FAILURE HANDLING AND EFFICIENT DATA TRANSPORT IN WIRELESS ETHERNET NETWORKS USING ADAPTIVE MODULATION

Torsten Mueller, Herbert Leuwer, Thorsten Kaiser, Thomas Alberty

Ericsson GmbH, Backnang, Germany

Abstract: Networks with microwave radio links shall ensure high service availability. The paper shows that in Ethernet-based networks with redundant links, rapid spanning tree protocol (RSTP) and multiple spanning tree protocol (MSTP) provide fast protection and have significant advantages over spanning tree protocol (STP). Traffic due to flooding of frames with unknown destination address may lead to congestion on low capacity link and must be considered for network planning depending on the network topology. Virtual LANs (VLANs) can be used to isolate flooding traffic. Adaptive modulation combines high service availability with high data rates on single links. Rerouting best effort traffic after modulation change using RSTP or MSTP achieves optimized network usage. Interruptions for best effort traffic can be kept short.

Keywords: Microwave radio, Adaptive modulation, Ethernet, Layer 2, Protection, Restoration, Convergence time, Spanning Tree, Rapid Spanning Tree, Multiple Spanning Tree

1. INTRODUCTION

Microwave radio is widely used in mobile radio backhaul networks. The capacity required for microwave radio links is expected to increase, e.g. due to the deployment of high speed packet access (HSPA), the long term evolution (LTE) strategy in mobile networks, and the usage in new application areas like fiber extension networks. Spectrum efficient modulation schemes are available to increase the data rates on microwave radio links. Even if the majority of the data traffic is best effort (BE), it is necessary to guarantee high availability for guaranteed traffic, e.g. real-time voice. For a single radio link this can be achieved using adaptive modulation. To increase the service availability in case of network element failures, radio networks may provide network redundancy using ring topologies. In order to rapidly detect failures and react on them, a protection protocol is needed.

At the same time, Ethernet technology is becoming popular in wireline as well as in wireless transport networks. Ethernet originally has been designed as a LAN technology with hubs and bridges connecting LAN segments. The Ethernet data plane itself has no protection against loops which are extremely critical in layer 2 networks. Therefore, bridges had to implement the spanning tree protocol (STP) which dynamically detects loops in the physical topology and opens them. This feature can be used for network protection purposes as well, but its recovery time is in the order of tens of seconds to minutes which is far too long for carrier grade applications.

Consequently, the rapid spanning tree protocol (RSTP) has been designed which provides significantly reduced recovery times. While RSTP recovery time is in the order of seconds in networks with half-duplex links, the majority of today's carrier is limited to full duplex links. In those cases, RSTP can achieve lower recovery times.

This paper investigates the recovery time of RSTP by use of discrete event simulation for various network topologies and discusses the significant factors. Besides the recovery time caused by the RSTP protocol itself, flooding traffic after errors has a significant effect as it reduces the network capacity usable for data traffic. Therefore, flooding has been considered in detail.

Microwave links work at frequencies which are subject to rain fading. While the usage of adaptive modulation on isolated radio links is well understood, this paper also investigates the effects on layer 2 networks. Furthermore, the paper introduces a mechanism which uses rerouting based on RSTP and multiple spanning tree protocol (MSTP) to achieve higher network efficiency.

The rest of the paper is organized as follows: Section 2 provides some background on the various flavors of the spanning tree protocol and its function, whereas section 3 investigates the recovery time by means of simulation. Section 4 explains the impact of broadcast traffic and flooding on network performance. Section 5 discusses the usage of adaptive modulation and introduces a mechanism to increase the network efficiency in networks which employ microwave radio links with adaptive modulation.

2. BACKGROUND ON THE SPANNING TREE PROTOCOLS

The algorithm for STP has been originally defined in [Perl85]. While the physical network topology might be arbitrarily meshed, STP forms a logical topology which is a spanning tree in the mathematical sense, thus eliminating loops. Furthermore, STP provides dynamic re-configuration in case of resource failures.

