

Rapid Spanning Tree in Industrial Networks

Michael Galea
RuggedCom Inc. - Industrial Strength Networks
Concord Ontario Canada

1. Introduction

This paper discusses the application of Rapid Spanning Tree to the construction of robust industrial networks. It is intended for networking professionals wishing to better understand the operation of the Rapid Spanning Tree Protocol (802.1w).

1.1 The Problem

Transparent bridges operate by forwarding traffic between their ports. The bridge examines each Ethernet frame and records (learns) its MAC address and the port upon which it resides. When a frame arrives for a given MAC address, the bridge "knows" on which outgoing port to send it.

If a frame arrives and its destination MAC address is unknown to the bridge, it will "flood" the frame out all of its ports. If bridges in the network are connected in a loop, flooding will lead to a situation in which a frame endlessly circulates. The looping traffic can easily consume all the bandwidth of links used.

It may seem that loops are evidence of configuration problems, but allowing loops in a topology is extremely useful. Without loops the topology has no redundancy. If a link fails then connectivity is lost. For this reason loops should be viewed not as misconfigurations but as good design strategy.

STP and Rapid Spanning Tree prevent accidental loops and allow redundant connections by detecting the loops and "opening" them.

2. Brief history of STP and RSTP

Spanning tree was designed to solve the fundamental problem of traffic loops created by the interconnection of LANs with redundant transparent bridges.

The key idea in STP is to prune (looping) links in order to reduce the network topology to that of a tree. The resulting tree "spans" (i.e. connects) all bridges, but eliminates loops. The steps in order to best accomplish this process are:

1. Allow all bridges to send messages to each other that convey their identity and link "cost".
2. Elect a single bridge, among all the bridges in the network to be a "root", or central bridge.
3. Let all other bridges calculate the direction and cost of the shortest path back to the root using messages received from bridges closer to the root. Each bridge must have only one "best" way to forward frames to the root.
4. If two bridges servicing the same LAN exchange messages with each other, the one with the lowest cost to the root will service the LAN. The other bridge

will discard all frames received from that LAN, thus opening the link and blocking a traffic loop.

The STP protocol has proved to be the tried and tested method for providing path redundancy while eliminating loops in bridged networks. The STP protocol does suffer from a number of drawbacks that limit its applicability, namely:

- STP has lengthy failover and recovery times. When a link fails in STP, a backup link to the root requires at least 30 seconds to recognize that it is the best (or only) path to the root and become usable.
- When a failed link returns to service, information about the "better" route will instantly cause a backup link to start blocking. But the portion of the network below the link that is returning to service will be isolated (for about 30 seconds) until that link becomes forwarding.
- Another problem with STP is that it requires that all links must pass through a lengthy period of address learning, even if the link is a point-to-point link to a device such as an IED or RTU.

2.1 Enter RSTP

RSTP solves STP's problem with failover time by a number of means. Whereas STP bridges store only the best path to the root bridge, RSTP bridges store all potential paths. When links fail, RSTP has pre-calculated routes to fall back upon. Additionally, unlike STP bridges, an RSTP bridge will respond to another bridge that advertises an inferior or incorrect route to the root bridge. This information allows the bridge with incorrect information to be rapidly trained.

RSTP solves STP's problem with lengthy recovery time by introducing a new procedure called proposing-agreeing. Proposing and agreeing works after a better path to the root is restored by "shuffling" the restored part of the network one hop at a time towards the network edge. This method also enables the network to come up quickly at inception.

RSTP also introduces a method for quickly bringing up ports at the edge of the network, while still protecting them against loops. If the port is designated as an "edge" type of port, RSTP will continue to send configuration messages out the port (in order to detect loops) but will allow traffic to flow as soon as the port rises. In the event of a loop, some looped traffic may flow before RSTP quickly seals the network. PC's, IEDs and RTUs connected via edge ports can send traffic without the extensive delays imposed by RSTP.

