P802.17/D0.2, January 2002

Proposed Draft Standard for Resilient Packet Ring Access Method & Physical Layer Specifications

(Transit path and fairness behavior)

Draft 0.2

NOTE — The editors recognize that the some text of this document has been taken from the draft Darwin. The purpose of this draft is to propose some changes to the Darwin draft in order to show how a behavior can be specified instead of possible implementations. The scope of this work is to show how the Darwin proposal can evolve into a behavior proposal and not to fix any inconsistency in the Darwin draft.

Comments on this proposal can be directed to the contributing editors:

Italo Busi
Alcatel
Via Trento, 30
20059 Vimercate (MI)
Italy
Phone: +39 039 686 7054
FAX: +39 039 686 3590
Email: italo.busi@alcatel.it

Table of contents

Abbreviations

MAC	Medium Access Control
PHY	Physical Interface

References

[B1] IEEE 802.3 – 2000 Edition

Carrier sense multiple access with collision detection (CSMA/CD) MAC and physical layer specification.

6. Media Access Control data path

6.1 Transit buffer

To be able to detect when to transmit and receive packets from the ring, RPR MAC makes use of a transit buffer as shown in Figure 6.1.

The structure of the transit buffer is an implementation option and is out of the scope of the IEEE 802.17 Standard. There are different implementations of the transit buffers. Some examples are shown in Annex K.

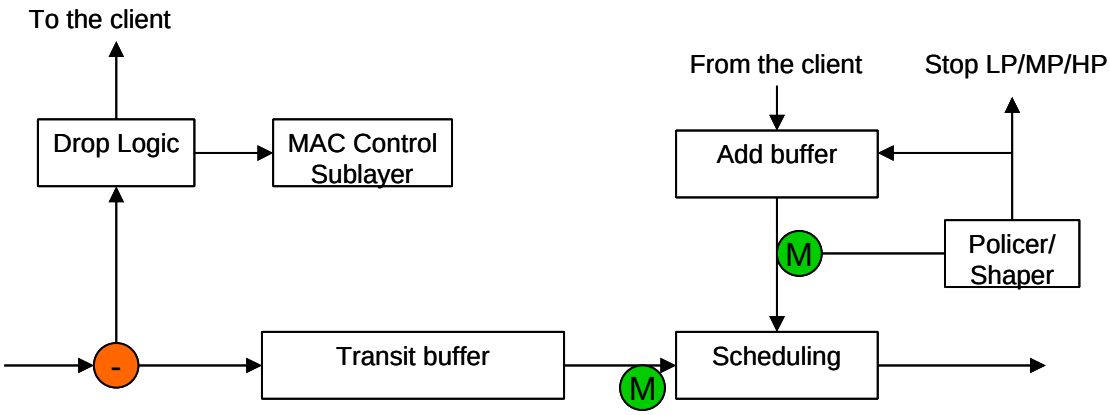

Figure 6.1 – MAC Data Path

6.2 Transmit and forwarding operation

A RPR MAC can transmit data packets from six possible flows:

1) High priority transit frame
2) Medium priority transit frame
3) Low priority transit frame
4) High priority frame from the client
5) Medium priority frame from the client
6) Low priority frame from the client

Note that Medium priority traffic is assigned a Committed Access Rate (CAR).. Traffic above the CAR will be treated as low priority, and will be referred to as “excess” MP traffic, or eMP.

The RPR MAC will decide which traffic to send based on a scheduling algorithm. The specification of such an algorithm is implementation specific and out of the scope of the IEEE 802.17 Standard. Some examples are shown in Annex K.

The scheduling algorithm shall ensure:

1) The commitments on the HP and cMP transit and add traffic.
2) A fair access between the LP and eMP transit and add traffic. The IEEE 802.17 Standard defines a per-station weighted fairness.
3) The LP and eMP add traffic should not exceed the allowed_rate parameter defined by the fairness protocol as specified in section 12.

6.3 Receive operation

Receive Packets entering a node are subject to the Header Error Check (HEC) test. If this test fails, the frame is stripped from the ring.

If the HEC test is passed, frames are then subject to the Destination Address (DA) match. If the DA matches, control frames are passed to the MAC Control Sublayer unit, while user data frames are passed to the client. If a DA matched packet is also a unicast, then the packet will be stripped.