STP has been defined and published as part of IEEE standard 802.1D (MAC Bridges). Various revisions exist, which can be distinguished by the year of their publication. In 2001, RSTP was approved in 802.1w which defines a new, much enhanced version of the basic spanning tree protocol. RSTP has been incorporated into [IEEE 802.1D-2004]. In parallel with the definition of RSTP, Multiple Spanning Tree Protocol (MSTP) was published as supplement 802.1s in 2002 and finally incorporated into [IEEE 802.1Q-2005]. It shares the enhanced protocol mechanisms of RSTP and introduces support for multiple loop-free topologies in parallel. User data traffic is mapped to particular topologies through its VLAN association. Multiple VLANs may be mapped to a single topology.

All flavors of STP share the basic spanning tree algorithm [Perl85] which uses the concept of path costs. By convention, a LAN segment of higher bandwidth has lower costs than one of lower bandwidth. Priority vectors are formed from a combination of identifiers and accumulated path costs. By comparison of different vectors, the better vector is chosen and finally the loop-free topology is formed as follows: First, one bridge is selected as the root bridge in the LAN, based on its identifier. All other bridges select the port with the lowest accumulated path cost towards the root bridge as their root ports. Those ports connecting remote areas of the LAN towards the root bridge become designated ports. Each bridge propagates its view of the LAN and its configuration through bridge protocol data units (BPDUs). A bridge which receives a BPDU compares the priority vector contained in it with its own current information and adapts to it if the received information is better.

Each bridge port can have one of the following states:

- Discarding – performs neither MAC learning nor forwarding,
- Learning – performs MAC learning but no forwarding,
- Forwarding – performs MAC learning and forwarding.

The original STP defines additional port states which have later been consolidated to the discarding state. The set of ports in the forwarding state and their attached LAN segments form the spanning tree, i.e. the "active topology" in the LAN.

Despite the commonalities, there are also great differences between RSTP/MSTP and STP:
The original STP follows a centralized approach. BPDUs are clocked through the root bridge. The other (designated) bridges in the LAN only transmit BPDUs on reception of BPDUs at their root ports. Failures of the root bridge or its ports are only recognized by a bridge due to missing BPDUs after a rather long timeout (default 20s). In the worst case, it takes 20s for a LAN to start re-configuration due to failure. Port state transitions are always timer controlled. After a re-configuration in the LAN, bridges need to delete the learned MAC addresses stored in the filtering databases (FDBs). The need for such reconfigurations is first propagated to the root Bridge which then pulses the command to flush the FDBs through the whole network. All FDBs will then be subject to accelerated ageing through shortened ageing timers but there is no immediate MAC address deletion. By default, MAC addresses will be deleted after 15s, but rapid ageing is applied for 35s in total. So only after 35s MAC address ageing with the standard ageing time is resumed.

RSTP significantly reduces the topology convergence times under certain conditions. For point to point links RSTP applies an event driven hand-shake mechanism between adjacent bridges which instantly aligns the configuration and allows rapid port state transitions. Bridge ports which do not connect to other bridges can be configured as edge ports. Edge ports are always in state forwarding. Additionally, MAC addresses learnt at edge ports are never deleted due to network reconfiguration. Each port individually transmits BPDUs controlled by per port timers. This allows for use of BPDUs as heartbeats between bridges. Bridge or link failures can thus be detected after at most 6s. However, if failures are detected by other means, network re-configuration may start immediately. Changes in the active topology are propagated away from the point in the LAN where they occurred. The root bridge has no special role in this process. MAC addresses are removed from the FDBs through immediate deletion (flushing) rather than rapid ageing. Not all MAC addresses but only those learnt at the required subset of bridge ports in the LAN are deleted.