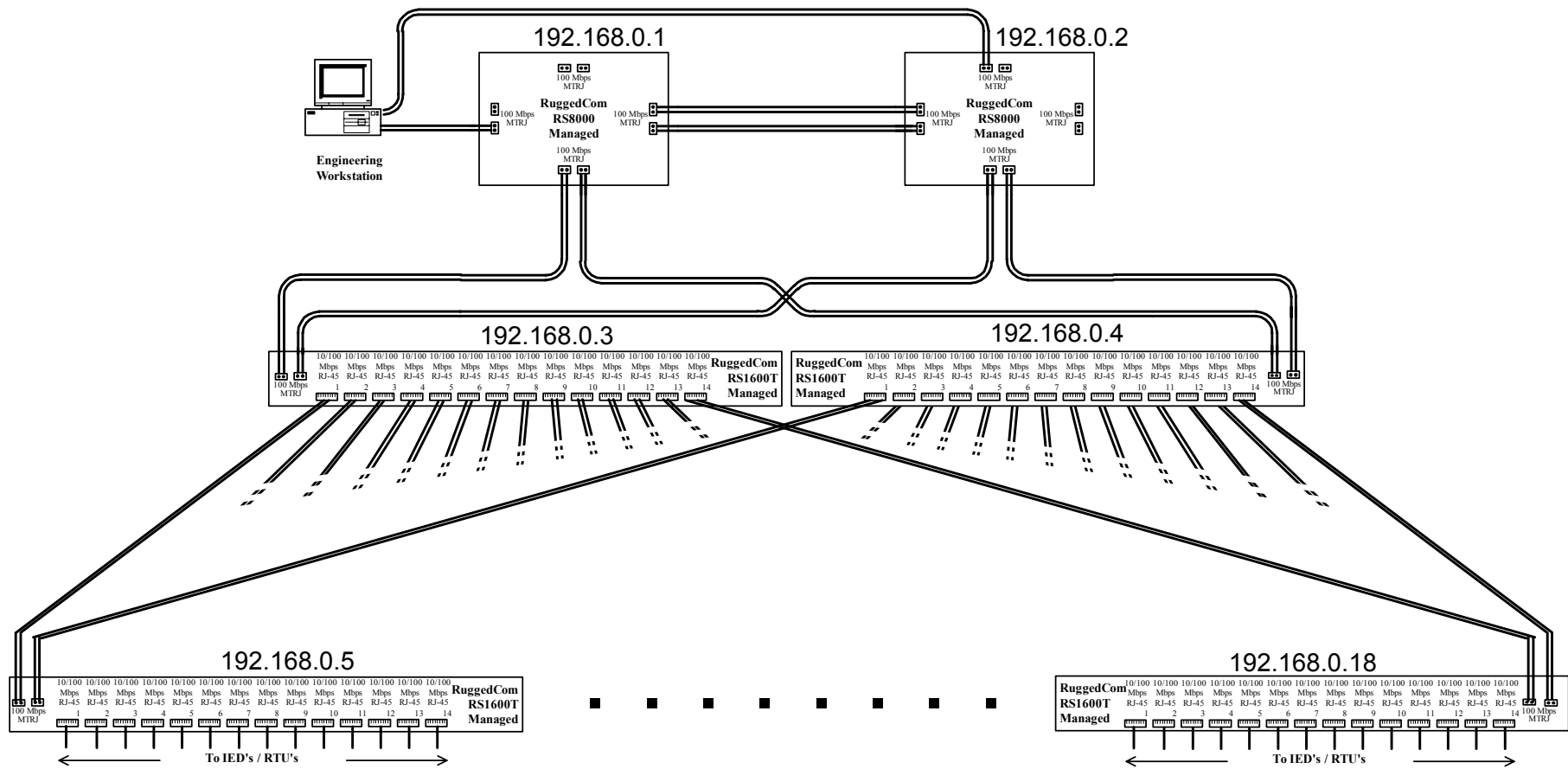


Figure 1 - A Well Connected Tree

3. Typical Architectures In Industrial Networks

3.1 Trees and Meshes

STP was originally designed to protect well connected tree topologies. Often, whole portions of the network are employed in a sparing mode. In Figure 1 switch 192.168.0.2 will carry traffic only if 192.168.0.1 fails. Similarly, switch 192.168.0.4 will carry traffic only if 192.168.0.3 fails.

If desired, the planner can balance the traffic in the network by manually adjusting link costs in the switches. As an example, by increasing the link cost to the 192.168.0.3 switch at switch 192.168.0.18, all traffic will flow to 192.168.0.4.

Tree networks offer the fastest failover and recovery times. In Figure 1 it can be seen that only two trunks are involved in the forwarding decision at any point in the network.

