

Redundant Link Management Technology in Gigabit Ethernet Networks

Ralf Aßmann
Dept. Research & Development
Schneider & Koch & Co. Datensysteme GmbH
Siemensstr. 23, 76275 Ettlingen

SK-internal Documentation
March 31, 1999

Scope

This document describes SysKonnnect's RLMT.

Document Control

Doc.number	Version	Date	Author
rlmt10.doc	1.0	31-Mar-1999	Ralf Aßmann

Contents

1. INTRODUCTION.....	2
1.1. PROBLEM	2
1.2. SOLUTION.....	2
2. DESIGN DESCRIPTION.....	3
2.1. THE PREFERRED PORT	3
2.2. RLMT MODES.....	3
2.2.1. RLMT Mode CLS.....	3
2.2.2. RLMT Mode CLP	3
2.2.3. RLMT Mode CLPSS	3
2.3. ADDRESSES	4
2.4. PACKET TYPES.....	4
2.5. PORT STATES	5
2.6. SCHEDULING SWITCHING DECISIONS.....	5
2.7. SWITCHING REASONS AND TIMES.....	6
2.8. PORT CRITERIA FOR SWITCHING.....	6
2.9. NETWORK SETUP	7
2.10. SNMP.....	10
3. RLMT AND EXISTING REDUNDANCY SOLUTIONS.....	12
3.1. RLMT VS. ADAPTER TEAMING.....	12
3.2. RLMT VS. FDDI'S STATION MANAGEMENT (SMT)	12
3.3. RLMT AND SWITCHES WITH REDUNDANT PORTS.....	12
4. RLMT EXAMPLE SESSION.....	12
5. FAQ.....	14

1. Introduction

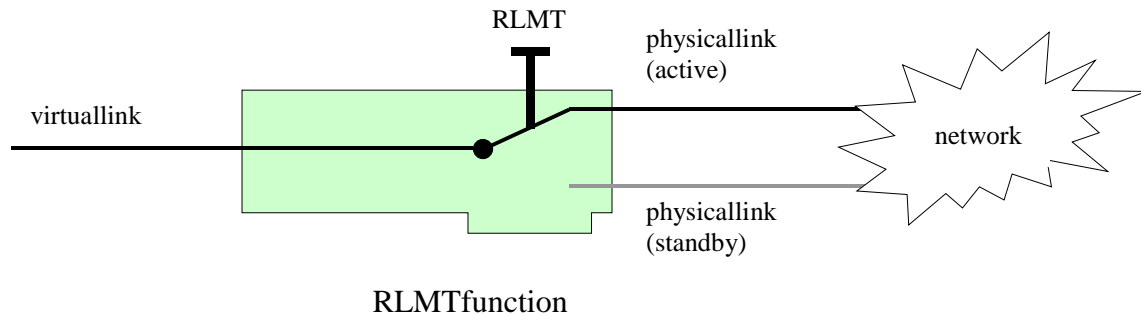
1.1. Problem

Gigabit Ethernet is widely supported in today's LANs, as it is upward compatible to Ethernet and Fast Ethernet. It has broad support through the existing Network Operating Systems (NOS). Typically it is used in backbones, connecting a company's servers. While it is one of the fastest LAN topologies today, it lacks one feature that is highly desirable in backbones: end-to-end redundancy. While it is possible to build up more than one path between every pair of switches, the link between a Gigabit Ethernet Network Interface Card (NIC) and a switch is still a single point of failure.

1.2. Solution

This is where SysKonnnect's solution hooks in: Gigabit Ethernet adapters providing two ports. Part of each driver is SysKonnnect's Redundant Link Management Technology (RLMT), which monitors the status of the ports. If the link of the active port fails, RLMT switches immediately to the standby link, thus keeping the virtual link up as long as at least one 'physical' link is up. In advanced modes, RLMT additionally monitors the network path between the two ports it drives and can detect and report network segmentation. Being

part of the driver, this function is independent of the operating system. It is just a function that is added together with the driver. This paper describes how to configure RLMT for your needs and gives insight into some technical details of RLMT.