3. SPANNING TREE RECOVERY TIME

As explained above, the STP ensures a loop-free logical topology by blocking ports of the physical topology. If a link or a node of the physical topology fails, the logical topology formed by the spanning tree changes in a way which ensures connectivity between all nodes if possible at all. This feature can be used for protection purposes. Between the failure event and the convergence of the spanning tree and the relearning of station MAC addresses, the communication of layer 2 clients might be disturbed.
An important parameter of any protection scheme is its recovery time which has to be as short as possible. For investigation we apply discrete event simulation

using the simulator YATS [YATS] extended by the Lua [LUA] programming language to describe simulation experiments.

We have performed studies using various topologies including ring, combination of ring and tree, as well as multiple rings, see Figure 1, Figure 2, and Figure 3. All bridges are interconnected using point to point links. The traffic is restricted to unicast.

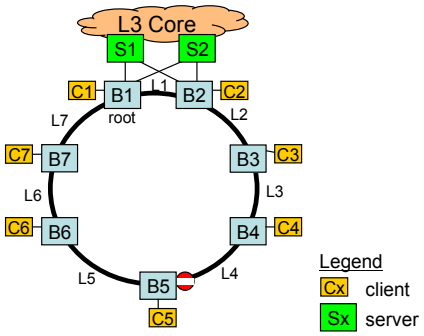

Figure 1: Ring topology (Ring-7)

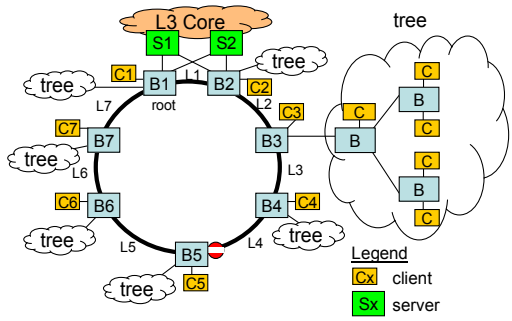

Figure 2: Ring and tree topology

In this paper, we use the following definitions:

- detection time, t_{detect}: the time difference between the failure event and the informing of the spanning tree process. If not stated otherwise, we assume that bridges can rapidly detect link or node failures and do not have to rely on RSTP heartbeats. Therefore, we assume the detection time to be zero.
- topology convergence time: the time difference between the failure event and the time when the last port in the network has transitioned to its final state. Please note that after the topology convergence time, FDBs may contain wrong entries so the communication for existing connections is still interrupted.
- total recovery time: the time difference between the failure event and the complete establishment of connectivity for new and existing connections.
- worst case total restoration time t_{worst}: Whether the communication after a failure is disturbed and how long depends on the exact position of the failure, the position of the communicating layer 2 stations, and whether the station MAC addresses had been already learned in the intermediate bridges. In order to simplify the results, we use t_{worst} which is the time period between the failure event and the instant when the last station is able to communicate.

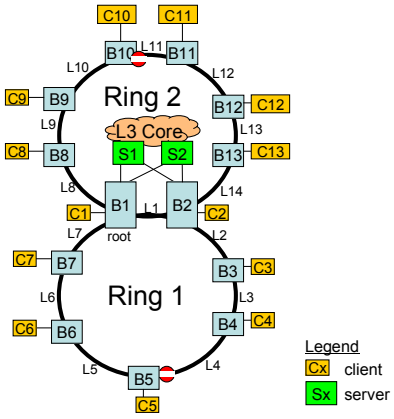

Figure 3: Multiple ring topology (2*Ring-7)

For STP, the topology convergence time and the total recovery time are

$$\text{ConvergenceTime} = \text{MaxAge} + 2 \cdot \text{ForwardDelay}$$

$$\text{TotalRecoveryTime} = \text{MaxAge} + 3 \cdot \text{ForwardDelay}$$

where MaxAge and ForwardDelay are STP parameters. Taking their value range into account, the the following minimum, default, and maximum values hold:

	Topology convergence time	Total recovery time
STP min.	6s+2*4s = 14s	18s
STP default	20s+2*15s=50s	65s
STP max.	40s+2*30s=100s	130s

As explained above, RSTP is much more efficient on point to point links than STP. Therefore, small recovery times can be achieved. Figure 4 compares the topology convergence time and recovery time for the ring network for all possible link failures. The BPDU processing time is assumed to be 5ms.