Latency in tree networks tends to be the lowest of all networks since there are fewer hops back to the root.

The size of tree type architectures can be arbitrarily large. RSTP recommends that the “bridge diameter” (i.e. the maximum number of hops from end to end of the network) be limited to seven. While a seven bridge diameter may seem like a small network, it can be quite large for a 16-port switch and range to hundreds of ports. Furthermore the RSTP recommendation is based upon a pessimistic assumption of the transit delay of each switch being up to a second. In practice the delay is much smaller and the maximum size of the network is correspondingly greater.

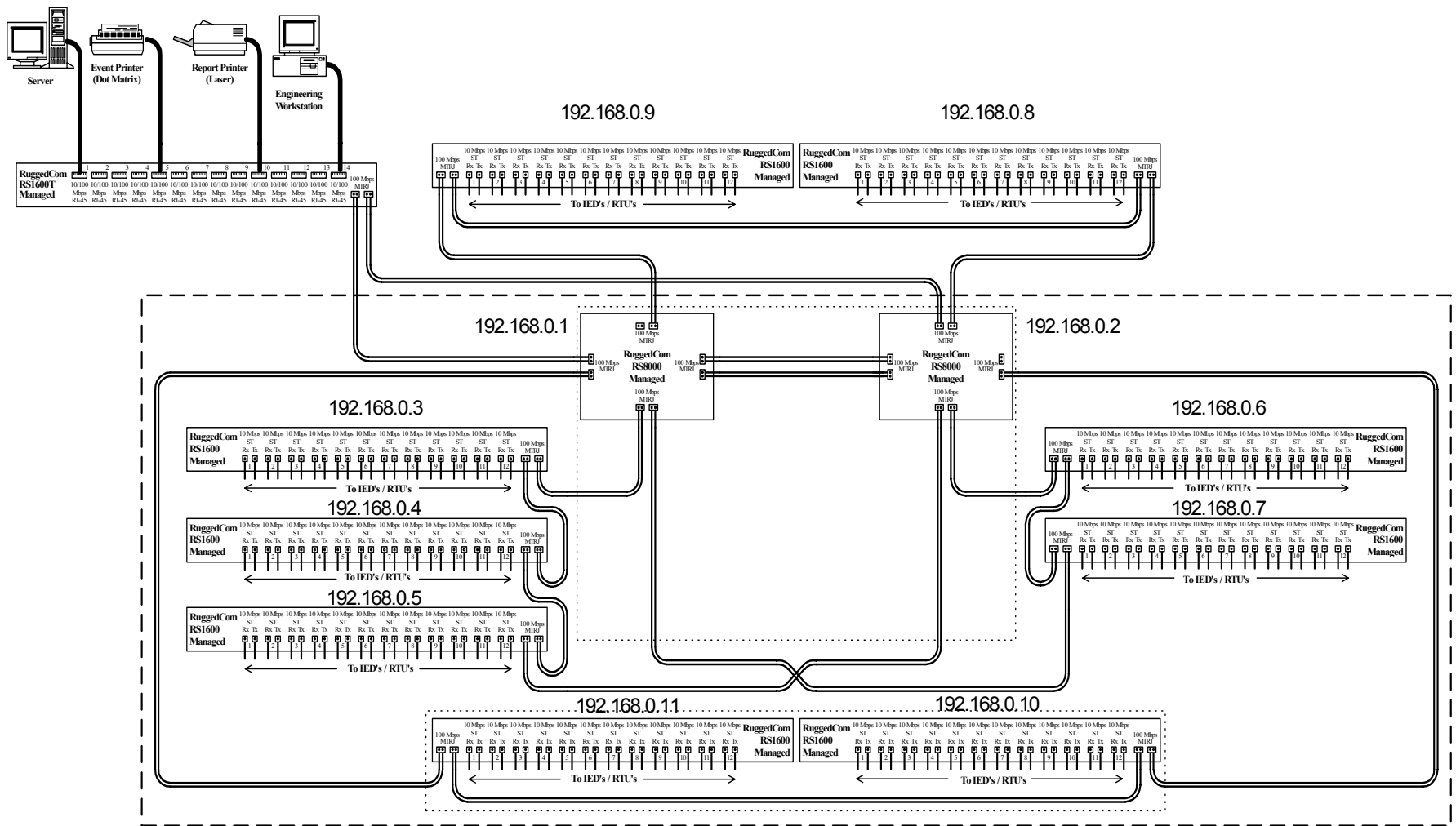


Figure 2 - Typical Industrial Network Using A Ring

3.2 Rings

More and more often, industrial networks are being implemented using rings. A ring topology offers built-in redundancy and is often the most economical in terms of interconnection costs. Two popular methods of implementing rings are collapsed backbone and distributed switch.

