P802.17/D1.0
January 18, 2002

Proposed Draft Standard for

Information Technology -
Telecommunications and information exchange between systems -
Local and metropolitan area networks -
Specific requirements -

Part 17: Resilient packet ring access method and physical layer Specifications

Submitted to IEEE 802.17 as the Proposal - Darwin

Draft 1.0-January 18, 2002

Sponsor
LAN MAN Standards Committee of the IEEE Computer Society

Abstract: The media access control characteristics for the shared medium in ring topology are described. A set of protocols for initializing the ring, transfering packets over physical and logical ring topologies is also specified. Specifications are provided for OC48 and OC192. System considerations and management information base (MIB)specifications.

Keywords: Local area network, metropolitan area network, resilient packet ring protocol, network management,

IEEE Standards documents are developed within the Technical Committees of the IEEE Societies and the Standards Coordinating Committees of the IEEE Standards Board. Members of the committees serve voluntarily and without compensation. They are not necessarily members of the Institute. The standards developed within IEEE represent a consensus of the broad expertise on the subject within the Institute as well as those activities outside of IEEE that have expressed an interest in participating in the development of the standard.

Use of an IEEE Standard is wholly voluntary. The existence of an IEEE Standard does not imply that there are no other ways to produce, test, measure, purchase, market, or provide other goods and services related to the scope of the IEEE Standard. Furthermore, the viewpoint expressed at the time a standard is approved and issued is subject to change brought about through developments in the state of the art and comments received from users of the standard. Every IEEE Standard is subjected to review at least every five years for revision or reaffirmation. When a document is more than five years old and has not been reaffirmed, it is reasonable to conclude that its contents, although still of some value, do not wholly reflect the present state of the art. Users are cautioned to check to determine that they have the latest edition of any IEEE Standard.

Comments for revision of IEEE Standards are welcome from any interested party, regardless of membership affiliation with IEEE. Suggestions for changes in documents should be in the form of a proposed change of text, together with appropriate supporting comments.

Interpretations: Occasionally questions may arise regarding the meaning of portions of standards as they relate to specific applications. When the need for interpretations is brought to the attention of IEEE, the Institute will initiate action to prepare appropriate responses. Since IEEE Standards represent a consensus of all concerned interests, it is important to ensure that any interpretation has also received the concurrence of a balance of interests. For this reason IEEE and the members of its technical committees are not able to provide an instant response to interpretation requests except in those cases where the matter has previously received formal consideration.

Comments on standards and requests for interpretations should be addressed to:

Secretary, IEEE Standards Board
445 Hoes Lane
P.O. Box 1331
Piscataway, NJ 08855-1331
USA

IEEE Standards documents are adopted by the Institute of Electrical and Electronics Engineers without regard to whether their adoption may involve patents on articles, materials, or processes. Such adoption does not assume any liability to any patent owner, nor does it assume any obligation whatever to parties adopting the standards

Patent Statement

The developers of this standard have requested that holder's of patents, that may be required for the implementation of the standard, disclose such patents to the publisher. However, neither the developers nor the publisher have undertaken a patent search in order to identify which, if any, patents may apply to this standard.

No position is taken with respect to the validity of any claim or any patent rights that may have been disclosed. Details of submitted statements may be obtained from the publisher concerning any statement of patents and willingness to grant a license under these rights on reasonable and nondiscriminatory terms and conditions to applicants desiring to obtain such a license.

Introduction

Comments on this document or questions on the Working Group status should be addressed to the Working Group Chair:

Mike Takefman
Cisco Systems, Inc.
2000 Innovation Dr
Kanata, Ontario
Canada K2K 3E8
Phone: +1.613.254.3399
FAX: +1.613.254.4867
Email:tak@cisco.com

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

Jim Kao
Cisco Systems Inc
170 W. Tasman Dr.
San Jose, CA 95134
Email: jkao@cisco.com

Supporters:

Gunes Aybay, Riverstone Networks

Amit Banerjee, AMCC

Mark Bordogna, Agere Systems

David Cheon, Sun Microsystems

Preminder Cohan, Infineon Technologies

Spencer Dawkins, Fujitsu Network Co

Joshua Etkin, Mindspeed

Jason Fan, Luminous

Vasan Karighattam, Intel

Yongbum Kim, Broadcom

Sateesh Kumar, Redwave Networks

Heng Liao, PMC-Sierra

Dave Meyer, Mindspeed

Chuck Lee, Appian Communications

Ashwin Moranganti, Appian Communications

Stuart Robinson, PMC-Sierra

Omer Goldfisher, Corrigent Systems

Martin Green, Cisco Systems

John Hawkins, Nortel Networks

Brian Holden, PMC-Sierra

Armin Schulz, AMCC

Raj Sharma, Luminous

Bob Sultan, DataNet Associates

David Zelig, Corrigent Systems

Contributors:

Costantino Bassias, Lantern Communications

Rhett Brikovskis, Lantern Communications

Leon Bruckman, Corrigent Systems

Italo Busi, Alcatel

Bob Castellano, Jedai Broadband Networks

Angela Faber, Telecordia

Marc Holness, Nortel Networks

Wai-Chau Hui, Nortel Networks

Jeanne De Jaegher, Alcatel

Jim Kao, Cisco Systems

Carey Kloss, Cisco Systems

John Lemon, Lantern Communications

Adisak Mekkittikul, Lantern Communications

Gal Mor, Corrigent Systems

Robin Olsson, Vitesse

Harry Peng, Nortel Networks

Necdet Uzun, Cisco Systems

Steven Wood, Cisco Systems

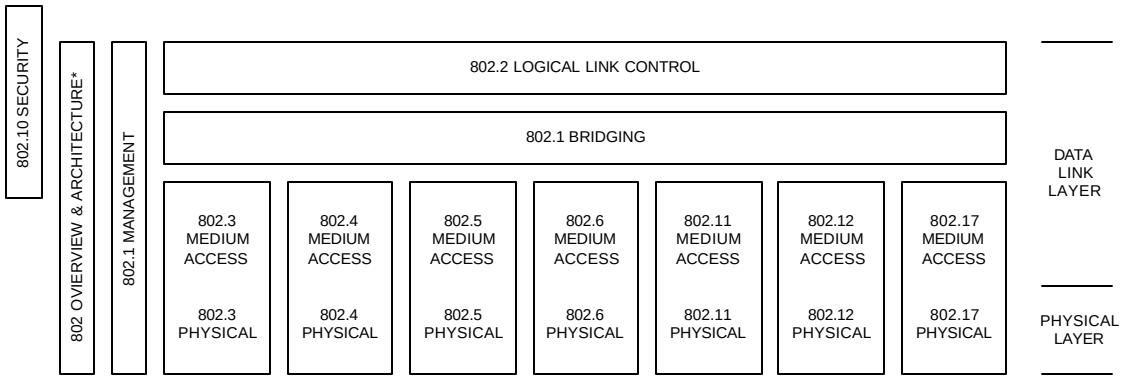
Donghui Xie, Cisco Systems

Mete Yilmaz, Cisco Systems

Pinar Yilmaz, Cisco Systems

Introduction to IEEE Std 802.17

This standard is a part of a family of standards for local and metropolitan area networks. The relastionship between the standard dn other members of the family is shown below. (The numbers in the figure refer to IEEE standard numbers.)



* Formerly IEEE Std 802.1A

This family of standards deal with the Physical and Data Link Layers as defined by the International Organizaiton for Standardization (ISO) Open Systems Interconnection (OSI) Basic Reference Model (ISO/IEC 7498-1:1994.) The access standards define (?xxx?) types of medium access technolgoies and associated physical media, each appropriate for particular applications of system objectives. Other types are under investigation.

The standards defineing the technologes noted above are as follows:

IEEE Std 802 Overview and Architecture. This standard provides an o erview to the family of IEEE 802 Standards.

.ANSI/IEEE Std 802.1B and 802.1k [ISO/IEC 15802-2]LAN/MAN Management.De .nes an OSI management-compatible architecture,and services and protocol elements for use in a LAN/MAN environment for performing remote management.

.ANSI/IEEE Std 802.1D Media Access Control (MAC)Bridges.Speci fies an architecture and protocol for the interconnection of IEEE 802 LANs below the MAC service boundary.

.ANSI/IEEE Std 802.1E[ISO/IEC 15802-4]System Load Protocol.Speci .es a set of services and protocol for those aspectsof management concerned with the loading of systems on IEEE 802 LANs.

.ANSI/IEEE Std 802.1F Common De .nitions and Procedures for IEEE 802 Management Information .

.ANSI/IEEE Std 802.1G [ISO/IEC 15802-5]Remote Media Access Control (MAC)Bridging .Speci .es extensions for theinterconnection,using non-LAN communication technologies,of geographically separated IEEE 802 LANs below the le el of the logical link control protocol.

.IEEE Std 802.1H [ISO/IEC TR 11802-5]Media Access Control (MAC)Bridging of Ethernet V2.0 in Local Area Networks .

.ANSI/IEEE Std 802.2 [ISO/IEC 8802-2]Logical Link Control.

ANSI/IEEE Std 802.3CSMA/CD Access Method and Physical Layer Specifications.

ANSI/IEEE Std 802.4 [ISO/IEC 8802-4]Token Passing Bus Access Method and Physical Layer Specifications.

ANSI/IEEE Std 802.5 [ISO/IEC 8802-5]Token Ring Access Method and Physical Layer Speci fications.

ANSI/IEEE Std 802.6 [ISO/IEC 8802-6]Distributed Queue Dual Bus Access Method and Physical Layer Specifications.

ANSI/IEEE Std 802.10Interoperable LAN/MAN Security.

ANSI/IEEE Std 802.11[ISO/IEC DIS 8802-11]Wireless LAN Medium Access Control (MAC)and Physical Layer Specifications.

ANSI/IEEE Std 802.12 [ISO/IEC 8802-12]Demand Priority Access Method, Physical Layer and Repeater Speci fications.

ANSI/IEEE Std 802.17 Resilient Packet Ring Access Method and Physical Layer Specifications.

In addition to the family of standards,the following is a recommended practice for a common Physical Layer technology:

.IEEE Std 802.7IEEE Recommended Practice for Broadband Local Area Networks.

Contents

1. Overview

1.1 Scope

This proposal defines the protocol and compatible interconnection of data communication equipment via a ring-topology Local and Metropolitan Area Network using resilient packet ring access method.