If a packet does not DA match or is a multicast and the packet does not Source Address (SA) match, then the packet is passed to the Transit Buffer (TB) for forwarding to the next node if the packet passes Time To Live tests.

6.4 Transit operation

A series of decisions based on the type of packet, source and destination addresses are made on the MAC incoming packets. Packets can either be control or data packets. The rules for reception and stripping are given below as well as in the flow chart in Figure 6.2.

The flowchart is TBD

Figure 6.2 – RPR Receive Flowchart

1) Received packets will be discarded if there is a HEC error.
2) Conditionally decrement TTL on receipt of a packet, discard if it gets to zero; do not forward. The conditions to decrement TTL are as follows: always decrement unless the ring is in the wrap state (anywhere) and the ring id in the packet an in the MAC do not match.
3) Strip unicast packets at the destination station. Control frames are passed to the MAC Control Sublayer while user data frames are passed to the MAC client.
4) Copy multicast control frames to the MAC Control Sublayer or multicast user data frames to the MAC client.
5) Do not process packets other than for TTL and forwarding if ring identifier bit is not matched for the direction in which they are received unless the node is wrapped.
6) Packets to be sent to the MAC client due to destination address match may be optionally discarded at the MAC if there is an FCS error.
7) Transit packets may be optionally discarded at the MAC if there is an FCS error.
8) Packets with source address and ring identifier bit match should be stripped. If the node is wrapped and source address matches then the packet should be stripped.

6.4.2 Transit operation in a Bridge (Promiscuous Mode)

When the RPR MAC is part of a bridge, all data packets that do not DA match are copied to the Bridge Relay Entity and forwarded to the transit buffer.

Optional behaviors to improve bridging performance include the use of a MAC Filtering Database to hold the DA and SA of stations that are located behind the bridge: In this case the DA and SA of the packet can be checked to determine if the packet is to be dropped or stripped. If the addresses are not found in the database then the same rules as promiscuous mode apply.

6.5 Circulating packet detection (stripping)

Packets continue to circulate when transmitted packets fail to get stripped. Unicast packets are normally stripped by the destination station or by the source station if the destination station has failed. Multicast packets are only stripped by the source station. If both the source and destination stations drop out of the ring while a unicast packet is in flight, or if the source node drops out while its multicast packet is in flight, the packet will rotate around the ring continuously.

The solution to this problem is to have a TTL or Time To Live field in each packet that is set to the number of nodes in the ring. As each node forwards the packet, it decrements the TTL. If the TTL reaches zero it is stripped off of the ring. In order to allow 256 nodes on a wrapped ring, the TTL is not decremented when the packet is on the opposite ring and the ring is still wrapped. Once the ring unwraps, TTL decrements are per-formed on all packets. This catches the case where the packet is stuck on the wrong ring.

The ring identifier bit is used to qualify all stripping and receive decisions. This is necessary to handle the case where packets are being wrapped by some node in the ring. The sending node may see its packet on the reverse ring prior to reaching its destination so must not source strip it.

A potential optimization would be to allow ring identifier bit independent destination stripping of unicast packets. One problem with this is that packets may be delivered out of order during a transition to a wrap condition. For this reason, the ring identifier bit should always be used as a qualifier for all strip and receive decisions.

6.6 Wrapping of data

Normally, transmitted data is sent on the same ring to the downstream neighbor. However, if a node is in the wrapped state, transmitted data is sent on the opposite ring to the upstream neighbor. Packets of type 0x3 are marked for steering only, and when they reach a wrap point they are stripped.

6.7 Pass-thru mode

An optional mode of operation is pass-thru mode. In pass-thru mode, a node transparently forwards data. The node does not source or sink packets. It may optionally decrement the TTL and adjust the HEC but does no other modifications to the packets that it forwards. The node does not source any control packets (e.g. topology discovery or protection switch protocol) and basically looks like a signal regenerator with delay (caused by packets that happened to be in the transit buffer when the transition to pass-thru mode occurred). A node can enter pass-thru mode because of an operator command or due to a error condition such as a software crash.

The justification for continuing with the TTL decremented operation is to prevent a packet from being delivered twice if the node that sourced the packet is the node that goes into pass-thru. This could cause packets to be stripped early when topology discovery has determined that the ring contains fewer stations and adjusts the TTL value down in magnitude.