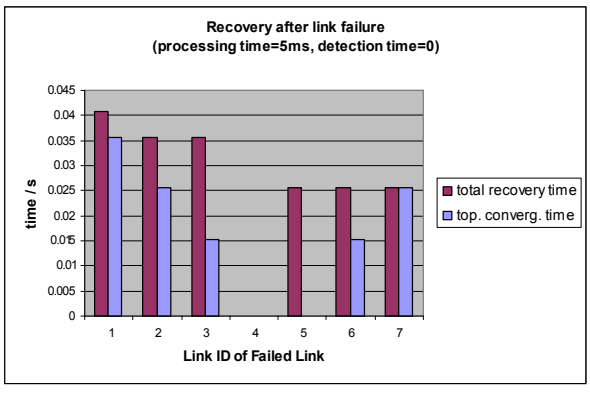

Figure 4: Comparison of topology convergence time and total recovery time in ring-7 network

As can be seen, the topology convergence time depends strongly on the distance of the failure from the blocked port which is on bridge B5 in the ring network of Figure 1. The total recovery time is longer and less dependent on the position of the failure be-

cause the length of the total recovery time is determined by the time when the last port in the network is flushed. Flushing of MAC addresses is initiated by BPDUs with topology change (TC) bit set. Those BPDUs need to pass through the complete ring. Please note that the difference of one processing time between link 1 and 2 is due to a different sequence of BPDU processing in one node. The difference between links 1,3,5 and 5,6,7 is due to an additional handshake of BPDUs. When link 4 fails, no change of topology is necessary and no BPDUs are generated.

The worst case total recovery time t_{worst} in the ring network can be calculated using:

$$t_{worst} = t_{det\ ect} + 2 \cdot N \cdot \left(t_{proc} + t_{queue} + t_{trans} + t_{prop}\right) + t_{flush}$$

where N is the number of nodes in the network, t_{proc} the processing time for a BPDU, t_{queue} queuing time of BPDUs, t_{trans}: transmission time of BPDUs, t_{prop} is propagation time between 2 adjacent bridges and t_{flush} is the time it takes to completely flush a FDB.

The times which depend on the number of nodes in the network are the most critical ones. While queuing, transmission, and propagation time are small or negligible in systems with data rates >10 Mbit/s, processing time is not. The linear effect of the processing time is shown in Figure 5.

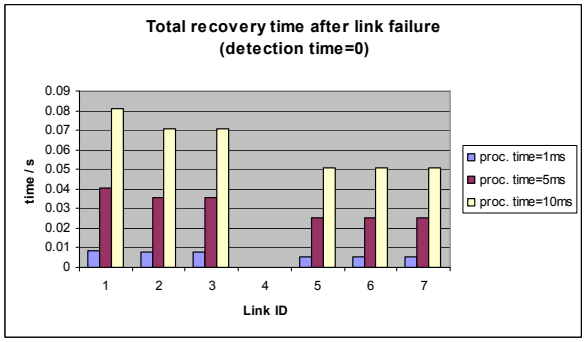

Figure 5 Total recovery time in the ring network for various processing times

Besides the traffic interruption after a link failure, there is also an interruption when the link is repaired. The interruption time is compared in Figure 6 and lies in the same range. Finally, for node failure scenarios, we have observed recovery times which are similar to link failure scenarios.