1.2 Purpose

The purpose of this protocol is to provide a scalable LAN/MAN/RAN/WAN architecture with shared access method, spatial re-use, and resiliency through fault protection method. Pursuant to this, the protocol will:

a) Support a minimum data rate of 155Mb/s, scalable to higher speeds.
b) Support for dual counter rotating ring over fiber optic and copper interconnects.
c) Efficient use of bandwidth by the use of spatial reuse and minimal protocol overhead.
d) Support for three traffic classes.
e) Scalability across a large number of stations attached to a ring
f) "Plug and play" design without a software based station management transfer (SMT) protocol or ring master negotiation as seen in other ring based MAC protocols.
g) Weighted Fairness among nodes using the ring (Each station can be assigned a proportion of the ring bandwidth).
h) Support for ring based redundancy (error detection, ring wrap, etc.) similar to that found in SONET BLSR specifications.
i) Provide media independent service interface from MAC to PHY layer.

1.3 Third-Party Operation

The RPR standard deals with ring networks that can scale from LAN to WAN environments. These networks could be operated by private organizations over their own infrastructure, private organizations over the facilities of a third party service provider or the network of a third party service provider as a public service. All three types of operations are supported.

Within the IEEE 802 family, most of the standards, such as the various versions of Ethernet, apply to local area networks, owned and operated by a single organization. In other cases, such as the 802.16 Broadband Wireless Access standard, a third-party operator is assumed. In that case, the open nature of the wireless medium makes it necessary to provide a high level of authentication and encryption, in order to restrict access to authorized users and to guarantee privacy of the transmissions.

In the case of RPR, the need for such security facilities is less, given that the difficulty of physical access to wireline facilities provides security absent in wireless systems. However, it is important to third-party service providers to be able to provide guarantees of service through service-level agreements (SLAs). The level of service provided to one user, in terms of throughput, delay, and other characteristics, should not be impacted by the actions of other users.

Thus this standard provides for a variety of classes of service and provides mechanisms to prevent users from degrading the service seen by others

1.4 Services

The applications that are expected to use an RPR network encompass the complete range of networking applications, including the familiar set of voice, video, and data. The resilient ring is ideal for critical appli-

cations that require high availability; since there are two paths between any two points on the ring, failure of a link need not prevent applications from running.

Typical uses of the network would include Internet access; in this case most traffic would exit from the RPR ring into a wide-area network or WAN for delivery to distant locations. Other uses would be for virtual private networks (VPNs) where bandwidth guarantees might be established to provide the equivalent of a hard-wired private line but with the cost savings associated with the ability of the network to recycle unused bandwidth to other users, or to allow a customer to burst at higher rates on an occasional basis

1.5 Network Properties

1.5.1 Network scale

RPR technology is optimized for the needs of metropolitan networks, in other words on a scale larger than a LAN but smaller than a typical WAN. Circumferences ranging up to several hundred kilometers are likely to be the most common. However, this does not preclude its use in other situations, for example as a building or campus backbone. It may also be used with ring circumferences ranging into thousands of kilometers with relatively little loss of responsiveness, should the fail-safe or bandwidth management properties of the network make it attractive to do so.

The data rates that can be accommodated by the RPR design cover a wide range. The protocols are designed to operate over a variety of physical layers, including SONET/SDH, Gigabit Ethernet (IEEE 802.3ab), and DWDM fiber. As higher-speed physical layers become available, it is expected that RPR will be able to work over them as well.

1.5.2 Topologies

This standard is intended for network configurations that have a ring topology, where there is a well-defined ring structure that offers two possible paths from any source to any destination. This is in contrast to a fortuitous loop in the cabling configuration, which is generally avoided in networking practice except as backup.

1.5.2.1 Dual ring

RPR uses a bidirectional ring. This can be seen as two symmetric counter-rotating ringlets. Most of the protocol finite state machines (FSMs) are duplicated for the two ringlets.

The bidirectional ring allows for two protection mechanisms to be implemented in case of media or station failure, Ring Wrap as in FDDI or SONET/SDH BLSR or Source Steering as in SONET UPSR, where the source station selects which ringlet will carry the packet.

To distinguish between the two ringlets, one is referred to as the “inner” ringlet, the other the “outer” ringlet. The RPR protocol operates by sending data traffic in one direction (known as “downstream”) and its corresponding control information in the opposite direction (known as “upstream”) on the opposite ringlet.

The nodes are able to send data on either ringlet, in other words clockwise or counter-clockwise. Generally the shorter of the two possible paths to a given destination is used, based on the node discovery scheme, but this is not required, due for instance to congestion or malfunction on one of the links. The structure of an individual node can be thought of as a dual add-drop multiplexer, as shown for a dual ring in Figure 1. The inputs to the switch include:

— Receive from upstream
— Local transmission queues

Outputs are:

— Packets for this node
— Transmit downstream

Note that multicast and broadcast packets are copied to the local receiver and transmitted downstream.

Figure 1—Switching functions at a dual-ring node.

The overall structure of the system is as shown in Figure 2.

Figure 2—Dual-ring MAN; individual nodes are as shown in Figure 1.

The use of a dual ring has substantial advantages in terms of the capacity of the system. Since there are two paths to each destination, it is possible to send traffic over the shorter path rather than the longer one; hence the longest path taken by any packet is half the ring circumference. If the destinations are randomly distributed along the ring, then the average path length is half of this, or a quarter of the circumference. Given that packets are stripped at the destination node (See “Spatial reuse” on page19.) extensive spatial re-use is possible, with the total capacity of the dual ring approaching 8 times that of a single fiber.

1.5.2.2 Shared medium

An RPR ring acts as a broadcast medium, in which a single transmission is capable of reaching all stations on the network. This means that the large number of mechanisms that have come into use with broadcast networks such as Ethernet will still work with RPR. Examples of these include the ARP protocol, the spanning tree protocol (IEEE 802.1D), and Layer 3 protocols in general.

Even under protection-switching conditions, where the rings operate as a bidirectional bus, the broadcast property still holds. Protection switching is an internal function of the RPR system and does not impact higher layers.

1.5.3 Packet-based operation

The basic unit of data on the RPR ring is a packet, consistent with current networking practice. This follows the precedent of the IEEE 802.5 Token-Passing Ring and the FDDI ring. It does differ, however, from the 802.6 Distributed Queue Dual Bus ring, which was based on cells, and SONET/SDH rings, which operate on a TDM basis with 8-bit granularity.

1.6 Resilience

The redundancy provided by two or more paths to the same destination provides resiliency in the face of fiber or equipment failure. This standard provides resiliency without the need to assign half of the fiber links to standby status. During normal operation, all paths carry traffic.

On the occurrence of network faults, due either to fiber breaks or equipment malfunction, the RPR system can continue to function. While throughput may be degraded, all functioning nodes continue to operate on the ring.

1.6.1 Source steering

Given that the ring provides two routes to any destination, if one is not operational, the other can be used. The source MAC simply must be aware of the existence of the failure; for the duration of the outage it can send all data over the operational path.

1.6.2 Wrapping protection

A RPR Ring is composed of two counter-rotating ringlets. If an equipment or fiber facility failure is detected, traffic going towards and from the failure direction is wrapped (looped) back to go in the opposite ringlet. Wrapping takes place on the nodes adjacent to the failure to re-route the traffic away from the failed span

1.7 Spatial reuse

Spatial Reuse is a concept used in rings to increase the overall aggregate bandwidth of the ring. This is possible because unicast traffic is only passed along ring spans between source and destination nodes rather than the whole ring as in earlier ring based protocols such as token ring and FDDI.

Figure3 below outlines how spatial reuse works. In this example, node 1 is sending traffic to node 4, node 2 to node 3 and node 5 to node 6. Having the destination node strip unicast data from the ring allows other nodes on the ring which are downstream to have full access to the ring bandwidth. In the example given this means node 5 has full bandwidth access to node 6 while other traffic is being simultaneously transmitted on other parts of the ring.

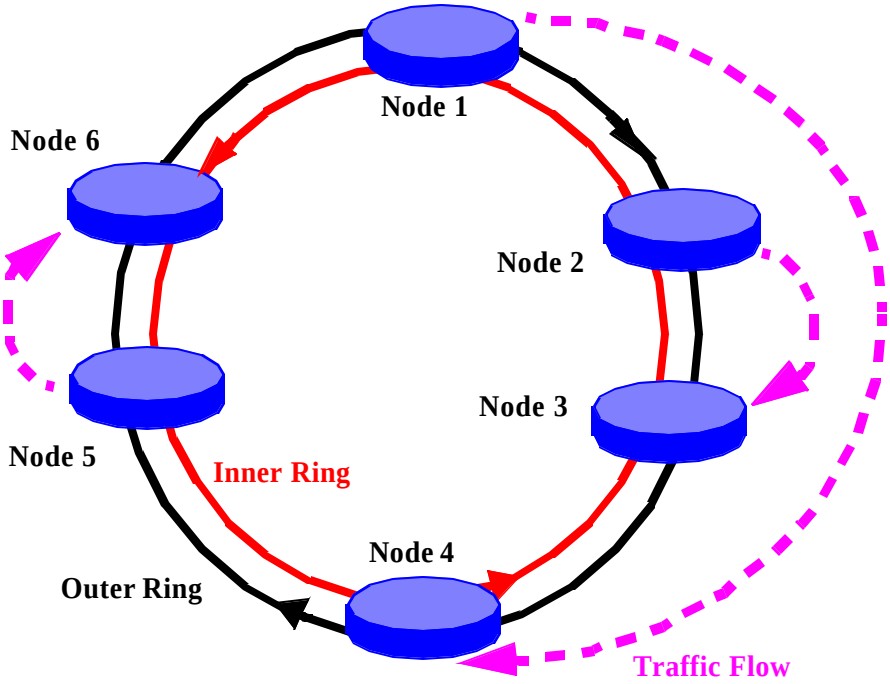


Figure 3—Global and Local Reuse

1.7.1 Destination stripping

RPR achieves a high degree of link utilization by having the destination node remove packets from the ring. This is called spatial reuse because it provides empty space on the ring which that node or another node down the fiber can use to send additional data. This is in contrast to the 802.5 Token-passing Ring or FDDI, in which packets which have already been received continue on the ring until they are removed by their original source node.

There is a provision for dealing with failed or non-existent destinations. A time-to-live or hop count field in the header makes it possible to detect packets circulating excessively and to remove them.

1.7.2 Multicast traffic

An exception to destination stripping is made for multicast and broadcast packets. Since there is no single destination to remove the packet, stripping is done by the original sender.

While the time-to-live field is decreased to zero the multicast packet also will be stripping from ring.

1.8 Bandwidth Management Features

In RPR systems, nodes cooperate to maximize traffic on the rings. They advertise information about current loads so that other nodes may determine how much traffic they can send through that node.

The bandwidth management algorithm is the heart of the RPR system; it is described in detail in Clause 11 and 12.