NOTE — We can use this mode also during auto-configuration to ensure that all the nodes on the ring support the same options. This will require this node to exchange topology data. Alternatively we can define another operation mode.

11. Media Access Control

11.1 Overview

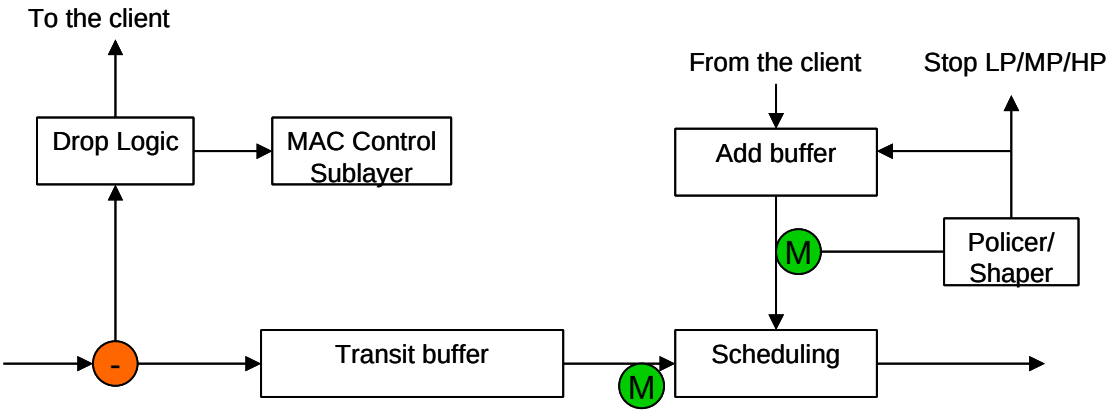

Figure 11.1 – MAC Reference Model

11.2 Traffic policing function

The policer should control that the eMP and LP traffic inserted to the ring and whose destination is beyond the choke point, does not exceed the rate allowed to this station by the fairness algorithm. A fairness message from a choke point is always transmitted with the MAC address of the choke point, thus every station receiving the message can determine exactly where the choke point is. Due the dynamic control nature of the fairness protocol, the rate at which a station is allowed to send traffic can vary in a wide range.

The policer is functionally equivalent to a leaky bucket, maximum token octets and token generation. The leaky bucket can have a size of MTU octets, which is usually provisioned for a maximum allowed burst transmission rate. The token octets are generated at the same rate the node is allowed to send eMP and LP traffic. For every octect of a eMP or LP packet that is transmitted, the token octect counter is reduced by one. When an eMP or LP packet is waiting for accessing the ring at the head of its queue, it will be granted ring access if its rate conforms to the station's fair rate and there is at least one octect token in the leaky bucket. When the number of tokens reaches the maximum token octets, token bucket shall saturate.

11.3 Pre-provision bandwidth for high priority traffic

For some application it is required to reserve certain bandwidth for high priority traffic. This reserved bandwidth can be provisioned and it cannot be allocated to eMP or LP traffic when not used. The fairness and scheduling algorithms will work as if the ring rate is decremented by the amount of the reserved bandwidth.

Bandwidth reservation is not mandatory in order to support high priority traffic.

11.4 RPR ring access operation

The fairness algorithm is used to regulate the access to the ring of the excess Medium Priority (eMP) and Low Priority (LP) traffic. High Priority and "within CAR" Medium Priority traffic does not follow the rules of the fairness algorithm and is transmitted at any time.

When a node detects congestion, it starts to advertise a normalized fair rate value to the upstream nodes. How the node detects congestion and how the normalized fair rate is determined are implementation specific details and are out of the scope of the IEEE 802.17 Standard. The advertised value should not violate the per-station weighted fairness allocation of the available bandwidth. The value that is advertised in the fairness message has to be normalized e.g. divided by the node's weight. Some implementation examples are shown in Annex K.

A node that receives a non-null fairness message will adjust its allowed rate for adding eMP and LP traffic according to the received value multiplied by its weight. This allows a node with a weigh of N to utilize N times as much bandwidth as a node with a weight of 1. If the source of the fairness message is the same node that receives it, the received value is treated as a null value.