2. Design Description

2.1. The Preferred Port

RLMT always tries to use one port for protocol traffic, the so-called “preferred port”. The “preferred port” is configurable, by default it is port A.

2.2. RLMT Modes

To meet the user’s demands, RLMT can be configured to different modes.

2.2.1. RLMT Mode CLS

RLMT’s default mode is “Check Link State” (CLS). In this mode RLMT just reacts on LinkUp and LinkDown events of ports reported by the hardware. This leaves the job of securing the path behind the connected component to the next component and the network layout.

2.2.2. RLMT Mode CLP

An advanced RLMT mode is “Check Local Ports” (CLP). In this mode, RLMT monitors the network path between the two ports of an adapter by regularly exchanging packets between them and by monitoring broadcast traffic on both ports. This mode requires a network configuration in which the two ports “see” each other (otherwise the network path between the two ports could not be monitored ...) to provide an advantage over the CLS mode.

2.2.3. RLMT Mode CLPSS

The other advanced RLMT mode is “Check Local Ports and Segmentation Status” (CLPSS). This mode does the same as the CLP mode, and additionally checks network segmentation by sending BPDU hello packets.

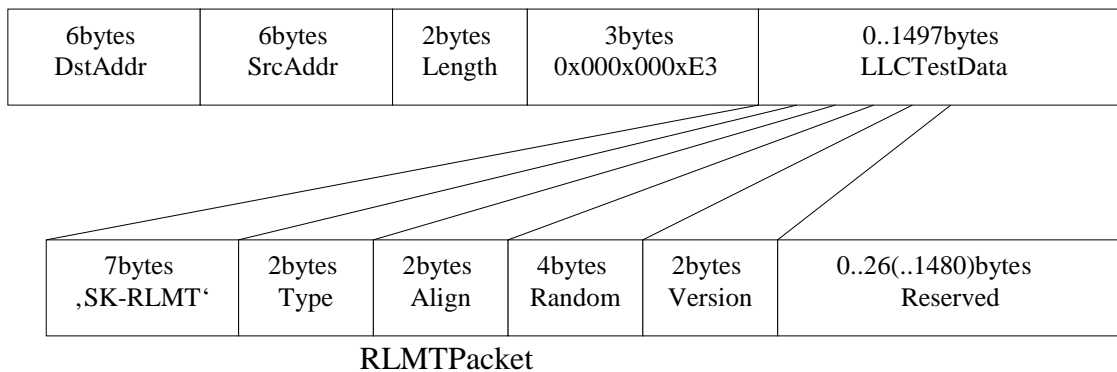
2.3. Addresses

For normal operation, RLMT assigns a unique ‘physical’ MAC address to each port it controls. This address is bound to one port and is different from the ‘logical’ MAC address that protocols use, which is always assigned to the active port. Thus the active port uses two MAC addresses—one for RLMT and one for protocols. SysKonnnect Gigabit Ethernet adapters currently use MAC addresses ending with bit combination ‘00’ for protocol usage, the same MAC address, but ending with bit combination ‘01’ is used for port A, and the same MAC address, but ending with bit combination ‘10’ is used for port B (on dual-link adapters). Bit combination ‘11’ is currently reserved.

To achieve a high level of confidence in selecting the best port that is currently available, RLMT sometimes tries to contact RLMT instances in other drivers by sending an “RLMT Check Request” packet (see below) to the RLMT multicast address 01-00-5A-52-4C-4D.

2.4. Packet Types

For communication RLMT uses LLC Test Packets ¹ with special data, further called RLMT packets:



- After each Link Up and after switching port, RLMT sends an RLMT Announce packet (Type == 1, see picture above) to the RLMT multicast address with the logical address as the source address. This packet causes the switch to learn the adapter address on the connected port.
- A change of a physical MAC address is reported using an RLMT Change packet (Type == 3).
- In CLP and CLPSS modes, RLMT Check Request packets (Type == 2) are sent between two links of an adapter that are up. In certain problem situations, this type of packet is also sent to the RLMT multicast address. This packet is always answered (best effort) with an RLMT Check Reply packet (SSAP == 1) to the port that sent the request. It is mandatory for LLC test packets that the data remain unchanged in the reply.
- In CLP and CLPSS modes, if a port does not receive any RLMT packets from another port that it is checking, it sends an RLMT Check Tx Line packet (Type == 4) to this

¹ cf. IEEE 802.2: Logical Link Control, chapter 5.4.1.1.3

port. A port receiving such a packet sends further RLMT Check packets to the RLMT multicast address until it receives a reply to its requests from somebody.