1.9 Fairness

Since the ring is a shared media, some sort of access control is necessary to ensure fairness and to bound latency. Access control can be broken into two types which can operate in tandem:

Global access control - controls access so that everyone gets a fair share of the global bandwidth of the ring.
Local access control - grants additional access beyond that allocated globally to take advantage of segments of the ring that are less than fully utilized.

As an example of a case where both global and local access are required, refer again to Figure3. Nodes 1, 2, and 5 will get 1/2 of the bandwidth on a global allocation basis. But from a local perspective, node 5 should be able to get all of the bandwidth since its bandwidth does not interfere with the fair shares of nodes 1 and 2

1.10 Jitter and Delay Considerations

1.11 Physical Layer Independence

The RPR design allows for wide latitude in the choice of physical layers. Various fiber-based layers, such as Gigabit Ethernet and SDH/SONET may be used. This standard includes Physical Layer Convergence Procedures (PLCP) to specify how RPR packets are carried over the physical layer, and how control information is passed between the layers. It is expected that new PLCP standards may be generated in the future as new physical layers come into use.

1.12 Application Areas

The applications environment for the resilient packet ring network is intended to be commercial and light industrial.

1.13 Conformance Requirements

Annex A will contain PICS Preform definitions for resilient packet ring network components.

1.14 Notation

This standard uses service primitives, finite state machines, state tables, and pseudo code, supplemented by prose descriptions and illustrative diagrams, to define the requirements of the protocol.

1.14.1 Service definition method and notation

The service of a layer or sublayer is the set of capabilities that it offers to a user in the next higher (sub)layer. Abstract services are specified here by describing the service primitives and parameters that characterize each service. This definition of service is independent of any particular implementation (see Figure 4).

Specific implementations may also include provisions for interface interactions that have no direct end-to-end effects. Examples of such local interactions include interface flow control, status requests and indications, error notifications, and layer management. Specific implementation details are omitted from this service specification both because they will differ from implementation to implementation and because they do not impact the peer-to-peer protocols.

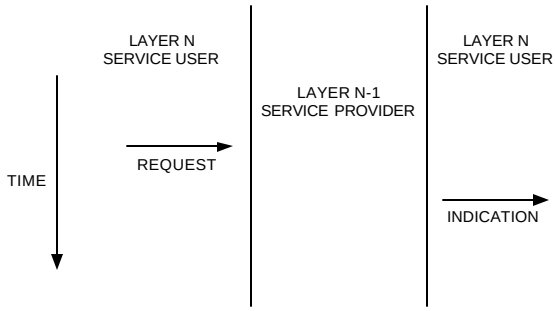


Figure 4—Service Definition

1.14.1.1 Classification of service primitives

Primitives are of two generic types:

a) REQUEST. The request primitive is passed from layer N to layer N-1 to request that a service be initiated.
b) INDICATION. The indication primitive is passed from layer N-1 to layer N to indicate an internal layer N-1 event that is significant to layer N. This event may be logically related to a remote service request, or may be caused by an event internal to layer N-1.

The service primitives are an abstraction of the functional specification and the user-layer interaction. The abstract definition does not contain local detail of the user/provider interaction. For instance, it does not indicate the local mechanism that allows a user to indicate that it is awaiting an incoming call. Each primitive has a set of zero or more parameters, representing data elements that shall be passed to qualify the functions invoked by the primitive. Parameters indicate information available in a user/provider interaction; in any particular interface, some parameters may be explicitly stated (even though not explicitly defined in the primitive) or implicitly associated with the service access point. Similarly, in any particular protocol specification, functions corresponding to a service primitive may be explicitly defined or implicitly available.

1.14.2 State diagram notation

The operation of a protocol can be described by subdividing the protocol into a number of interrelated functions. The operation of the functions can be described by state diagrams. Each diagram represents the domain of a function and consists of a group of connected, mutually exclusive states. Only one state of a function is active at any given time (see Figure5)

Each state that the function can assume is represented by a rectangle. These are divided into two parts by a horizontal line. In the upper part the state is identified by a name in capital letters. The lower part contains

the name of any ON signal that is generated by the function. Actions are described by short phrases and enclosed in brackets.

All permissible transitions between the states of a function are represented graphically by arrows between them. A transition that is global in nature (for example, an exit condition from all states to the IDLE or RESET state) is indicated by an open arrow. Labels on transitions are qualifiers that must be fulfilled before the transition will be taken. The label UCT designates an unconditional transition. Qualifiers described by short phrases are enclosed in parentheses.

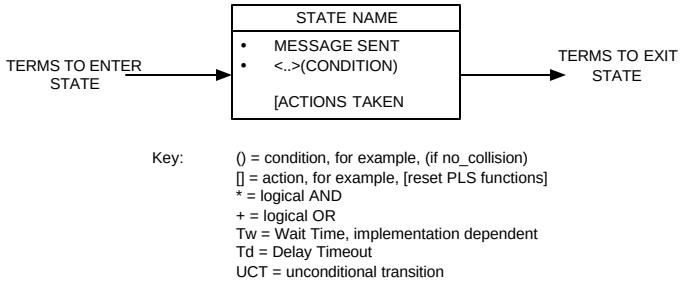


Figure 5—-State diagram notation example

State transitions and sending and receiving of messages occur instantaneously. When a state is entered and the condition to leave that state is not immediately fulfilled, the state executes continuously, sending the messages and executing the actions contained in the state in a continuous manner.

Some devices described in this standard are allowed to have two or more ports. State diagrams that are capable of describing the operation of devices with an unspecified number of ports, required qualifier notation that allows testing for conditions at multiple ports. The notation used is a term that includes a description in parentheses of which ports must meet the term for the qualifier to be satisfied (e.g., ANY and ALL). It is also necessary to provide for term-assignment statements that assign a name to a port that satisfies a qualifier. The following convention is used to describe a term-assignment statement that is associated with a transition:

a)) The character "**:**" (colon) is a delimiter used to denote that a term assignment statement follows.
b)) The character "<=" (left arrow) denotes assignment of the value following the arrow to the term preceding the arrow.

The state diagrams contain the authoritative statement of the functions they depict; when apparent conflicts between descriptive text and state diagrams arise, the state diagrams are to take precedence. This does not override, however, any explicit description in the text that has no parallel in the state diagrams.

The models presented by state diagrams are intended as the primary specifications of the functions to be provided. It is important to distinguish, however, between a model and a real implementation. The models are optimized for simplicity and clarity of presentation, while any realistic implementation may place heavier emphasis on efficiency and suitability to a particular implementation technology. It is the functional behavior of any unit that must match the standard, not its internal structure. The internal details of the model are useful only to the extent that they specify the external behavior clearly and precisely.

2. Normative references

The following standards contain provisions which, through reference in this text, constitute provisions of this standard. At the time of publication, the editions indicated were valid. All standards are subject to revision, and parties to agreements based on this standard are encouraged to investigate the possibility of applying the most recent editions of the standards indicated below. Members of IEC and ISO maintain registers of currently valid International Standards.

> [**Editor's note**: The following references are lifted from other 802 specifications, and suitably modified (deletions). As IEEE 802.17 includes external PHY layer references, the respective standard must be added here as a normative reference. This paragraph is to be kept during draft process and deleted before publication]

ANSI X3T9 FDDI Specification

[Bellcore GR-1230, Issue 4, Dec. 1998, "SONET Bidirectional Line-Switched Ring Equipment Generic Criteria".

ANSI T1.105.01-1998 "Synchronous Optical Network (SONET) Automatic Protection Switching"

Malis, A. and W. Simpson, "PPP over SONET/SDH", RFC 2615, June 1999.

Simpson, W., "PPP in HDLC-like Framing", STD 51, RFC 1662, July 1994.

ANSI/TIA/EIA-568-A, Commercial Building Telecommunications Cabling Standard.CISPR 22: 1993, Limits and Methods of Measurement of Radio Interference Characteristics of Information Technology Equipment.[3]

IEC 60060 (all parts), High-voltage test techniques.[4]

IEC 60068, Basic environmental testing procedures.

IEC 60096-1: 1986, Radio-frequency cables, Part 1: General requirements and measuring methods and Amd. 2: 1993.

IEC 60793-1: 1995, Optical fibres—Part 1: Generic specification.

IEC 60793-2: 1992, Optical fibres—Part 2: Product specifications.

IEC 60794-1: 1996, Optical fibre cables—Part 1: Generic specification.

IEC 60794-2: 1989, Optical fibre cables—Part 2: Product specifications.

IEC 60825-1: 1993, Safety of laser products—Part 1: Equipment classi.cation, requirements and user's guide.

IEC 60825-2: 1993, Safety of laser products—Part 2: Safety of optical fibre communication systems.

IEC 60874-1: 1993, Connectors for optical fibres and cables—Part 1: Generic specification.

IEC 60874-10: 1992, Connectors for optical fibres and cables—Part 10: Sectional specification, Fibre optic connector type BFOC/2.5.

IEC 60950: 1991, Safety of information technology equipment.

IEC 61000-4-3, Electromagnetic Compatibility (EMC)—Part 4: Testing and measurement techniques—Section 3: Radiated, radio-frequency, electromagnetic field immunity test.

IEC 61754-4: 1997, Fibre optic connector interfaces—Part 4: Type SC connector family.

IEEE Std 802-1990, IEEE Standards for Local and Metropolitan Area Networks: Overview and Architecture. [5]

IEEE Std 802.1F-1993 (Reaff 1998), IEEE Standards for Local and Metropolitan Area Networks: Common Definitions and Procedures for IEEE 802 Management Information.

IEEE P802.1Q/D11 (July 30, 1998), Draft Standard for Local and Metropolitan Area Networks: Virtual Bridged Local Area Networks.[6]

IETF RFC 1155, Structure and Identification of Management Information for TCP/IP-based Internets, Rose, M., and K. McCloghrie, May 1990.[7]

IETF RFC 1157, A Simple Network Management Protocol (SNMP), Case, J., Fedor, M., Schoffstall, M., and J. Davin, May 1990.

IETF RFC 1212, Concise MIB Definitions, Rose, M., and K. McCloghrie, March 1991.

IETF STD 17, RFC 1213, Management Information Base for Network Management of TCP/IP-based internets: MIB-II, McCloghrie K., and M. Rose, Editors, March 1991.

IETF RFC 1215, A Convention for Defining Traps for use with the SNMP, M. Rose, March 1991

IETF RFC1662: PPP in HDLC Like Framing, W. Simpson, July 1994

IETF RFC 1901, Introduction to Community-based SNMPv2, Case, J., McCloghrie, K., Rose, M., and S. Waldbusser, January 1996.