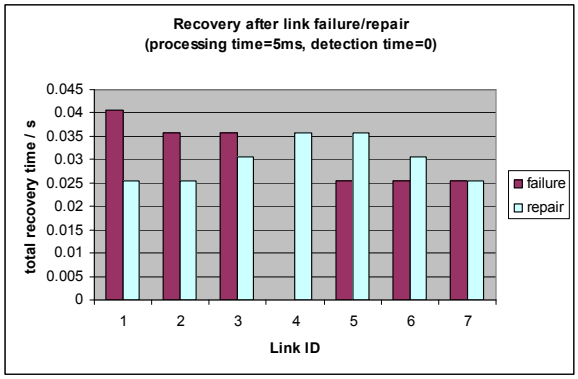

Figure 6 Total recovery time in the ring-7 network after link failure and link repair

4. NETWORK FLOODING

Layer 2 technology has no configured knowledge about the location of end stations. Instead bridges learn the source MAC addresses of frames to gain knowledge about the topological position of end stations. If a bridge receives a frame with an unknown destination MAC address, it floods the frame to all of its ports except the port where the frame was received. This mechanism floods frames with unknown MAC address along the active topology throughout the whole LAN. The same applies to broadcast and multicast traffic which must be delivered to all stations attached to the LAN. Therefore, the LAN is also called broadcast domain. The broadcast domain can be further restricted through VLANs.

The amount of flooding traffic depends directly on the number of end stations in the broadcast domain. Therefore, the network planner will typically limit the number of end stations per broadcast domain which subsequently limits the flooding traffic to a small percentage of the network bandwidth.

This approach may be problematic if the network contains links with very differing data rates. In this case, links with low data rate need to carry a significant portion of flooding traffic. Example: 200 end stations are connected to the broadcast domain via a 100 Mbit/s Fast-Ethernet LAN. The flooding traffic is assumed to be 100 kbit/s. Further end stations are connected using POTS with 64 kbit/s. While 100 kbit/s are just 0.1% of the Fast Ethernet LAN, the flooding traffic exceeds the bandwidth of the POTS, hence no communication is possible with stations connected using POTS.

One approach to circumvent this problem is to use VLANs and separate the network into different broadcast domains. However, this approach may lead to very high administrative effort. Another possibility is to use layer 3 routing to connect stations with slow links.

Apart from broadcast and multicast frames, traffic flooding happens after failure/repair of a link/node

which changes the active topology. After flushing the FDBs, all traffic will be flooded until the MAC addresses have been relearned. During this period, the network has a decreased capacity for user traffic.

Figure 7 qualitatively illustrates the utilization of network capacity after a link failure.

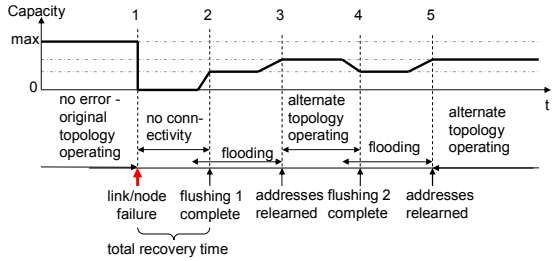

Figure 7: Qualitative illustration of usable network capacity after a link failure

After the link failure (1), the usable network capacity is zero until the spanning tree has converged to the new active topology. The RSTP protocol leads to two flushing periods. When the first flushing period has completed (2), all stations can communicate, but due to the flooding, the capacity is limited until the flooding period has completed (3). Two seconds after the first flushing wave, there is a second flushing cycle independently triggered by each station (4) which limits the capacity again. When all addresses are relearned, the flooding period finishes and the full capacity available for the currently active topology can be used.

Typically, transport networks need to carry traffic of different services. In the following we differentiate between guaranteed traffic and BE traffic:

- Guaranteed traffic has a constant data rate and requires low loss ratio, delay, and delay variation. Its delivery must be guaranteed with a high availability. One example for guaranteed traffic is voice traffic.
- On the other hand, services using BE traffic can deal with changing bandwidth (elastic traffic) and accept a wider range of loss, delay, and delay variation. The available bandwidth can be shared between different connections.