In CLPSS mode, a Bridge Protocol Data Unit (BPDU) hello packet is sent after Link Up or when RLMT wants to check network segmentation.

01-80-C2-00-00-00	6 bytes SrcAddr	2 bytes Length	3 bytes 0x420x420x03	0..1497 bytes BPDUData
-------------------	--------------------	-------------------	-------------------------	---------------------------

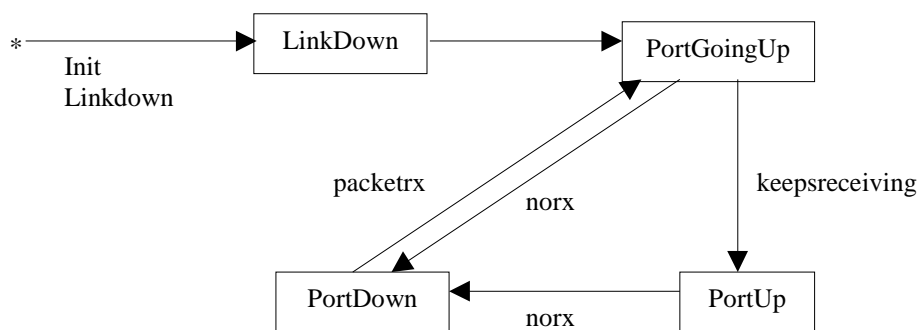
BPDU Hello Packet

In the BPDU hello packet, the port announces itself as root bridge with very low priority. Typically, the connected switch tells the real root bridge. If the root bridges reported on both ports differ, RLMT reports network segmentation. Once network segmentation is detected, it is checked and reported again every 15 minutes until the condition stops.

2.5. Port States

In RLMT, the ports are assigned different states:

- LinkDown: the link is down.
- PortDown (CLP and CLPSS modes only): the link is up, but a port did not receive for too long (i.e. 1.9 seconds) and the other port is not marked as suspicious.
- PortGoingUp: the link just came up (all modes) or the port just started receiving while in the PortDown state (CLP and CLPSS modes).
- PortUp: the link is up long enough (2.5 seconds). In CLP and CLPSS modes, this also indicates that the port received something in the last 1.9 seconds or that the port received something after it marked the other port as suspicious.



RLMT's Port State Machine (CLP and CLPSS modes)

2.6. Scheduling Switching Decisions

Definitions:

A port is called SuspectTx if it received an RLMT Check Tx Line packet and did not receive an RLMT Check Reply packet after this.

A port is called SuspectRx if it did not receive anything for too long.

² cf. IEEE P802.1d: Media Access Control (MAC) Bridges (for Spanning Tree), chapter 9,3

There are several reasons that cause RLMT to schedule as switching decisions:

- Link Up and Port Up at a port
- norx for 1.9 seconds (CLP and CLPSS modes)
- a port received while in Port Down state and not Suspect Tx (CLP and CLPSS modes)
- a port received while Suspect Rx (CLP and CLPSS modes)
- Port Down (also Link Down) of the active port
- the preferred port changes (CLP and CLPSS modes)

A possible result of a switching decision is to stay on the currently active port.

