node blocks a port, that segment of the ring is not utilized (except for control traffic) wasting valuable resources. MRP solves this problem by allowing multiple logical rings over the same physical links.

The second optimization is the use of topology groups. In networks with a large number of VLANs, the overhead to run a separate instance of a control protocol could increase load on the CPU. To prevent high CPU load, Foundry switches and routers use topology groups that share bridge topologies among different VLANs. Each topology group can have independent interface states, so one topology group can be blocking on an interface while another is forwarding. A topology group controls the state of specific switch ports in use by VLANs or group of VLANs.

Using topology groups is an efficient method of decreasing CPU load on switches by applying topology information learnt on a "control" VLAN to other members that are in the same topology group as the "control" VLAN.

The above two optimizations may also be combined to get even further benefits. Each topology group can have a unique MRP Master switch. When the Master blocks a port, it does so for only the specified control VLAN (and all VLANs controlled by that topology group). The illustration in Figure 9 below shows links blocked for certain VLANs while forwarding for others. Combining these two thus provides efficient link utilization.

Integration with an MPLS core

One of the most proven and resilient methods of deploying carrier-grade Ethernet service is over an MPLS core. Customer VLANs in an access ring may thus be carried over an MPLS core by mapping them to a VPLS instance. By using MRP in access rings that terminate on the Provider Edge router, a VPLS service can be offered with end-to-end resiliency that seamlessly integrates with the resiliency offered by a MPLS network.

Scalability of a Ring

There is theoretically no upper limit to the number of nodes that may be present in a ring running MRP. In practice, the size of a ring is limited by the **ring latency**. Ring latency is the time taken by a RHP packet to travel around the ring and come back to the MRP Master node. The ring latency comprises of the per-node forwarding latency and the sum of the latencies on the physical link between each member node in a ring. For proper MRP operation, it is recommended that ring latency be less than the hello interval. In no case must the ring latency be greater than "hello interval + dead timer interval". Rings containing 16 nodes or greater can

be achieved and have been successfully implemented in production networks. MRP is also flexible to allow ring nodes to be interconnected by a Link Aggregation Group (LAG), should additional capacity in the ring be desired.

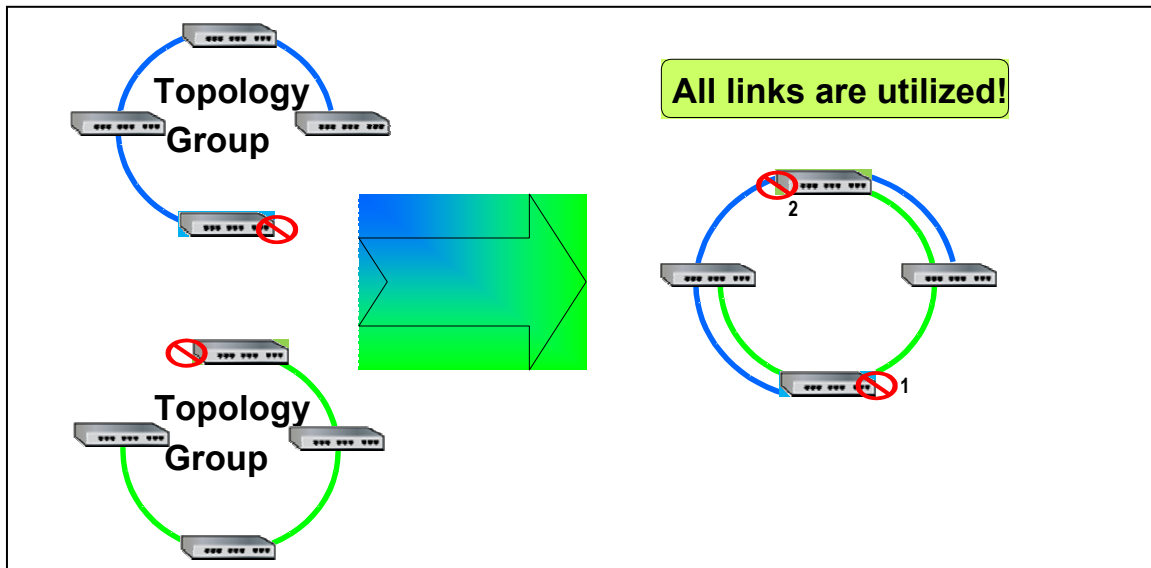


Figure 9: Increasing ring utilization using Topology Groups

MRP Support on Foundry Products

MRP is currently supported on a broad range of Foundry switches and routers, in both fixed form factor and modular chassis products. For details on MRP support among Foundry products, please refer to the appropriate product data sheet or configuration guide.

Summary

Foundry's MRP offers an alternative to spanning tree and provides super resiliency in ring networks. Its support for different types of ring topologies has made it a popular choice for providers who wish to balance advanced resiliency requirements with cost-effective loop-prevention mechanism in Layer 2 networks. Further, its integration with VSRP, VPLS, topology groups and VLAN groups maximizes utilization of the ring and allows end-to-end resiliency to be achieved when delivering a Carrier-Grade Ethernet service.

Author: Ananda Rajagopal
Document version 2.0

Foundry Networks, Inc.
4980 Great America Parkway
Santa Clara, CA 95054-1200
U.S. and Canada Toll-free: (888) TURBOLAN
Telephone: +1 408.207.1700
Email: info@foundrynet.com
Web: <http://www.foundrynet.com>

Foundry Networks, AccessIron, BigIron, EdgeIron, FastIron, IronPoint, IronView, IronWare, JetCore, NetIron, ServerIron, Terathon, TurboIron, and the "Iron" family of marks are trademarks or registered trademarks of Foundry Networks, Inc. in United States and other countries. All other trademarks are the properties of their respective owners.

Although Foundry has attempted to provide accurate information in these materials, Foundry assumes no legal responsibility for the accuracy or completeness of the information. More specific information is available on request from Foundry. Please note that Foundry's product information does not constitute or contain any guarantee, warranty or legally binding representation, unless expressly identified as such in a duly signed writing.

©2002-2008 Foundry Networks, Inc. All Rights Reserved.