The above differentiation can easily be extended to multiple classes but this is outside of the scope of this paper.

Transport networks must be planned to carry guaranteed traffic also in the failure case. The question is whether and how far flooding reduces the quality of service and how relevant this issue is for microwave radio networks. We have limited our work to the case of aggregation networks where traffic is only carried between clients and servers. We have also assumed symmetric traffic in both directions (upstream, downstream). The capacity on all links is normalized to 1.
The capacity during the flooding period strongly depends on the network topology:

Single ring: Apart from the reduced capacity due to the failure, there is no further reduction in capacity due to flooding. However, this holds only, if all ports which connect to the ring are edge ports. In this case, while the upstream traffic is flooded, the downstream traffic is correctly delivered without flooding.

Multiple rings: There is a capacity decrease during the flooding period which depends on the number of rings. To prevent the decrease, the rings must be separated using VLANs. Even then, if only a single spanning tree instance is used, all rings are flushed after a failure However, the flooding of frames is limited to the scope of single rings. Therefore, the capacity is as in the single ring case. If the reaction to failures must be limited to single rings, MSTP with one instance per ring has to be provided. Please note that this statement assumes a single MST region.

Rings with trees: Failures in the trees can be isolated from the ring using the configuration option “restrictedTCN”. In this case, topology change BPDUs will not enter the ring. Critical are errors in the ring itself as ports which connect tree and ring cannot be configured as edge ports.

Table 1: Available capacity per station (link capacity=1, capacity equ. distr. to all stations)

	Ring-7 edge	Ring-7 no edge	2*Ring-7
No error, no flooding	0.33	0.33	0.33
No error, flooding	0.2	0.077	0.1
Error, no flooding	0.2	0.2	0.2
Error, flooding	0.2	0.077	0.1

Table 1 gives an overview on the capacity for the various network types under different conditions. The link failure leading to the lowest capacity per end station has been used as error scenario. The ring network is the only topology, where the capacity in the flooding case does not decrease below the error case which is typically used as capacity for network planning issues. Therefore, flushing needs to be taken into account or its influence must be avoided using the methods described above.

5. SPANNING TREE AND ADAPTIVE MODULATION

Microwave links work at frequencies which are subject to rain fading. Adaptive modulation allows to dynamically change the PHY mode which is a combination of modulation and coding scheme depending on weather conditions. During good weather conditions, an efficient PHY mode is used providing a high data rate. During heavy rain, adaptive modulation uses a more robust PHY mode to guarantee the availability of the link at the cost of a reduced data rate.

The network needs to be dimensioned to carry all guaranteed traffic under the worst case condition where all links use the most robust PHY mode and

where a link has failed. BE traffic may be dropped using priority CoS queuing. On the other hand, dropping BE traffic is inefficient if there is still capacity in the network to carry the traffic on other links. To gain from this available capacity, however, the routing must be done in an optimal way which takes into account the changed capacity of the links.

For illustration of the possible gain we consider the following simplified example: In the ring scenario we assume that all can use two PHY modes: 1) 4 QAM(2/3), 25 Mbit/s, 2) 64 QAM(1/1), 116 Mbit/s. For each node, the demand for guaranteed traffic is 8 Mbit/s, all other available capacity is used by BE traffic and is equally shared between the nodes.