IETF RFC 1902, Structure of Management Information for Version 2 of the Simple Network Management Protocol (SNMPv2), Case, J., McCloghrie, K., Rose, M., and S. Waldbusser, January 1996.

IETF RFC 1903, Textual Conventions for Version 2 of the Simple Network Management Protocol (SNMPv2), Case, J., McCloghrie, K., Rose, M., and S. Waldbusser, January 1996.

IETF RFC 1904, Conformance Statements for Version 2 of the Simple Network Management Protocol

(SNMPv2), Case, J., McCloghrie, K., Rose, M., and S. Waldbusser, January 1996.

IETF RFC 1905, Protocol Operations for Version 2 of the Simple Network Management Protocol (SNMPv2), Case, J., McCloghrie, K., Rose, M., and S. Waldbusser, January 1996.

IETF RFC 1906, Transport Mappings for Version 2 of the Simple Network Management Protocol (SNMPv2), Case, J., McCloghrie, K., Rose, M., and S. Waldbusser, January 1996.

IETF RFC 2233, The Interfaces Group MIB using SMIv2, McCloghrie, K., and F. Kastenholz, November 1997.

IETF RFC 2271, An Architecture for Describing SNMP Management Frameworks, Harrington, D., Presuhn, R., and B. Wijnen, January 1998.

IETF RFC 2272, Message Processing and Dispatching for the Simple Network Management Protocol (SNMP), Case, J., Harrington D., Presuhn R., and B. Wijnen, January 1998.

IETF RFC 2273, SNMPv3 Applications, Levi, D., Meyer, P., and B. Stewart, January 1998.

IETF RFC 2274, User-based Security Model (USM) for version 3 of the Simple Network Management Protocol (SNMPv3), Blumenthal, U., and B. Wijnen, January 1998.

IETF RFC 2275, View-based Access Control Model (VACM) for the Simple Network Management Protocol (SNMP), Wijnen, B., Presuhn, R., and K. McCloghrie, January 1998.

IETF RFC2615:PPP over SONET/SDH, A. Malis, W. Simpson, June 1999

ISO/IEC 15802-1: 1995, Information technology—Telecommunications and information exchange between systems—Local and metropolitan area networks—Common specifications—Part 1: Medium Access Control (MAC) service definition.[8]

ISO/IEC 2382-9: 1995, Information technology—Vocabulary—Part 9: Data communication.

ISO/IEC 7498-1: 1994, Information technology—Open Systems Interconnection—Basic Reference Model: The Basic Model.

ISO/IEC 7498-4: 1989, Information processing systems—Open Systems Interconnection—Basic Reference Model—Part 4: Management Framework.

ISO/IEC 8824: 1990, Information technology—Open Systems Interconnection—Specification of Abstract Syntax Notation One (ASN.1).

ISO/IEC 8825: 1990, Information technology—Open Systems Interconnection—Specification of basic encoding rules for Abstract Syntax Notation One (ASN.1).

ISO/IEC 9646-1: 1994, Information technology—Open Systems Interconnection—Conformance testing methodology and framework—Part 1: General concepts.

ISO/IEC 9646-2: 1994, Information technology—Open Systems Interconnection—Conformance testing methodology and framework—Part 2: Abstract test suite specification.

ISO/IEC 10040: 1992, Information technology—Open Systems Interconnection—Systems management overview.

ISO/IEC 10164-1: 1993, Information technology—Open Systems Interconnection—Systems management —Part 1: Object Management Function.

ISO/IEC 10165-1: 1993, Information technology—Open Systems Interconnection—Management information services—Structure of management information—Part 1: Management Information Model.

ISO/IEC 10165-2: 1992, Information technology—Open Systems Interconnection—Structure of management information: De.nition of management information.

ISO/IEC 10165-4: 1992, Information technology—Open Systems Interconnection—Management information services—Structure of management information—Part 4: Guidelines for the definition of managed objects.

ISO/IEC 10742: 1994, Information technology—Telecommunications and information exchange between systems—Elements of management information related to OSI Data Link Layer standards.

ISO/IEC 11801: 1995, Information technology—Generic cabling for customer premises.

ISO/IEC 15802-2: 1995 [ANSI/IEEE Std 802.1B-1992 and IEEE Std 802.1k-1993], Information technology —Telecommunications and information exchange between systems—Local and metropolitan area networks —Common specifications—Part 2: LAN/MAN Management.

ISO/IEC 15802-3: 1998 [IEEE Std 802.1D, 1998 Edition), Information technology—Telecommunications and information exchange between systems—Local and metropolitan area networks—Common specifications —Part 3: Media Access Control (MAC) bridges.[9]

ITU-T Recommendation G.957 (1995) Digital line systems—Optical interfaces for equipments and systems relating to the synchronous digital hierarchy.[10]

ITU-T Recommendation I.430 (1995), Basic user-network interface—Layer 1 specification.

System Packet Interface Level 3 (SPI-3): OC48 System Interface for Physical and Link Layer Devices. Implementation agreement: OIF-SPI3-01.0

System Packet Interface Level 4 Phase 1 (SPI-4.1): OC192 System Interface for Physical and Link Layer Devices. Implementation agreement: OIF-SPI4-01.0

System Packet Interface Level 4 Phase 2 (SPI-4.2): OC192 System Interface for Physical and Link Layer Devices. Implementation agreement: OIF-SPI4-02.0

3. Terms and Definitions

Footnotes will be deleted from final document unless otherwise noted.

Text appearing in angle brackets will be removed prior to ballot on the 802.17 standard. This material is maintained for discussion purposes during the standardization effort.

Each definition listed below will contain a comment at the end of the definition giving the first place of usage within this standard. The comments will be of the form ì(See IEEE 802.17, Clause 10.)î or ì(See IEEE 802.17, 4.5.6.7.)î. All definitions not referenced to a usage in this standard shall be removed without need of vote before balloting of the standard.

3.1 802.17: See IEEE Std. 802.17.

3.2 agent: [802.3-2000 1.4.30 (modified)] A network management entity (NME) which can be used to configure the station and/or collect data describing operation of that station.

3.3 <all-stations MAC address[1]: TBD >

3.4 backpressure: The sending of a *control frame* in the *upstream* direction, to stop or slow the flow of *data traffic*.

3.5 <bandwidth: Note: The term *bandwidth* is applicable to the *physical layer* and should **not** be used in reference to the *MAC layer*. The term *data-rate or capacity* is used instead.>

3.6 best-effort service (BES): A service not providing any *QoS* guarantee.

3.7 bit error ratio (BER): [802.3-2000 1.4.47] The ratio of the number of bits *received* in error to the total number of bits *received*.

3.8 bit rate: [ISO/IEC2382-09 9.03.01 (modified)] The speed at which bits are transferred.

3.9 bridge: [IEC2382-25 25.01.12 (modified)] A functional unit that interconnects two networks that use the same logical link control protocol but may use different *medium access control protocols*. *Local area networks (LANs)* and *metropolitan area networks (MANs)* are example of *networks* that a *bridge* may interconnect.

3.10 bridged network: [(C/LM) 10038-1993, 802.1G-1996 (modified)] A concatenation of individual *networks* interconnected by *MAC bridges*.

3.11 broadcast address: [ISO/IEC2382-25 25.01.13] A *group address* that identifies the set of all *stations* on the *network*.

3.12 broadcast: [802.5-1998 1.3.10] The act of sending a *frame* addressed to all *stations*.

3.13 buffer insertion ring (BIR): An *access technique* for ring media that gives absolute priority to passthru traffic except when transmission of an ingress frame is in progress.

3.14 buffer: An area of memory used for temporary storage of *frames*.

[1] Brian Holden to investigate.

3.15 bursty (burstiness): Characterization of traffic as using the maximum data-rate of a channel only intermittently.

3.16 capacity: The maximum *data-rate* supported by a *medium* or *channel*.

3.17 channel: see *transmission channel*.

3.18 class of service (CoS): The categorization of traffic according to *relative* delivery priority.

3.19 closed user group (CUG): [09.08.14, 610.7-1995] A specified group of *network* users who are permitted communications among themselves but not with other *network* users.

3.20 committed burst size (Bc): [ITU I.233.1 A.5 (modified)] The maximum amount of data (in bits) that the *network* agrees to *transfer*, under normal conditions, during a time interval *Tc*.

3.21 committed information rate (CIR): [ITU I.233.1 A.8 (modified)] The information transfer rate which the network is committed to transfer under normal conditions. The rate is averaged over a minimum interval of time (*Tc*).

3.22 committed rate measurement interval (Tc): [ITU I.233.1 A.7 (modified)] The time interval during which the user is allowed to send only the committed amount of data (*Bc*) and the excess amount of data (*Be*).

3.23 congestion avoidance: [ITU I.233.1 A.11 (truncated)] Procedures initiated at or prior to the onset of mild congestion in order to prevent congestion from becoming severe.

3.24 congestion control: [ITU I.233.1 A.9] Real-time mechanisms to prevent and recover from congestion during periods of coincidental peak *traffic* demands or *network* overload conditions (e.g. resource failures). *Congestion control* includes both *congestion avoidance* and *congestion recovery* mechanisms.

3.25 congestion management: [ITU I.233.1 A.10] This includes *network* engineering, OAM procedures to detect the onset of congestion, and real-time mechanisms to prevent or recover from congestion. Congestion management includes, but is not limited to, *congestion control*, *congestion avoidance*, and *congestion recovery*.

3.26 congestion recovery: [ITU I.233.1 A.12 (truncated)] Procedures initiated to prevent congestion from severely degrading the end user perceived *quality of service*(s) delivered by the *network*.

3.27 congruent ringlets: *Ringlets* that share the same set of *stations*, but a distinct set of *links*, such that the order of station traversal via the links is identical or is exactly reversed.

3.28 control frame: A *frame* carrying only *MAC sublayer* control information.

3.29 control latency: Interval between the time that a *control frame* is sent from a *station* and the time that the effect of that *control frame* is observable at the issuing *station*.

3.30 conversation: [IEEE 100 (C/LM) 802.3 ad-2000] A set of *MAC frames transmitted* from one *end station* to another, where all of the *MAC frames* form an ordered sequence, and where the communicating *end stations* require the ordering to be maintained among the set of *MAC frames* exchanged.

3.31 copy (copying): Replication of an *inbound frame* by the *MAC sublayer* (independent of whether or not the *frame* is *stripped*).

3.32 customer separation: The property that data associated with one group of network users (e.g. a customer organization) is not communicated to a different group of *network* users.