2.7. Switching Reasons and Times

Switching Reason	Typical Switching Time	Max. Switching Time
Link Down on active port (e.g. due to rx line broken or tx line broken on a link using auto-negotiation) and other port is up.	Immediately	
Tx line broken on a link using no auto-negotiation.	<2sec	<=2.4sec.
Port Up on preferred port (while the other port is active).	at Port Up (CLP mode: 2.5sec. after Link Up; CLP and CLPSS modes: at least 2.5sec. after start of non-BPDU traffic, depending on network traffic)	
Nonon-BPDU traffic on active port, but traffic on another port.	<2sec after receive stops	<=2.4sec after receive stops
Due to network segmentation, a broadcast packet is only received at one port.	0.5sec. after broadcast receive	<=1sec. after broadcast receive
Protocol traffic stops due to network segmentation and the standby port's RLMT check packet, which is sent to the RLMT multicast address, is answered by a second SK-NET GE adapter in its segment, whereas the active port gets no reply.	about 1sec.	<=2.4sec.

Table: Switching reasons and times

2.8. Port Criteria for Switching

A switching decision works as follows:

- The ports that meet the highest criteria (see below) are determined.
- If the preferred port is among these ports, it is used.
- Otherwise, if the active port is among these ports, it is used.
- Otherwise, the port with the lowest index (program-internal) of these ports is used.

Port criteria in RLMT mode CLS (best to worst):

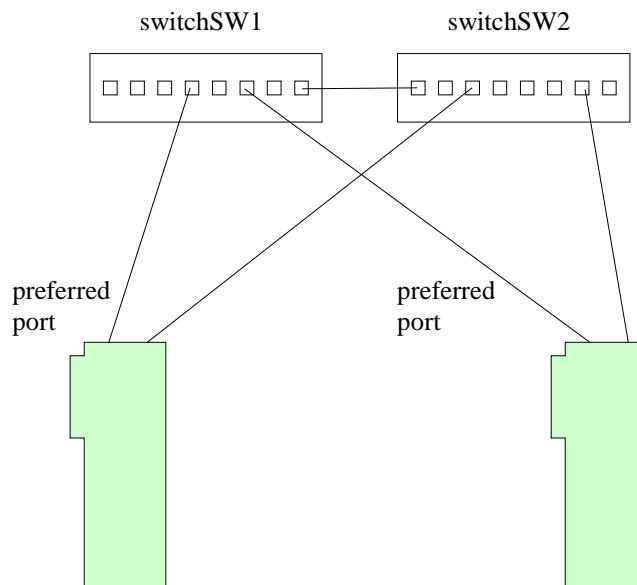
1. The port is in PortUp state, with successful auto-negotiation.
2. The port is in PortUp state, without successful auto-negotiation.
3. The port is in PortGoingUp state longer than all other ports that are PortGoingUp and the link is up with successful auto-negotiation.
4. The port is in PortGoingUp state longer than all other ports that are PortGoingUp and the link is up without successful auto-negotiation.
5. The port is in PortDown state with successful auto-negotiation.
6. The port is in PortDown state without successful auto-negotiation.

Port criteria in RLMT modes CLP and CLPSS (best to worst):

1. The port is in PortUp state and is the only port that received the last broadcast packet.
2. The port is in PortUp state and not Suspect Rx.
3. The port is in PortUp state, with successful auto-negotiation.
4. The port is in PortUp state, without successful auto-negotiation.
5. The port is in PortGoingUp state longer than all other ports that are PortGoingUp and the link is up with successful auto-negotiation.
6. The port is in PortGoingUp state longer than all other ports that are PortGoingUp and the link is up without successful auto-negotiation.
7. The port is in PortDown state with successful auto-negotiation.
8. The port is in PortDown state without successful auto-negotiation.

2.9. Network Setup

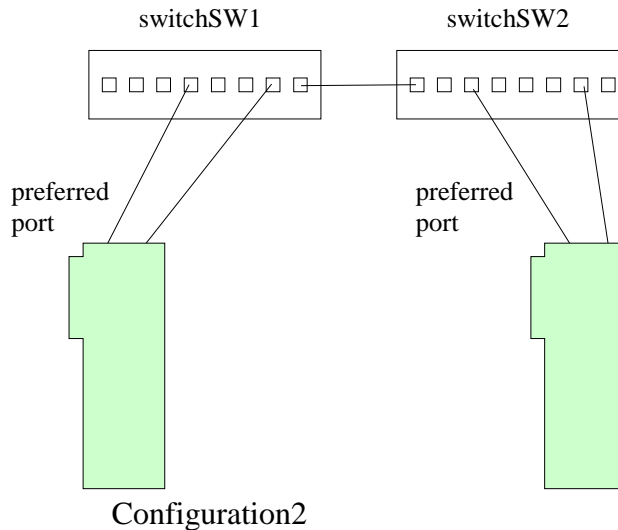
Here is a commented overview of some possible network setups that shall help you using RLMT most efficient.