The distributed switch method, or simple ring, is employed when traffic sources are geographically distributed. The traffic sources at each location are aggregated onto switches, which are organized into a ring. The connections between switches in the ring may be made using dual redundant links to obviate the possibility of failure at a fiber, connector or port level.

The collapsed backbone method (See Figure 2) is usually employed when a large number of traffic sources are located in close proximity to one another. The traffic sources are aggregated onto switches, the switches organized into a number of rings and all rings terminated in a common root node.

Quite often the network topology is a mixture of both methods, such as a ring of rings.

Traffic in a ring tends to be balanced. The ring will open itself with an equal number of switches on either side of the root bridge (given an odd number of switches in the ring).

Latency in ring networks tends to be greater than in tree networks as there are usually more hops to pass through in order to go anywhere useful. The worst case occurs when switches on either end of the blocked link at the “bottom” of the ring need to forward to each other. In this case traffic must flow through every switch in the ring.

Ring networks offer only slightly slower failover and recovery times than tree networks. The worst case link failure in ring networks occurs on a port at the root. In this failure case half of the switches in the ring must retrain themselves to face their root port in a completely opposite direction after a link failure or recovery. The other half of the network must reverse the direction of transmission to switches in the failing half.

The size of the ring is in theory limited by the RSTP bridge diameter, which assumes a pessimistic transit delay of one second per switch. In practice the maximum number of switches in an optimized ring occurs when the number of priority bridge levels has been exhausted. This limits the size of the ring to 31 switches. Rings of more than 31 switches are still possible but will failover and recover in a slower fashion. The RuggedSwitch™ User Guide fully details the steps in planning a ring and the issues involved in implementing large rings.

4. Detailed Examples Of The Failover And Recovery Process

4.1 Dual Link arrangement

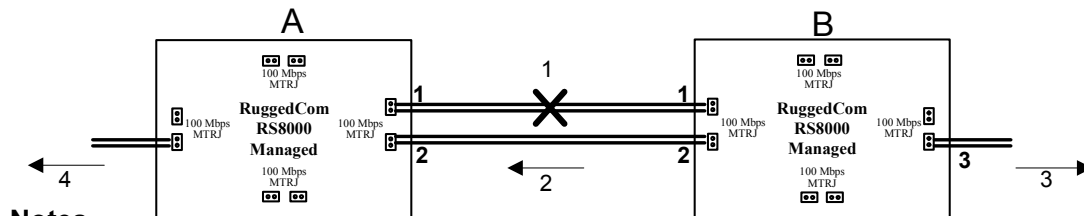
Figure 2 presents two switches protected by a dual link arrangement, and the series of events that occur after a link failure.

Both bridge's detects failure of link 1 simultaneously and immediately age out the learned MAC address entries for these ports.

Bridge B has been receiving periodic transmissions of BPDUs on link 2. This information allows it to evaluate link 2 as its best path to the to the root bridge. Bridge B immediately sets its root port to 2.

RSTP procedure requires a topology change when adding a path to the topology. Bridge B "sees" the new root port as an added path and floods topology changes out its ports. Though not strictly necessary in this case, they cause no ill effects.

Including the time to recognize the link failure (an process that takes less than a millisecond) the switches failover to link 2 in less than 5 milliseconds.