3.33 <cut-thru: The *passthru* of a *frame* through a *station* such that the first bit of the *frame* is *retransmitted* before the last bit is *received*[2].>

3.34 cyclic redundancy check (CRC): A form of error check used to ensure the accuracy of *transmitting* a message. Note: The *CRC* is the result of a calculation carried out on the set of *transmitted bits* by the *transmitter*. The *CRC* is encoded into the *transmitted signal* with the data. At the *receiver*, the calculation creating the *CRC* may be repeated, and the result compared to that encoded in the signal.

3.35 data delivery ratio (DDR): [FRF.13 section 5 (modified)] Reports the network's effectiveness in transporting offered data. The DDR is a ratio of successful payload octets received to attempted payload octets transmitted.

3.36 data frame: A *MAC frame* carrying data supplied by the *MAC client*.

3.37 data-rate: The rate at which information is transferred, measured in bits-per-second.

3.38 <data-stream (stream): [(C) 610.10-1994 (modified)] A continuous stream of data elements being *transmitted*.>

3.39 delivered duplicated frames: [ITU I.233.1 A.15 (modified)] A *frame received* at a *destination* such that the frame was not generated by the *source station* identified by the *source address* and the *frame* is exactly the same as a *frame* that was previously delivered to that *destination*.

3.40 delivered errored frames: [ITU I.233.1 A.14 (modified)] The number of *frames* for which the value of one or more of the bits in the *frame* is in error, or when some, but not all, bits in the *frame* are lost bits or extra bits (i.e. bits that were not present in the original signal).

3.41 delivered out-of-sequence frames: [ITU I.233.1 A.16 (modified)] A frame (Ft) arriving at a *destination station* after a *frame* F_{t+1}, F_{t+2}, F_{t+3}Ö., F_n in a sequence of *frames* F_1, F_2, F_3, Ö. , F_n sent from a *source station*.

3.42 destination station (destination): [from IEEE 100 (C/BA) 1355-1995 (modified)] The *station(s)* on a network that is(are) the intended recipient(s) of an *802.17 frame*.

3.43 discard eligibility (DE): [FRF glossary (modified)] A bit indicating that a *frame* may be discarded in preference to other frames if congestion occurs, in order to maintain the committed *quality of service* within the *network*.

3.44 downstream: The direction of data flow.

3.45 dual-ring: A *ring* composed of exactly two *congruent ringlets* having opposite orientations.

3.46 1.46 <dynamic bandwidth allocation: Candidate definition should be proposed. The T&D Ad Hoc noted that some people think of this as a synonym for fairness.>

3.47 1.47 <effective transfer rate: [ISO/IEC2382-09 9.05.22 (modified)] The number of bytes *transferred* between two points per unit time and accepted as valid at the *destination.*>

[2] LAN switches typically perform cut-thru of 802.3 frames after reception of the destination *MAC address* (first six bytes of the *frame*). LAN switch cut-thru is not described in 802.1D but is left as a device specific feature. The frame check is not performed for frames that are cut-thru.

3.48 egress queue delay: TBD by Performance Ad Hoc

3.49 egress stripping: The removal of *frames* by the *egress station*.

3.50 egress: The direction towards the *MAC client* from the *ring* or *MAC sublayer*.

3.51 encapsulation: A process by which an entity places a *header* and, optionally, a *trailer* on an *SDU*.

3.52 <encapsulating bridge[3]:> (TBD)

3.53 end station: [802.3-2000 1.4.111 (modified)] A station attached to a *network* that is an initial source or a final destination of MAC frames transmitted across that *network*. A network layer router is, from the perspective of the *LAN*, an *end station*; a *MAC Bridge*, in its role of *forwarding MAC frames* from one *LAN* to another, is not an *end station*

3.54 excess burst size (Be): [ITU I.233.1 A.6 (modified)] The maximum amount of data by which a user can exceed *Bc* during a time interval *Tc*. This data (*Be*) is delivered in general with a lower probability than *Bc*.

3.55 fairness: The assignment of *ring ingress rates* such that available *capacity* is shared according to a specified algorithm.

3.56 flow control: A *congestion control* mechanism allowing one *station* to communicate to another *station* the information that *frame transmission* should be reduced or halted in order to avoid *buffer* overrun, or other conditions associated with congestion, at the *receiving station* and allowing the resumption of normal levels *frame transmission* when the condition is resolved.

3.57 flow: The collection of *frames* associated with a *conversation* that can be identified by one or a combination of specific values carried in the *protocol headers* at the *MAC layer* or above.

3.58 rame check sequence (FCS): [IEEE 100 (C/LM) 802.12-1995] *A Cyclic Redundancy Check (CRC)* used by the *transmit* and *receive* algorithms to detect errors in the bit sequence of a *MAC frame*

3.59 frame delivery ratio (FTR): [FRF.13 section 4] The ratio of successful *frame receptions* to attempted *frame transmissions*.

3.60 frame transfer delay (FTD): [FRF.13 section 3] The difference in milliseconds between the time a *frame* exits a *source station* and the time the same *frame* enters the *destination station*.

3.61 frame transmission time: TBD by Performance Ad Hoc

3.62 frame: (see *MAC frame*)

3.63 global fairness: TBD by working group.

3.64 global spatial reuse: The utilization of *ring capacity* by *stations* on the *ring* when the *station* to which the *capacity* is assigned does not utilize that *capacity.*

3.65 group address: [ISO/IEC2382-25 25.01.15] An *address* that identifies a group of *stations* on a *network*.

3.66 guaranteed-service (GS): Service that assures conformance to specific QoS parameter values.

[3] Bob Castellano to investigate.

3.67 IEEE Std. 802.17 (802.17): The IEEE *resilient packet ring* standard.

3.68 inbound: The direction of *frame* arrival at a *station* from a *ringlet*.

3.69 individual address: [ISO/IEC2382-25 25.01.14] An *address* that identifies a particular *station* on a *network*.

3.70 ingress queue delay: TBD by Performance Ad Hoc

3.71 ingress rate control: *Rate control* performed at the *ring ingress*.

3.72 ingress stripping: The removal of *frames* by the *ingress station*.

3.73 ingress: The direction from the *MAC client* towards the *ring* or *MAC sublayer*.

3.74 insert[4] (insertion): The placement of an *ingress frame* on the *ring* by a *station*.

3.75 interconnected rings: Non-*congruent rings* that intersect at one or more *stations.*

3.76 jitter: Variation in *delay* associated with the *transfer* of *frames* from one point in the *network* to another.

3.77 latency: The time required for information to be *transferred* between two points. Synonymous with *delay* for purposes of the 802.17 specification.

3.78 link aggregation group[5]: [IEEE 802.3-2000 1.4.154 (modified)] A group of *links* that appear to a *MAC client* as if they were a single *link*. All *links* in *a link aggregation group* connect the same pair of aggregation systems. One or more *conversations* may be associated with each *link* that is part of a *link aggregation group*.

3.79 link partner: The device at the opposite end of a *link* from the local *station.*

3.80 link: [IEEE 100 (C/LM) 802.5c-1991] A unidirectional physical and media connection between two *stations.*

3.81 local area network (LAN): [adapted from IEEE 100 (C/DIS) 1278.2-1995, 1278.3-1996[6]] A communications network designed for a user premises, typically not exceeding a few kilometers in length, and characterized by moderate to high data transmission rates, low delay, and low bit error rates.

3.82 local fairness: TBD by working group.

3.83 local spatial reuse: The utilization of common *ring capacity* by *stations* communicating across non-overlapping *segments* of the *ring.*

3.84 logical link control (LLC) sublayer: [C/LM 8802-5-1992s] That part of the *data link layer* that supports *media* independent *data link* functions, and uses the services of the *MAC sublayer* to provide services to the *network layer*.

3.85 lost frames: [ITU I.233.1 A.17 (modified)] A *frame* not delivered to the intended *destination* user within a specified time-out period, and the *network* is responsible for the non-delivery.

[4] (C/LM) 11802-4-1994 is not applicable. That definition uses the term to specify device, rather than frame, insertion in the ring..

[5] 802.17 link aggregation is not necessarily identical to that specified by 802.1ae.

[6] 'moderate sized geographic area' replaced by 'user premises, typically not exceeding a few kilometers in length,'

3.86 MAC client: The *protocol layer* (or *sublayer*) immediately above the *MAC sublayer*. Generally, the *network layer* or *logical link control (LLC) sublayer.*

3.87 MAC end-to-end delay: TBD by Performance Ad Hoc

3.88 MAC frame (frame): [IEEE 100 (C/LM) 802.12-1995] The logical organization of control and data fields (e.g., addresses, data, error check sequences) defined for the *MAC sublayer*[7]. Note: The term *frame* can be prefixed with an orientation (*ingress, egress, inbound, outbound*) or an operation (*inserted, copied, stripped, passedthru*).

3.89 management information base (MIB): [802.3-2000 1.4.163] A repository of information to describe the operation of a specific network device.

3.90 maximum frame size (MFS): The maximum number of bytes in a *frame*.

3.91 maximum transfer unit (MTU): [IEEE 100 610.7-1995 (modified)] The largest *payload* that can be *transferred* across a given *physical network* in a single *frame*.

3.92 medium: See *transmission medium*.

3.93 medium access control (MAC) sublayer: (1) [IEEE 100 (C/LM) 8802-5-1995] The portion of the *data link sublayer* that controls and mediates the access to the *ring.* **(2)** [802.3-2000 1.4.167] The *data link sublayer* that is responsible for *transferring data* to and from the *physical layer*[8]. **(3)** [ISO/IEC 15802-1] The MAC service provider.

3.94 medium access delay: TBD by Performance Ad Hoc

3.95 <medium[9] **agnostic:** Denotes a *MAC sublayer* that can operate with arbitrary *physical layer* alternatives, requiring a *reconciliation sublayer* for each specific *PHY* type supported.>

3.96 <medium interface connector (MIC): [802.5-1998 1.3.36 (modified)] A connector interface at which signal transmit and receive characteristics are specified for attaching stations.>

3.97 metropolitan area network (MAN): [IEEE 100 (C/LM) 8802-6-1994] A *network* for connecting a group of individual *stations* and *networks* [for example, *local area networks (LANs)*] located in the same urban area. Note: A *MAN* generally operates at a higher speed than the networks interconnected, crosses network administrative boundaries, may be subject to some form of regulation, and supports several access methods.

3.98 misdelivered frames: [ITU I.233.1 A.16 (modified)] A *frame transferred* from a *source* to a *destination* user other than the intended *destination* user. It is considered inconsequential whether the information is correct or incorrect in content.

3.99 <MTU transparency: The ability to *passthru frames* without regard to *MTU* size.>

3.100 multicast address: [ISO/IEC2382-25 25.01.16]. A *group address* that identifies a subset of the *stations* on a *network.*

3.101 multicast: The act of sending a *frame* addressed to a group of *stations*.