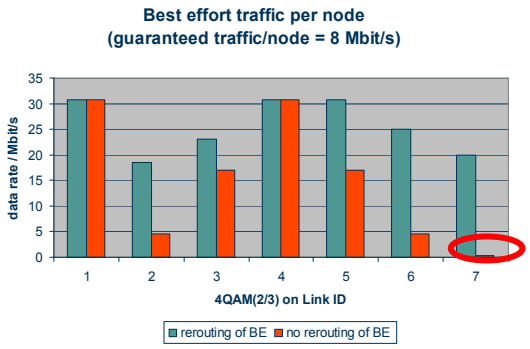

Figure 8: BE traffic per node in ring scenario, one link PHY mode 4QAM(2/3)

The data rate for BE traffic per node is shown in Figure 8. In the worst case, the rain cell crosses link L7. Even if no rerouting of BE takes, place, all guaranteed traffic can be carried. However, the BE traffic per node is limited to 333 kbit/s in this case. If we apply rerouting of BE traffic, the BE traffic per node can be increased to 20.4 Mbit/s. For other links, the effect is less dramatic but the increase is still considerable.

To achieve the optimum utilization of network capacity, BE traffic can be routed separately from guaranteed traffic. There are different options:
The guaranteed traffic is statically routed, either as TDM or using static packet cross-connections. The spanning tree is only used for BE traffic. Consequently, other mechanisms must be applied to guarantee the absence of loops for guaranteed traffic and to achieve rerouting in the failure case.

Guaranteed traffic and BE traffic are differentiated using VLANs. MSTP is applied where guaranteed traffic and BE traffic are mapped to different MST instances as illustrated in Figure 9. In case of total link or node failure, also guaranteed service can be rerouted using MSTP.
Both options ensure that guaranteed traffic is not rerouted as result of a PHY mode change, thus avoiding any disruption of that traffic. For BE traffic, short interruption times may be acceptable.

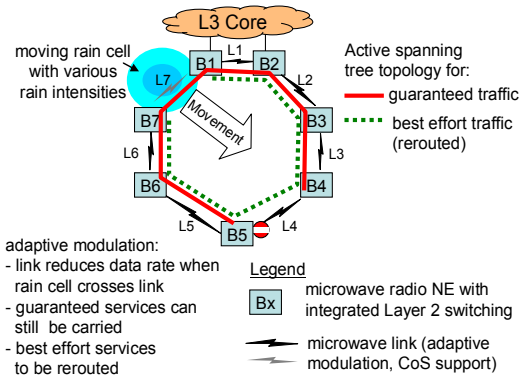

Figure 9 Illustration of MSTP use for adaptive modulation

The rerouting can be achieved by dynamically adjusting the path cost of the spanning tree for the BE traffic. The costs associated with this approach are the interruption which happens during the spanning tree recovery and the reduced capacity during the flooding period.

Adaptive modulation changes the PHY mode based on the measurement of the signal to noise ratio (SNR). Assuming that the SNR changes with bounded speed, the margin which is applied to SNR ensures that adaptive modulation can change the PHY mode without traffic hit before an error occurs. For the calculation of the total recovery time, we can therefore assume the detection time for spanning tree recovery time to be zero. The flow of BPDUs which can be observed after a path cost change is very similar to the case of a link failure. The only difference is that an additional BPDU with TC bit set is issued leading to a FDB flush at the link where path cost increases leading to a slightly higher recovery time, see Figure 10. In general, the total recovery time can be kept short if the systems are optimized for this case.

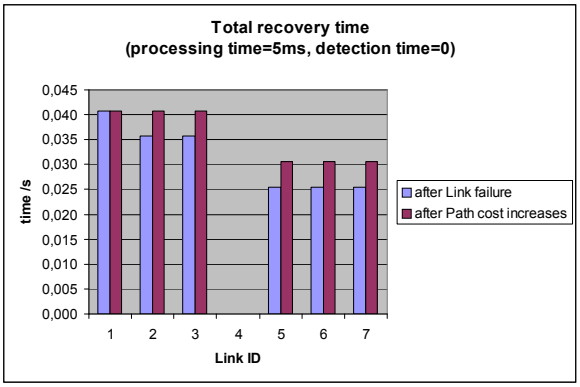

Figure 10 Comparison of total recovery time for link failure and path cost change, ring network

The above procedure causes automatic network reconfiguration. To ensure useful communication within the network, it is important to limit the dynamics. Therefore, dampening procedures must be applied which are outside of the scope of this paper. A simple mechanism is to apply a wait to restore time.