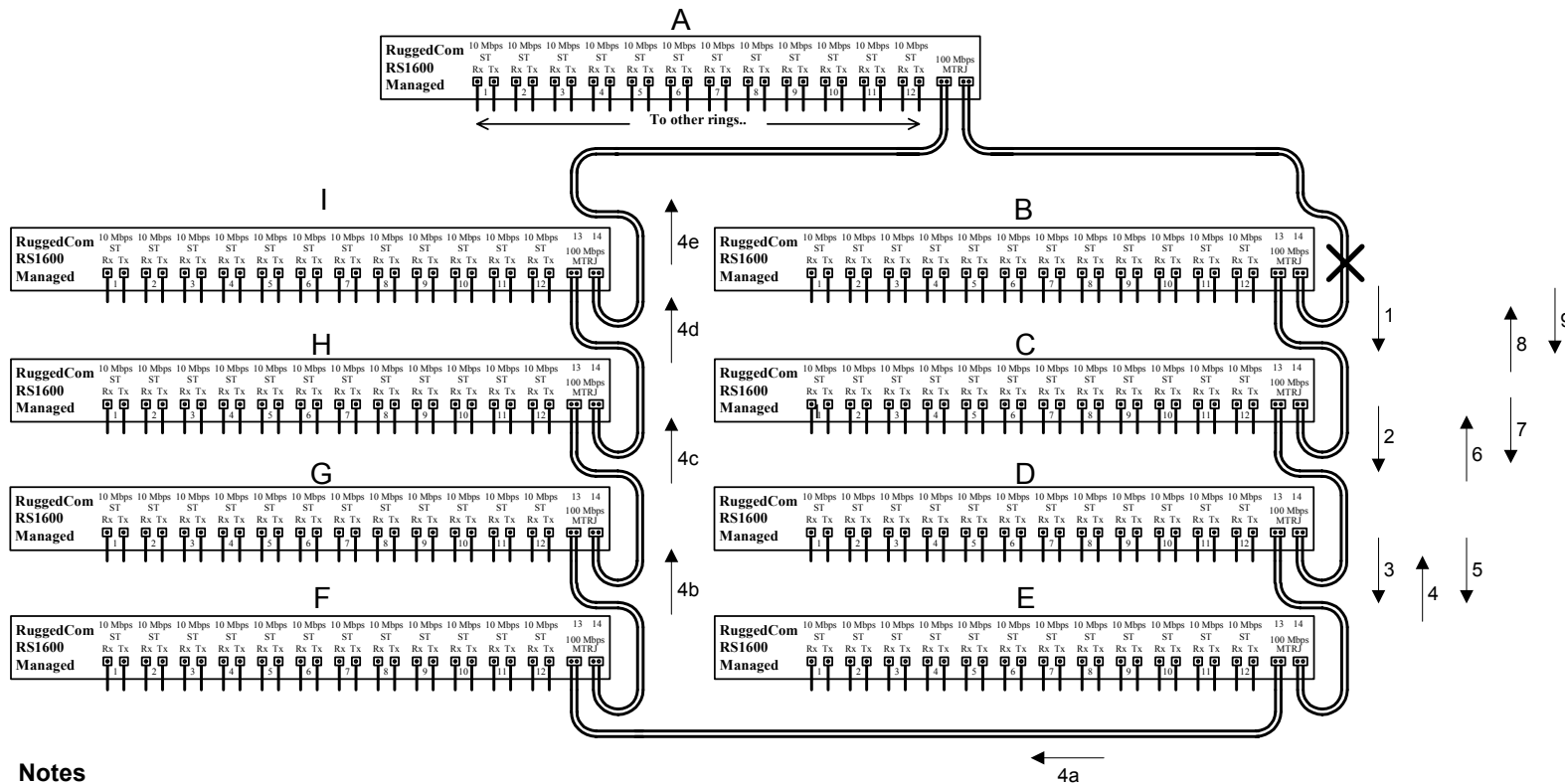
Notes

Before the link between A and B fails, bridge A is the root and link 2 is blocking at B. Both of A's ports are designated. Port 1 of switch B is root and port 2 is an alternate.

- 1) The link fails.
- 2) Bridge B transmits a BPDU to A on link 2 signifying it has changed it's port role from Alternate to Root. Bridge B places the link in the forwarding state. At this point connectivity is restored. The BPDU also has the topology change bit set.
- 3) Bridge B issues BPDUs with the topology change bit set to any of its designated ports.
- 4) Bridge A issues BPDUs with the topology change bit set to any of its designated ports.