---

[7] Omitted sentence: ìThe MAC frame may be constructed in either ISO/IEC 8802-3 or ISO/IEC 8802-5 format.

[8] This appears as a definition of medium access control, but it is clearly a definition of medium access control sublayer.

[9] Standards documents are split on whether this is medi*a* or medi*um.* Medium, the singular, seems more appropriate and is used here.

3.102 multi-ring: A *ring* composed of multiple *congruent ringlets*, at least two of which are *opposing ringlets*.

3.103 neighbor: [BH] A *station* that is exactly one *link* away from a given *station*.

3.104 network control host: [802.3-2000 1.4.176] A network management central control center that is used to configure *agents*, communicate with *agents*, and display information collected from *agents*.

3.105 network: A generic designation for a bridged LAN, MAN, RAN, or WAN.

3.106 opposing ringlet: A *ringlet* whose traffic circulates in the direction opposite that of a given *ringlet*.

3.107 outbound: The direction of *frame* departure from a *station* to a *ringlet*.

3.108 packet: A *frame* to which has been added those fields that are medium dependent[10].

3.109 <packetization delay: TBD by Performance Ad Hoc>

3.110 partition: One set of communicating *stations* on a *partitioned ring*.

3.111 partitioned ring: A *ring* having two or more points of failure resulting in two or more non-communicating sets of *stations*.

3.112 passthru buffer delay: TBD by Performance Ad Hoc

3.113 passthru delay[11]: TBD by Performance Ad Hoc

3.114 passthru queuing: A method of *passthru* in which *passthru traffic* is queued to allow *passthru* or *insertion* of *traffic* of higher priority and to allow a *transmission* in progress to complete.

3.115 passthru: The passing of a *frame* through a *station* via the *ring*[12].

3.116 path: The specific sequence of *stations* and *links* traversed by a *frame* in *passthru* between two *stations*.

3.117 <pause: [802.3-2000 1.4.209 (modified)] A mechanism associated with the *IEEE 802.3 MAC* specification for providing full duplex *flow control*.>

3.118 <payload agnostic: Denotes a *MAC sublayer* that is not sensitive to the contents of the *payload transferred* to/from the *MAC client*.>

3.119 physical layer (PHY): [(C/LM) 8802-5-1995] The layer responsible for interfacing with the *transmission medium*. This includes conditioning signals received from the *MAC* for *transmitting* to the *medium* and processing *signals received* from the *medium* for sending to the *MAC*.

3.120 plug-and-play: The requirement that a *station* perform *passthru*, *strip*, and *ring* control activities without manual intervention except for what may be needed for connection to the *ring*. The station may additionally *copy* and *insert frames*.

[10] This is the IEEE view of a packet. It is entirely different from the IEEE view of a packet as an L3 PDU.

[11] The term *station* is used to qualify the term *transit-delay* since the term *transit-delay* is used by *frame relay* to indicate *end-to-end transit delay*.

[12] Includes the case of wrapping, if supported.

3.121 <port: (1) The point of *ingress* for *inbound frames* and the point of *egress* for *outbound frames* with respect to the *data station.* **(2)** [IEEE 100 (C/LM) 802.1G-1996, 8802-5-1995 (modified)] A signal interface provided by stations that is generally terminated at a *medium interface connector (MIC)*.[13]>

3.122 <preemption: The interruption of a *frame* in *transmission* for the purpose of *transmitting* a *frame* of higher priority.>

3.123 propagation delay: TBD by Performance Ad Hoc

3.124 <protocol agnostic: Denotes a *MAC sublayer* that can operate with arbitrary *upper-layer protocol* alternatives.>

3.125 protocol data unit (PDU): [802.5-1998 1.3.46] Information delivered as a unit between peer entities that contains control information and, optionally, data.

3.126 protocol implementation conformance statement (PICS): 1.3.47: A statement of which capabilities and options have been implemented for a given Open Systems Interconnection (OSI) protocol.

3.127 <protocol stack delay: TBD by Performance Ad Hoc>

3.128 <QTag prefix: [802.3-2000 1.4.222] The first four octets of an Ethernet-encoded Tag Header. The Ethernet-encoded Tag Header is defined in IEEE P802.1Q.>

3.129 quality of service (QoS): One or a combination of measurable properties (parameters) defining the requirements of a given data service.

3.130 rate control: Limitation of the *traffic rate* in bytes over a specified time interval.

3.131 receive (receipt, reception): The action of a station taking a frame from the medium.

3.132 reconciliation sublayer (RS): [adapted from IEEE 100 (C/LM) 802.3 –1998 modified] A mapping function that reconciles the *signals* at the media independent interface (MII) to the *media access control (MAC)* – physical signaling sublayer (PLS) service definitions.

3.133 regional area network (RAN): (1) A *network* for connecting a group of individual *stations* and *networks* [for example, *metropolitan area networks (MANs)*] located in multiple contiguous urban areas. (2) A *MAN* spanning multiple urban areas.

3.134 <residual error rate: [ITU I.233.1 A.13 (modified[14])] As applied to *MAC layer* service: (1 – (total correct MAC SDUs delivered)/(total offered MAC SDUs)).>

3.135 resilient packet ring (RPR): (1) A *connectionless* ring-based *MAC protocol* as defined by IEEE 802.17, appropriate for *LAN*, *MAN*, or *RAN*[15]deployment[16]. (2) A collection of *stations* conforming to the *resilient packet ring protocol*, and the links forming the *ring*.

3.136 ring end-to-end delay: TBD by Performance Ad Hoc

3.137 ring latency: TBD by Performance Ad Hoc

[13] Removed 'Ports may or may not provide physical containment of channels' What, exactly, does this mean?
[14] Need help with this one.
[15] Why did we exclude WAN?
[16] Or should this be specifically the protocol standardized by IEEE802?

3.138 ring medium: The abstraction of a *ring* as a continuous closed path *transmission medium.*

3.139 ring segment (segment): The portion of a *ring* bounded by two *stations* interconnected by one or more *links*.

3.140 ring topology (topology): [IEEE 100 610.7-1995 (modified)] The logical and/or physical arrangement of *stations* on a *ring*.

3.141 ring: (1) The collection of *stations* and *links* forming a *resilient packet ring*. **(2)** The set of *congruent ringlets* forming a *resilient packet ring.*

3.142 ringlet: A closed unidirectional *path* formed by an ordered set of *stations*, and the *links* interconnecting *stations*, such that each *station* has exactly one *link* entering the *station* and one *link* exiting the *station*.

3.143 <round trip propagation time: > TBD by Performance Ad Hoc

3.144 segment: (see *ring segment*)

3.145 service data unit (SDU): [802.5-1998 1.3.59] Information delivered as a unit between adjacent entities that may also contain a PDU of the *upper layer*.

3.146 service level agreement (SLA): Contract between a *network* service provider and a customer that specifies, in measurable terms, what services the *network* service provider will furnish.

3.147 shared access: The capability of two or more *stations* to share the *capacity* of the *ring medium*[17].

3.148 simple-fairness: A class of *fairness algorithm* that assigns *equal* shares of *ring capacity*.

3.149 simultaneous access: The *insertion* of *traffic* onto the *ring medium* by two or more *stations* at the same instant in time.

3.150 source station (source): The *station* that originates an 802.17 MAC *frame* with respect to a *network.*

3.151 spatial reuse: The utilization of *ring capacity* by a *station* different from the *station* to which the *capacity* was nominally assigned.

3.152 <spatial reuse protocol (SRP): A protocol, described in IETF informational RFC 2892, August 2000.>

3.153 station (data station): [IEEE 100 1073.3.1-1994, 1073.4.1-1994, 8802-5-1995 (modified)] A device that may be attached to a *network* for the purpose of *transmitting* and *receiving* information on that *network*.

3.154 station management (SMT): [802.5-1998 1.3.66] The conceptual control element of a station that interfaces with all of the layers of the station and is responsible for the setting and resetting of control parameters, obtaining reports of error conditions, and determining if the station should be connected to or disconnected from the medium.

3.155 station passthru delay: TBD by Performance Ad Hoc

3.156 steering: The *transmission* of a *frame* on a specific *ringlet* at the *ingress station* based on knowledge of the *ring topology.*

[17] The definition of ring latency in ISO/IEC 2382-25 25.04.03 suggests that the ring is modeled as a shared medium even if it is not a continuous physical medium.

3.157 store-and-forward: A method of *passthru* such that all bits of the *frame* are *received* and *buffered* before re*transmission* begins[18].

3.158 <stream: (see *data-stream*)>

3.159 strip (stripping): The removal of a *frame* from the *ring*.

3.160 <tagged MAC frame: [802.3-2000 1.4.269] A frame that contains a QTag Prefix.>

3.161 throttle: The sending of a *control frame* to a specific *station,* to stop or slow the flow of *data traffic*.

3.162 throughput: [ITU I.233.1 A.1(modified)] The number of data bits contained in the MAC frame payload successfully transferred from *source station* to *destination station* per unit time. A frame is successfully transferred if the *FCS* check for the frame is satisfied.

3.163 time-to-live (TTL): Value carried in the *protocol header* of a *frame* in order to allow the stripping of a *frame* that has *passedthru* a sufficient number of stations. The *TTL* value is generally set to an initial value at the source and decremented at each subsequent hop. The *frame* is *stripped* when the *TTL* value reaches zero.

3.164 topology: (see *ring topology*)

3.165 traffic class: A grouping of traffic that is to be processed by a distinct set of rules.

3.166 train: A collection of two or more contiguous *frames* on the ring.

3.167 transfer: [ISO/IEC2382-09 9.03.01 (modified)] The movement of an *SDU* from one layer to an adjacent layer. Also used generally to refer to any movement of information from one point to another.

3.168 <transfer rate: [ISO/IEC2382-09 9.05.21] The number of bytes *transferred* per unit time between two points.>

3.169 transmission channel (channel): [ISO/IEC2382-9 09.02.14] A means of transmission of a signal in one direction between two stations where the signals are physically or logically isolated from the signals in other channels.

3.170 transmission medium (medium): [IEEE 100 (C/LM) 8802-6-1994, 802.5-1998 1.3.34] The material on which *information signals* may be carried; e.g., optical fiber, coaxial cable, and twisted-wire pairs.

3.171 transmission: (see *transmit*)

3.172 transmit (transmission): [(C/LM) 802.5-1989s, 8802-5-1995 (modified)] The action of a *station* placing a *frame* on the *medium.*

3.173 transparent bridging: [(C/LM) 8802-5-1995] A *bridging* mechanism in a *bridged network* that is transparent to the *end stations*.