Nodes that are not congested and receive a non-null fairness message propagate the received value to the upstream nodes.Nodes that are congested propagate the smaller of the value of the rate the node would propagate according to its internal state and the value received from the downstream node.

12. MAC fairness

12.1 Overview

The fairness algorithm provides a fair access for all stations on the ring. It consists of three components:

1) Determine congestion status
2) Determine the advertise_rate parameter
3) Determine the station allowed_rate parameter

The fairness algorithm applies to only to LP and eMP traffic coming from the MAC client. Each station is assigned a weight, which allows the user to allocate more ring bandwidth to certain station in congested saturation.

If a node experiences congestion, it will advertise a fair rate to upstream nodes via the opposite ring. How the fair rate is calculcated is implementation dependent and out of the scope of the IEEE 802.17 Standard. Some examples are shown in Annex K. The fair rate counter is run through a low pass filter function and divided by a weighting function (e.g. local station weight). The low-pass filter stabilizes the feedback, and the division by weight normalizes the transmitted value to a weight of 1.0. When the upstream stations receive an advertised fair rate, they will adjust their transmit rates of the eMP and LP whose destination is beyond the choke point, so as not to exceed the advertised value (adjusted by their weights). Nodes also propagate the advertised value received to their immediate upstream neighbor.

Stations receiving advertised values who are also congested propagate the minimum of their normalized low pass filtered advertised fair rate and the received fair rate.

The client can also take advantage of Virtual Destination Queueing (VDQ) by utilizing the multi-choke concept, which is made available to the client by fairness algorithm. VDQ combined with fairness algorithm can increase ring utilization.

The multi-choke concept deals with the case where a node wants to send traffic to a destination that is closer than a congested link. As an example, consider the case where node 1 wants to send traffic to node 2, and the link between nodes 2 and 3 is congested. The fairness algorithm will allow node 1 to send as much traffic as it wants to node 2, and will only limit traffic to nodes beyond the congested link to the fair rate.

In a multi-choke implementation, each client will track advertised fair rates for congested nodes. A node is allowed to send unlimited traffic to any node between itself and the first congested node (choke point). It can send traffic to nodes between the first and second choke point based on the first choke point's advertised fair rate. In general, a node can send traffic to a particular destination if it has satisfied the fair rate conditions for all choke points between itself and the destination.

12.2 Congestion detection

Congestion detection is strictly dependent on the actual queuing and the scheduling implementations.

Thus congestion detection mechanisms are implementation specific and out of the scope of IEEE 802.17 Standard. Some examples are shown in Annex K.

12.3 RPR fairness packet format

RPR fairness packets are sent out periodically to propagate allowed rate information to upstream stations in a unicast packet format. The recommended fair rate period is between the decay interval and 1 MTU transmission time.

See figure 23 in Darwin

Figure 12.1 – Fairness Packet format

A fair rate of all ones indicates a value of NULL.

12.3.2 When generated

The fairness messages are generated periodically.

12.3.3 Version field (3 bits)

This field is to specify the version number of fairness packet. Table 12.1 shows the fairness message version values. Type 1 fairness message is used to implement the basic algorithm. Type 2 fairness message is only needed to support multi-choke implementation. Type 1 messages are propagated hop by hop and contain the SA of the most congested node on its way while type 2 messages are broadcast and contain the SA of the node that they are originated by. Type 1 messages are processed by MAC Control block and information contained is passed to the MAC clients whereas type 2 messages are not processed by the MAC Control block and passed to the MAC clients as well.

See table 9 in Darwin

Table 12.1 – Version Values

12.3.4 Reserved field (12 bits)

It is set to 0 for type 0x01 and 0x02 fairness packets.

12.3.5 Length field (Optional 8 bits)

This is optional field within reserved field to specify the length of fairness packet. It is set to 0x00 for type 1 and 2 fairness packets.

12.3.6 Control values (16 bits)

This field is to carry the fair rate (total number of bytes/normalization_factor added to the ring by the node) to the upstream node while congestion is detected. The normalization factor is 1 for OC-48 and below, 16 for OC-192 and proportional thereafter. A value of 0xFFFF indicates the availability of up to line rate bandwidth.

Annex K Implementation Guidelines

(Informative)

This section is TBD.

NOTE — Some text can be grabbed from the Darwin proposal to show the Darwin implementations as possible example of IEEE 802.17 standard compliant implementations.