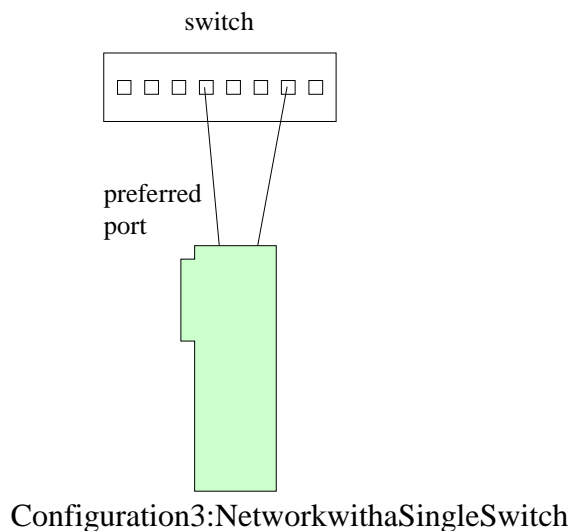
Configuration 1: Dual Homing

Dual Homing is the preferred configuration for dual-port adapters. If a switch fails, each driver will switch over to the other switch and work can continue. If the link between the

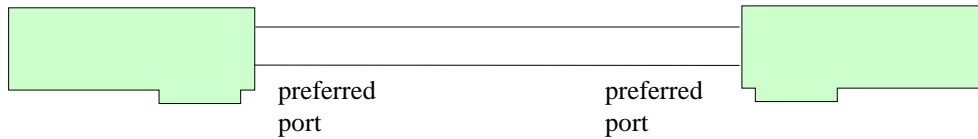
two switches fails, work will continue over switch 1. Here is a possible drawback of this configuration: if there are more switches in the net and switch 1 is segmented after the link failure, the connected stations are also. So you should carefully plan your inter-switch connections. When using several switches, the preferred ports should be grouped and the non-preferred ports should be grouped.



You can also use configuration 2. As long as the switches work, all ports will see their respective partner. But as switch failure will cause failure of all stations connected to it.

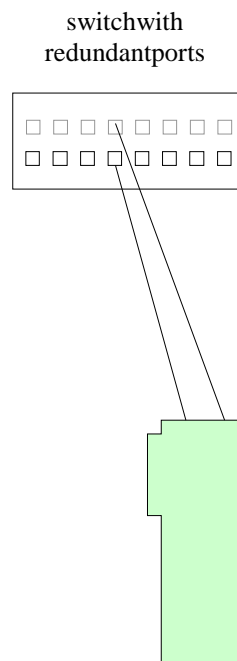


If you have only one switch, there is not much to plan: connect both adapter ports to switch ports. This configuration helps against failure of one link per adapter.