3.174 unicast: The act of sending a *frame* addressed to a single *station.*

[18] The definition that appears in 09.07.13 ISO/IEC 2382-9 1995, 'A mode of operation of a *data network* in which data are temporarily stored before they are *retransmitted* toward the *destination*', is ambiguous, as it is not clear whether 'data' refers to a complete *frame* or some portion of a *frame*.

3.175 unknown unicast: The act of sending a *frame* addressed to a single *station,* where the location in the *network* is unknown.

3.176 upper-layers: The collection of *protocol layers* above the *data-link layer*.

3.177 upstream: The direction opposite that of the *downstream* direction.

3.178 <upstream neighbor's address (UNA): [802.5-1998 1.3.77 (modified)] The address of the *station* immediately *upstream* from a given *station*.>

3.179 verified frame: [802.5-1998 1.3.79 (modified)] A valid *frame* addressed to the *station*, for which the information field has met the validity check.

3.180 virtual LAN (VLAN): [IEEE 100 (C/LM 802.1Q-1998)] A subset of the active *topology* of a *bridged local area network.* Associated with each *VLAN* is a *VLAN Identifier (VID).*

3.181 virtual medium (VMedium): A logical partition of the *network* intended to provide *customer separation*.

3.182 weighted-fairness: A class of *fairness algorithm* that allows the assignment of unequal shares of *ring capacity*.

3.183 wide area network (WAN): [IEEE 100 (C/DIS) 1278.2-1995] A communications network designed for large geographic areas. Sometimes called *long-haul* network.

3.184 wrapping: In the case of a *dual ring*, the *transmission* of a *frame* on the ringlet opposing the ringlet on which it was received.

4. Abbreviations and acronyms

All abbreviations or acronyms not used in this standard shall be removed without need of vote before balloting of the standard.

This standard contains the following abbreviations and acronyms:

802.17	IEEE Std 802.17
Bc	committed burst size
Be	excess burst size
BER	bit Error Ratio
BES	best Effort Service
BIR	burst Information rate
CIR	committed information rate
CoS	class of service
CRC	cyclic redundancy check
CUG	closed User Group
DDR	data delivery ratio
DE	discard eligibility
EIR	excess information rate
FCS	frame check sequence
FTD	frame transfer delay
FTR	frame delivery ratio
GS	guaranteed-service
LAN	local area network
LLC	logical link control
MAC	medium access control
MAN	metropolitan area network
MFS	maximum frame size
MIB	management information base
MTU	maximum transfer unit
PDU	protocol data unit
PHY	physical layer
PICS	protocol implementation conformance statement
QoS	quality of service
RAN	regional area network
RPR	Resilient Packet Ring
RS	reconciliation sublayer
SDU	service data unit
SLA	service level agreement
SMT	station management
Tc	committed rate measurement interval
TTL	time-to-live
VLAN	virtual LAN
VDQ	virtual destination queue
WAN	wide area network

5. Media Access Control (MAC) reference and service model

5.1 Scope

This clause specifies the services provided by the MAC sublayer and the MAC Control sublayer to the client of the MAC (see Figure6). MAC clients may include the Logical Link Control (LLC) sublayer, Bridge Relay Entity, or other users of ISO/IEC LAN International Standard MAC services (see Figure7). The services are described in an abstract way and do not imply any particular implementations any exposed interface.There is not necessarily a one-to-one correspondence between the primitives and the formal procedures and interfaces.

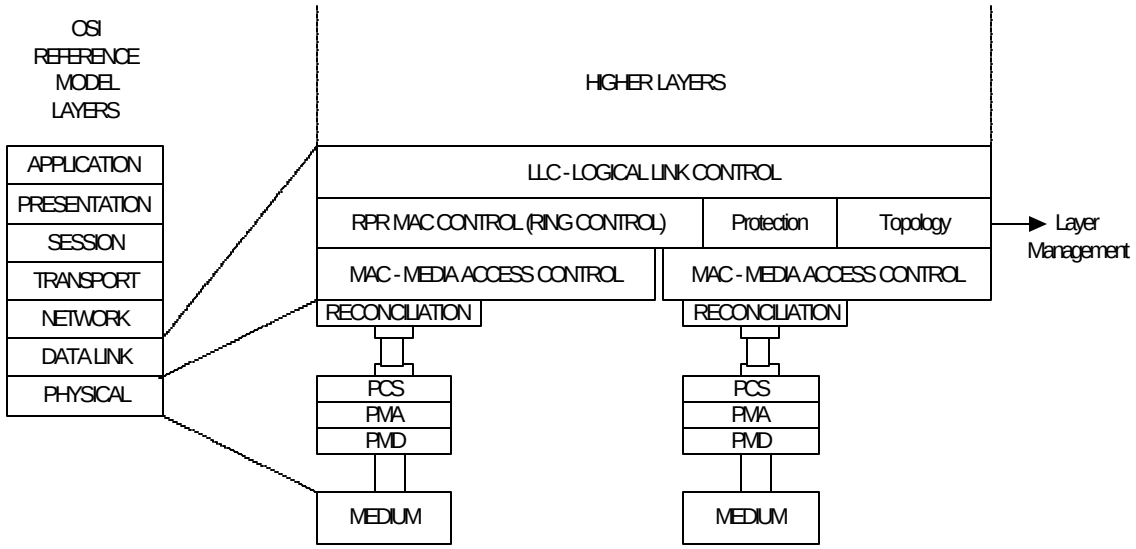


Figure 6—Service specification relation to the LAN model

5.2 Overview of MAC Services

The services provided by the MAC sublayer allow:

a) The local LLC sublayer in an end node to exchange data with peer LLC sublayer entities
b) The local LLC sublayer in an end node to exchange resilient packet ring parameters with local MAC entities.
c) The relay entity in a bridge to change data with local MAC entities in the bridge.
d) A non LLC MAC Client sublayer in an end node to exchange data with peer MAC client sub entities

The MAC sublayer is presents four service access points for the exchange of MAC client PDUs between MAC client entities. The service access points provide access to 4 logical channels at the MAC layer: high priority channel, medium priority channel, low priority channel and control channel.

5.2.1 Provisioned Service

The 802.17 MAC provides a mechanism to reserve ring bandwidth. This bandwidth will be invisible to the RPR fariness algorithm and must be wholly managed by the RPR MAC client. Traffic admitted in the other 3 traffic classes will not be able to utilized the bandwidth allocated to the Provisioned class, even when that bandwidth idle.

By definition, if this service is offered in the RPR network, the network operator must set the provisioned bandwidth allocation consistently amongst all stations on the ring and then set allocation the provisioned bandwidth amongst nodes in the network. How this badwidth allocation is performed is the responsibility of the MAC client and beyond the scope of this specification.

The transmit data requests to the provisioned service access point will be policed by the RPR MAC to meet the statically provisioned rate for this service. The service access point will also provide an indication to the MAC client of the status of the underlying channel. This information includes whether the service is currently operative (up or down).

The bandwidth allocated to the provisioned service may be set to zero allowing network designers the option of not offering such a service.

5.2.2 High Priority Service

The MAC provides a high priority delivery service. This service is intended to support application which require bounded end-to-end delay and jitter specifications.

The MAC assumes that traffic requesting high priority service will be shaped at ingress to meet provisioned values for CIR, BIR and EIR by the MAC client. The MAC sublayer will implement a policing function as part of the high priority service to ensure that provisioned service parameters are not violated.

The high priority service is an engineered service and must be provisioned by the network designer.

The service access point also provides an indication to the MAC client of the status of the underlying channel. This information includes whether the service is currently operative (up or down) and whether there is dynamic backpressure from the media to indicate that traffic cannot currently be accepted.

5.2.3 Medium Priority Service

The Medium priority service is provided to implement a traffic class for applications which are not delay sensitive but which require BW guarantees.

It is similar in implementation to the High Priority service in that it expects the client to provide shaped ingress traffic stream that conforms to provisioned CIR and EIR limits. However traffic is treated differently with respect to the RPR-fa depending on whether it meets its CIR/EIR profile.

Those frames determined to be in-profile will be marked as such with a bit in the RPR header at ingress to the ring. These frames will not be counted as part of the RPR fairness algorithm at ingress to the ring or when transitting through stations on the ring. They are essentially invisible to the fairness algorithm in the same way that high priority packets are.

Those frames determined to be out-of-profile will be marked as such with a bit in the RPR header at ingress to the ring. Out-of profile frames will be counted as part of the RPR fairness algorithm both at ingress to the ring and while transitting stations on the ring.

Regardless of whether the frame is marked in- or out-of-profile, the frames are still sent on the low priority transit channel on the ring.

For in-profile traffic, applications using the medium priority service will not be blocked by ingress rate pro-licing from the RPR fairness algorithm (like the high priority service) but will still incur the increased end-to-end delay of the low priority transit channel.

For out-of-profile traffic, applications using the medimu priority service will receive the same service level as if they had used the low priority service access point.

The service access point also provides an indication to the MAC client of the status of the underlying chan-nel. This information includes whether the service is currently operative (up or down) and whether there is dynamic backpressure from the media to indicate that traffic cannot currently be accepted.

5.2.4 Low Priority Service

The Low Priority service is provided to implement a Best Effort Traffic Class (BETC). It is transmitted on the MAC Low Priority Transit Path and is not sensitive to end-to-end delay or jitter.

The service access point also provides an indication to the MAC client of the status of the underlying chan-nel. This information includes whether the service is currently operative (up or down) and whether there is dynamic backpressure from the media to indicate that traffic cannot currently be accepted.

5.3 MAC Peer-to-Peer Services

Since 802.17 MAC network is a ring based, shared media network, each MAC has a transit service to frames that are not destined or sourced from the MAC client, hosted by that particu.lar MAC. This traffic passes through the MAC sublayer on one of three channels: high priority, low priority or controlth

5.3.1 High Priority Transit Channel

The MAC implements a high priority transit channel to support the High priority traffic services. The high priority transit channel provides a worst-case per-station transit delay of one frame-time in order to bound the maximum delay for the network on the high priority service class.

The high priority transit channel does not support preemption of transmission of either the transit or ingress frames.

5.3.2 Low Priority Transit Channel

The MAC implements a low priority transit channel to support both medium and low priority service classes. All low priority traffic and medium priority traffic travels through the Low Priority Transit Channel on the ring.

The Low Priority Channels implements a lossless service on the ring.

5.3.3 Control Channel

The MAC implements a control channel through which control messages are passed between MAC entities on the RPR network. Traffic for this channel has the highest priority in terms of scheduling at ingress to the ring and does not participate in nor is policed by the RPR fairness mechanism.

Control traffic on the ring transit path is treated the same as traffic in the High Priority channel for the purposes of transit path scheduling and ring access (i.e. it has priority over all ingress traffic).