Figure 3 - Failover In A Dual Link Arrangement

The recovery process for this example is quite straightforward. When link 1 is restored, bridge's A and B will transmit BPDUs on it. Bridge A will ignore the BPDU from bridge B. Bridge B will use the bridge A BPDU to place its link 2 in blocking and then change its root port towards A. Afterwards, bridge B will signal a topology change to bridge A. At this point the network will be healed. The recovery process introduces an outage of less than 5 milliseconds.



Notes

Before the link between A and B fails, bridge A is the root and the link between E and F is blocking. All ports that face in the direction of bridge A are root ports and all ports that face away from A are designated ports.

- 1) Bridge B detects a failure of the link to bridge A. It propagates information about a new root (itself) to C.
- 2) Bridge C propagates information about the new root to D.
- 3) Bridge D propagates information about the new root to E.
- 4) Bridge E knows that root bridge A is still alive. It discards D's BPDU and sends it a BPDU with a port role of designated and the proposal flag set.
- 4a) Bridge E takes the link to F out of blocking and signals a topology change to F. Bridge F, G, H and I propagate the change upward.

Any frames sent by these bridges are now flooded in both directions around the ring.

- 5) Bridge D blocks the link to C and moves its root port to E by sending it a BPDU with a port role of root and the agreement flag set.
 - 6) Bridge D sends a BPDU to C with a port role of designated and the proposal flag set.
 - 7) Bridge C blocks the link to B and moves its root port to D by sending it a BPDU with a port role of root and the agreement flag set.
 - 8) Bridge C sends a BPDU to B with a port role of designated and the proposal flag set.
 - 9) Bridge B converts its link to C from a designated port to root port by sending C a BPDU with the agreement flag.
- The ring is now healed.

Figure 4 - Failover In A Ring Topology

4.2 Failover In Rings

Figure 3 presents a network of nine bridges organized in a ring topology. The figure details the sequence of steps to heal the ring after the link between bridges A and B fails.

Initially, bridge B has information only about root bridge A. All information about the root bridge flows towards the break between bridge E and F. After link AB fails bridge B recognizes the failure and must conclude that it is the root bridge, propagating the information towards C.

The information will continue to propagate around the ring until it reaches the portion of the network that is still aware a path to bridge A exists (i.e. bridge E).

Bridge E propagates correct information towards bridges D, C and B. Since these bridges are changing the identity of their root ports, they must use the proposal-agreement process to achieve rapid forwarding.

Typically, each step in the process involves a protocol “think time” and a frame transmission time, the sum of which is less than about 3 milliseconds. This leads to a total failover time for the ring of about 27 milliseconds. There is also the time required to signal topology change to bridges F-A. In this example the topology change time is interleaved with the failover process and does not contribute to the failover time.

The recovery process for this example is quite straightforward. When link AB is restored, bridge A will transmit a BPDU down it. Bridge B will change its root port towards A, and then signal a topology change. Bridge B will propagate the new root information towards bridge C. Bridge C will change its root port and will train bridge D. Bridge D will train bridge E. Bridge E will attempt to train bridge F but bridge F will see a lower path cost from bridge G and will discard the BPDU from E. At this point the network will be healed.

When bridge A receives the topology change from B it propagates the topology change towards bridges I-F. During the recovery process bridge A will continue to forward a number of frames for bridges B-E in the direction of bridge I. At some point these frames will encounter a newly blocked link on bridge C-E.

Fortunately, bridge A will use the topology change to start flooding frames, as will bridges I through F. Bridge A will lose about 2 milliseconds worth of frames, bridge I 4 milliseconds, bridge H 6 milliseconds, bridge G 8 milliseconds and bridge F 10 milliseconds worth of frames.

5. Conclusions

RSTP is well equipped to deal with deal with link failures and to provide rapid startup in industrial networks. RSTP may be employed effectively in tree type or ring type architectures.

Dual link arrangements (where one link serves as a hot standby for another) provide rapid failure recovery, typically in less than 5 milliseconds.

RSTP rings also provide rapid recovery. Practical rings should be limited to 31 switches. A useful rule of thumb is to budget 3 milliseconds of recovery time for every switch in the ring.

These performance levels are available on RuggedCom products due to optimizations of the RSTP protocol that improve performance while maintaining interoperability with other vendors.

6. References:

1. The Switch Book, Rich Seifert, Wiley
2. ANSI/IEEE Std 802.1D, 1998 Edition
3. ANSI/IEEE Std 802.1W, 2001 Edition