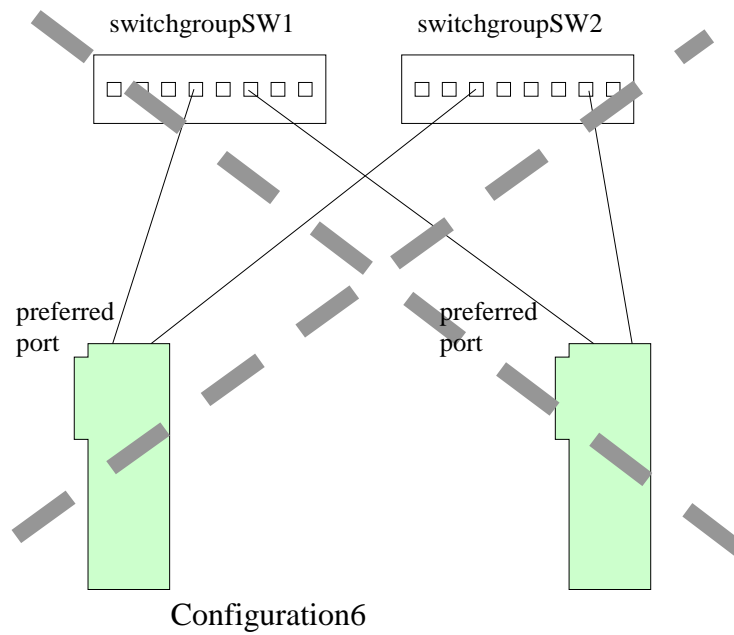
Configuration 4: Back-to-Back

An even simpler case: back-to-back connection of two dual-port adapters. Just ensure to connect both preferred ports.



Configuration 5: Switch with Redundant Ports

If you use a switch with redundant ports you may connect your dual-port adapter as shown above. But this will not give you the same protection as dual homing, as these switches do not fully activate the standby port until they want to use it. So you lose the possibility to monitor the other port over the net, and the switching time will be higher, as auto-negotiation is done after activation of the standby port.



In configuration 6, there is a possible problem: if the two switch groups SW 1 and SW 2 are not connected and one switch of a group fails, all stations connected to this switch will use the corresponding switch of group SW 2 and will thus be segmented from the other station that remains in group SW 1. Conclusion: connecting the two ports of a dual-port adapter to isolated net segments is risky and does not give you the level of redundancy that could be reached with other configurations.

2.10. SNMP

SysKonnnect provides monitoring and configuration support for RLMT with a private MIB.

Configuration Table (per adapter)

skGeRlmtEntriesNumber	Stores the number of skGeRlmtTable entries. The number corresponds to skGeNumber.												
skGeRlmtIndex	Stores the index of the adapter card that corresponds with this entry. It is identical to skGeIndex.												
skGeRlmtMode	Stores the working mode of RLMT. <table border="0" style="margin-left: 20px;"> <tr> <td>Mode</td> <td>Value</td> <td></td> </tr> <tr> <td>CLS</td> <td>0</td> <td>Always active.</td> </tr> <tr> <td>CLP</td> <td>1</td> <td>Sends RLMT Check Requests.</td> </tr> <tr> <td>CLPSS</td> <td>3</td> <td>Sends RLMT Check Requests and monitors BPDHello packets.</td> </tr> </table>	Mode	Value		CLS	0	Always active.	CLP	1	Sends RLMT Check Requests.	CLPSS	3	Sends RLMT Check Requests and monitors BPDHello packets.
Mode	Value												
CLS	0	Always active.											
CLP	1	Sends RLMT Check Requests.											
CLPSS	3	Sends RLMT Check Requests and monitors BPDHello packets.											
skGeRlmtPortActive	Stores the index of the currently active port. The value 0 indicates that currently no port is active.												
skGeRlmtPortPreferred	If RLMT evaluates that all links are well working this value is used to decide which port should be used. The value '0' indicates automatic mode and lets RLMT do the decision.												

skGeRlmtChangeCts	Counts the number of times RLMT switched from one port to another which indicates an error.
skGeRlmtChangeTimeStamp	Stores the time in hundreds of seconds since the last reboot when the connection was switched to another port.
skGeRlmtChangeEstimate	Stores the number of port switches per hour.
skGeRlmtChangeThreshold	Stores the threshold in number of port switches per hour. If the threshold is exceeded a trap is generated.
skGeRlmtPortNumber	Stores the number of physically present ports.