It should be noted, that the rerouting of BE traffic described above is not the result of an error scenario but an automatic optimization which happens under "normal" operating conditions. Typically, rerouting is either based on an error scenario or is due to a planned optimization step. To achieve the more optimum network usage, the operator must be willing to accept increased dynamics in the network.

CONCLUSIONS

The rapid spanning tree protocol (RSTP) and the multiple spanning tree protocol (MSTP) are standardized mechanisms which provide very fast recovery after errors in realistic network topologies. Both have significant advantage over the Spanning Tree Protocol (STP) provided that systems are optimized to achieve fast recovery.
Flooding is a crucial mechanism for layer 2 networks. To prevent flooding traffic from severely impacting network performance, its consequences need to be taken into account during the network planning process. In ring networks, the amount of flooding traffic can be limited by configuration of ports as edge ports. In other network topologies, the size of broadcast domains can be limited using VLANs.
Adaptive modulation combines high link availability with high data rate during periods of good propagation conditions. While adaptive modulation improves the performance for microwave radio links, a further increase of network utilization is possible using dynamic rerouting. During bad propagation conditions, RSTP is able to reroute best effort traffic leading to increased network efficiency. Interruptions are only for best effort traffic and can be kept short. Additionally, the usage of MSTP allows rerouting of guaranteed traffic in case of total link or node failure.

REFERENCES

[Perl85] Radia Perlman, "A Protocol for Distributed Computation of a Spanning Tree in an Extended LAN", Ninth Data Communications Symposium, Vancouver, 1985

[IEEE 802.1D-2004] IEEE Std. 802.1D-2004, IEEE Standard for Local and Metropolitan Area Networks: Media Access Control (MAC) Bridges, June 2004

[IEEE 802.1Q-2005] IEEE Std 802.1Q, 2005 Edition, IEEE Standards for Local and metropolitan area networks— Virtual Bridged Local Area Networks, 2005

[YATS] Yet another tiny simulator, TU-Dresden, http://www.ifn.et.tu-dresden.de/tk/Resources/Software/YATS/yats.html

[LUA] http://www.lua.org

ACKNOWLEDGEMENTS

The work reported in this paper has been supported by the German Ministry of Education and Research (BMBF) within the Eibone project framework under contract number 01 BP 553. The responsibility for the content of this paper is with the authors.

BIOGRAPHY

Torsten Mueller received his PhD from Dresden University of Technology, Chair for Telecommunication in 2002 where he worked on ATM, TCP/IP, and QoS. In 2001 he joined Marconi R&D and worked in the area of ASTN/GMPLS and Ethernet/SDH. He is currently with the microwave radio department at Ericsson.

Herbert Leuwer received his Dipl. Ing. (FH) from Fachhochschule Koblenz. In 1986 he joined the R&D department of ANT Nachrichtentechnik developing optical fibre measurement technology. Since 1998 he works on ATM and Ethernet switching for broadband access, next generation SDH and microwave systems. He is currently with the microwave radio department at Ericsson.

Thorsten Kaiser received his Dipl.-Ing. from RWTH Aachen. In 1993 he joined the R&D department of ANT Nachrichtentechnik developing signalling protocols and participating in standardization activities. Since 2002 he works on Ethernet switching. He is currently with the microwave radio department at Ericsson

Thomas Alberty received his Dipl.-Ing. degree in electrical engineering from Aachen University of Technology and his PhD from Munich University of Technology. He joined ANT Telecommunications in 1986. Starting with development of synchronization algorithms, he covered in the meantime all parts of signal processing in microwave systems and is now mostly interested in cross-layer aspects. He is currently with the microwave radio department at Ericsson.