Statistic Table (per port)

skGeRlmtStatEntriesNumber	Number of table elements in skGeRlmtStatTable.
skGeRlmtStatGeIndex	Index of the adapter card.
skGeRlmtStatIndex	Physical index of the port.
skGeRlmtStatStatus	The working state of the port: standby (1), active (2), error (3).
skGeRlmtStatTxHelloCts	Number of hello packets sent in CLP mode.
skGeRlmtStatRxHelloCts	Number of hello packets received in CLP mode.
skGeRlmtStatTxSpHelloReqCts	Number of packets sent in CLPSS mode to force a BPDU hello packet.
skGeRlmtStatRxSpHelloCts	Number of BPDU hello packets received in CLPSS mode.

Trap definition

skGeRlmtChangeThresholdCondition	This trap is generated when the port switches exceed the threshold.
skGeRlmtChangeCondition	This trap is generated when the connection is switched to another port. The variables passed with the trap correspond to the new active port.
skGeRlmtPortDown	This trap is generated when RLMT detects that a port went logically down.
skGeRlmtPortUp	This trap is generated when RLMT detects that a port went logically up.
skGeRlmtSegmentation	This trap is generated when RLMT works in spanning-tree mode (skGeRlmtMode) and two root bridges are detected, which means a segmentation of the net. One of these segments may be isolated from the rest of the net.

3. RLMT and existing Redundancy Solutions

3.1. RLMT vs. Adapter Teaming

Competing redundancy solutions in the (Gigabit) Ethernet area are designed differently: they use existing drivers and add a new level above them which provides the redundancy.

Advantages of SysKonnnect Dual-Port Gigabit Ethernet adapter over these solutions are:

- less resources needed (IRQ, I/O, memory, CPU time)
- higher port density / less PCI slots used

The RLMT approach (redundancy software integrated into drivers) adds the following advantages:

- OS independence
- protocol independence
- faster (protocol stack is not involved in the decision if a different port must be used), resulting in less network downtime

3.2. RLMT vs. FDDI's Station Management (SMT)

A topology with built-in redundancy is Fiber Distributed Data Interface (FDDI), meanwhile available over fiber and copper. In FDDI, the Station Management (SMT) communicates with neighbor stations to build a consistent view of neighbor relationships. Unlike (Gigabit) Ethernet, FDDI is designed to build a double ring, which works despite the failure of one station. Failure of a second station starts network segmentation.

As all FDDI drivers implement SMT, a station does not need to check explicitly if a path between its two ports exists, but just checks its neighbors.

Dual homing can be realized with RLMT if you use the appropriate network configuration.

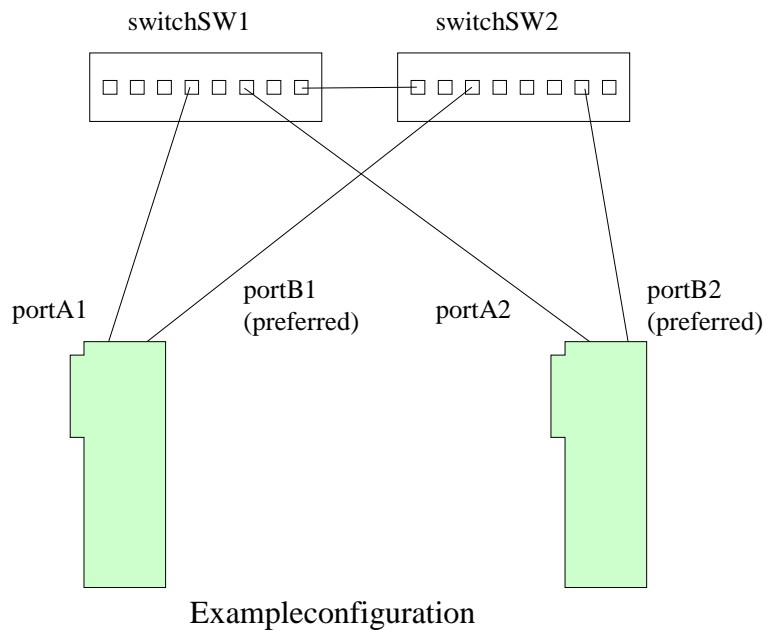
3.3. RLMT and Switches with Redundant Ports

Switches with redundant ports do not fully activate the standby port, so if the active port fails, the standby port has to spend some time in finalizing auto-negotiation. Thus, RLMT works if connected to such a switch, but it loses some control. Only CLS mode makes sense, as the standby port cannot send or receive.

4. RLMT Example Session

Configuration

- Two adapters, both A ports (A1 and A2) are connected to switch SW1, both B ports (B1 and B2) are connected to switch SW2.
- All switch ports are configured to Spanning Tree mode.
- Both drivers are configured to RLMT mode CLPSS.
- On both adapters, B is configured as the preferred port.



Initialization (in each driver)

- RLMT is started with PrefPort set to BandRlmtMode set to CLPSS.
- The link of port B comes up, causing the driver to switch to port B.
- The link of port A comes up, causing RLMT to check the segmentation state and to start sending RLMT packets.
- After 30 sec. (the time until the spanning tree reaches the forwarding state) the switch is forwarding to port B and data traffic starts.

The link B1-SW2 fails.

- RLMT is notified of port B's LinkDown and switches to port A immediately. Data traffic is not interrupted noticeably.

The link B1-SW2 is re-established.

- RLMT is notified of port B's LinkUp. As the switch is not yet forwarding data traffic, the port remains PortDown.
- After 30 sec. ³ the switch starts forwarding data traffic. Port B changes to PortGoingUp.
- After 2.5 sec. Port B gets PortUp and becomes the active port.

Up to this point, the same behavior will be seen in CLP mode.

The link between the two switches fails.

- The ports of each adapter don't see each other and start sending their RLMT requests to the RLMT multicast address. The other adapter answers these requests, so each port knows it is working.

³ This is the time the switch port needs in spanning tree mode to go into the forwarding state. It may be configured to a shorter or longer time.

- But as the desired partner doesn't answer, a BPDU hello packet is sent from each port, triggering the switches to send BPDU hello packets to each port. As these packets report different root bridges, the drivers report segmentation using the error log feature.
- After 15 min. the driver checks and reports segmentation again using the error log feature.

The link between the two switches is re-established.

- RLMT packets travel again between the ports of each adapter. No segmentation is reported anymore.

5. FAQ

1. *Does the NIC need the spanning tree protocol?*
No. In CLS and CLP modes it is not needed at all. In CLPSS mode it is only needed for the segmentation check to work. So if you configure RLMT to CLPSS mode but do not enable spanning tree at the switch, RLMT will give you the same protection you get in CLP mode.
2. *Does the NIC work with the spanning tree protocol?*
Yes; it even uses the spanning tree protocol in RLMT mode CLPSS.
3. *How much overhead does RLMT add to a driver?*
In CLS mode: sending one small packet at Link Up.
In CLP mode (dual-link adapter): CLS plus two sends and two receives every second plus a few decisions. RLMT does not send any broadcasts
In CLPSS mode (dual-link adapter): CLP plus one send and one receive at Link Up plus one send and one receive if RLMT assumes the net may be segmented.
4. *How much network traffic does RLMT produce?*
In CLS mode: one small packet at Link Up.
In CLP mode (dual-link adapter): CLS plus two small packets every second. These packets are usually directed packets, in case of any problem RLMT sends these packets as multicast packets. RLMT does not send any broadcasts.
In CLPSS mode (dual-link adapter): CLP plus two BPDU packets at Link Up plus two BPDU packets if RLMT assumes the net may be segmented.
5. *What are the failover times?*
With link failure: immediately; with timeout-based switches: about 2 seconds.
6. *What does RLMT cause to fall back to the preferred port?*
Preferred port's link is up and preferred port is rated better than other port.
7. *How many MAC addresses does RLMT use?*
Two—one for each port; the protocol uses a third address.
8. *How many IP addresses does RLMT use?*
None—RLMT is not IP-based.
9. *How does a switch recognize that the station is reachable over a different port?*
RLMT sends an RLMT Announce packet at Link Up and after a switch. This causes connected switches to learn the new route to the adapter.
10. *How is RLMT supported by management applications?*
SysKconnect provides monitoring and configuration support for RLMT with a private MIB.