The MAC provides no policing of traffic destined for this channel except for inter-station synchronization purposes.

5.4 MAC services to the MAC Client Layer

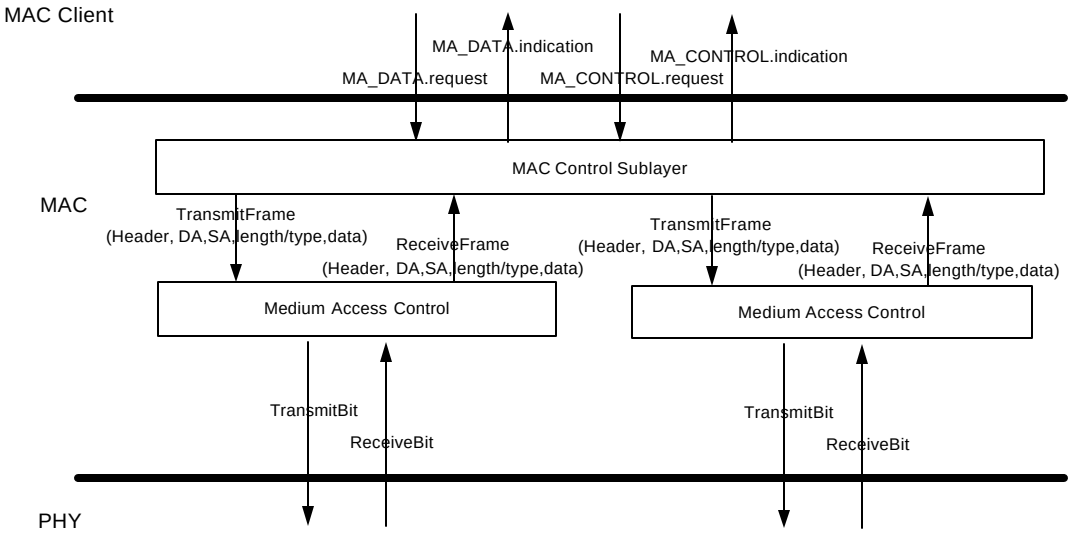


Figure 7—MAC Service Model

5.4.1 Overview of the interactions

Four service primitives are defined for the LLC interfaces.

- MA_DATA.request
- MA_DATA.indication.
- MA_CONTROL.request (used by MAC Control sublayer).
- MA_CONTROL.indication (used by MAC Control sublayer).

The formats for the M_DATA.indications and M_DATA.requests are the same as formats for MA_DATA.indication and MA_DATA.requests, except for the addition of an optional parameter for the FCS. This parameter may be used to preserve the FCS when bridging between LANs using like formats.

5.4.2 Basic services and options

The MA_DATA.request, MA_DATA.indication service, MA_CONTROL.request and MA_CONTROL.indication primitives described in this subclause are mandatory.

5.4.3 Detailed service specification

5.4.3.1 MA_DATA.request

5.4.3.1.1 Function

This primitive defines the transfer of data from a MAC client entity to a single peer entity or multiple peer entities in the case of group addresses.

5.4.3.1.2 Semantics of the service primitive

The semantics of the primitives are as follows:

```
MA_DATA.request (header,
         destination_address,
         source_address,
         m_sdu,
         service_class,
         ringlet_id)
```

The header parameter may specify one or the other ring medium, priority, Time To Live (TTL), and unicast or multicast. The destination_address parameter may specify either an individual or a group MAC entity address. It must contain sufficient information to create the DA field that is pre-appended to the frame by the local MAC sub- layer entity and any physical information. The source_address parameter, if present, must specify an individual MAC address. If the source_address parameter is omitted, the local MAC sublayer entity will insert a value associated with that entity. The m_sdu parameter specifies the MAC service data unit to be transmitted by the MAC sublayer entity. There is sufficient information associated with m_sdu for the MAC sublayer entity to determine the length of the data unit. The service_class parameter indicates a quality of service requested by the MAC client The ringlet_id parameter, if present, allows the MAC client to optionally specify the desired ring on which to transmit the m_sdu. The MAC will obey this request except when the ringlet status shows that the it is down for a protection event.

5.4.3.1.3 When generated

This primitive is generated by the MAC client entity whenever data shall be transferred to a peer entity or entities. This can be in response to a request from higher protocol layers or from data generated internally to the MAC client, such as required by Type 2 LLC service.

5.4.3.1.4 Effect of receipt

The receipt of this primitive will cause the MAC entity to insert all MAC specific fields, including header, DA,SA, and any fields that are unique to the particular media access method, and pass the properly formed frame to the lower protocol layers for transfer to the peer MAC sublayer entity or entities.

5.4.3.1.5 Additional comments

The RPR MAC protocol provides four qualities of service in service_class requested.

5.4.3.2 MA_DATA.indication

5.4.3.2.1 Function

This primitive defines the transfer of data from the MAC sublayer entity (through the MAC Control sub-layer) to the MAC client entity or entities in the case of group addresses.

5.4.3.2.2 Semantics of the service primitive

The semantics of the primitive are as follows:

```
MA_DATA.indication (header,
         destination_address,
         m_sdu,
         ringlet_id,
         reception_status)
```

The header parameter may specify one or the other ring medium, priority, Time To Live (TTL), and unicast or multicast. The destination_address parameter may be either an individual or a group address as specified by the DA field of the incoming frame. The source_address parameter is an individual address as specified by the SA field of the incoming frame. The m_sdu parameter specifies the MAC service data unit as received by the local MAC entity. The reception_status parameter is used to pass status information to the MAC client entity. The ringlet_id parameter indicates, to MAC clients which optionally use the information, which ringlet the m_sdu was received from.

5.4.3.2.3 When generated

The MA_DATA.indication is passed from the MAC sublayer entity (through the MAC Control sub-layer) to the MAC client entity or entities to indicate the arrival of a frame to the local MAC sublayer entity that is destined for the MAC client. Such frames are reported only if they are validly formed, received without error, and their destination address designates the local MAC entity.Frames destined for the MAC Control sublayer are not passed to the MAC client if the MAC Control sublayer is implemented.

5.4.3.2.4 Effect of receipt

The effect of receipt of this primitive by the MAC client is unspecified.

5.4.3.2.5 Additional comments

If the local MAC sublayer entity is designated by the destination_address parameter of an MA_DATA.request, the indication primitive will also be in oked by the MAC entity to the MAC client entity. This characteristic of the MAC sublayer may be due to unique functionality within the MAC sublayer or characteristics of the lower layers (for example, all frames transmitted to the broadcast address will invoke MA_DATA.indication at all stations in the network including the station that generated the request).

5.4.3.3 MA_CONTROL.request

This primitive defines the transfer of control requests from the MAC client to the MAC Control sublayer.

5.4.3.3.1 Function

This primitive defines the transfer of control commands from a MAC client entity to the local MAC Control sublayer entity.

5.4.3.3.2 Semantics of the service primitive

The semantics of the primitive are as follows:

```
MA_CONTROL.request (header
         destination_address,
```

```
        opcode,
        request_operand_list)
```

The destination_address parameter may specify either an individual or a group MAC entity address. It must contain sufficient information to create the DA field that is preappended to the frame by the local MAC sublayer entity. The opcode specifies the control operation requested by the MAC client entity. The request_operand_list is an opcode-specific set of parameters. The valid opcode and their respective meanings are described in Table 1— on page47.

Table 1—Control Request Opcodes

Opcode	Operand	Meaning
0x00	none	No Request
0x01	none	Request Network Topology
0x02	Service_Class	Request Service Status
0x03	Station_MAC_Address	Request Station Configuration
0x04	Station_MAC_Address	Request Transit Path Congestion Status
0x05	Station_MAC_Address	Request Current Topology Database
0x06	none	Pause Message
0x07-0xFF	TBD	TBD

5.4.3.3.3 When generated

This primitive is generated by a MAC client whenever it wishes to use the services of the MAC Control sublayer entity.

5.4.3.3.4 Effect of receipt

The effect of receipt of this primitive by the MAC Control sublayer is opcode-specific.(See Clause TBD.)

5.4.3.4 MA_CONTROL.indication

5.4.3.4.1 Function

This primitive defines the transfer of control status indications from the MAC Control sublayer to the MAC client.

5.4.3.4.2 Semantics of the service primitive

The semantics of the primitive are as follows:

```
MA_CONTROL.indication(header,
        opcode,
        indication_operand_list)
```

The elements of the indication_operand_list are opcode-specific, and specified in Table2.

Table 2—Control Indication Opcodes

Opcode	Operand	Meaning
0x01	Network Topology Data Structure	Network Topology Change
0x02	Service_Class, Status (ok_to_send, do_not_send)	Service Status Change
0x03	configuration_parameter_list	Request Station Configuration
0x04	normalized_bandwidth_value	Request Transit Path Congestion Status
0x05-0xFF	TBD	TBD

5.4.3.4.3 When generated

The MA_CONTROL.indication is generated by the MAC Control sublayer under conditions specific to each MAC Control operation.

5.4.3.4.4 Effect of receipt

The effect of receipt of this primitive by the MAC client is unspecified.

5.5 Physical layer service access point (PSAP) definition

5.5.1 PSAP

The IEEE 802.17 MAC supports the following Physical Service Access Point (PSAP) primitives:

— PHY_LINK_STATUS.indication
— PHY_DATA.request
— PHY_DATA.indication
— PHY_DATA_VALID.indication
— PHY_READY.indication

5.5.1.1 PHY_LINK_STATUS.indication

This interface provides a means to indicate some the status of the physical link to the MLME. The setting of this MLME MIB attribute can cause the MLME to perform an action and/or request that an action be performed by the MAC.

The semantics of the primitive are as follows:

```
PHY_LINK_STATUS.indication (
          link_status
          )
```

The link_status parameter may have the values of OK, DEGRADE, or FAIL. All physical layers shall support OK and FAIL. Support for generation of DEGRADE is optional.

5.5.1.2 PHY_DATA.request

This interface defines the transfer of an octet of data from the MAC to the RS.

The semantics of the primitive are as follows:

```
PHY_DATA.request (
         output_unit
         )
```

The output_unit parameter contains an octet_of_data, or DATA_COMPLETE.

5.5.1.3 PHY_DATA.indication

This interface defines the transfer of an octet of data from the RS to the MAC.

The semantics of the primitive are as follows:

```
PHY_DATA.indication (
         input_unit
         )
```

The input_unit parameter contains an octet_of_data.

5.5.1.4 PHY_DATA_VALID.indication

This interface indicates whether the parameter of PHY_DATA.indicate contains valid data.

The semantics of the primitive are as follows:

```
PHY_DATA_VALID.indication (
         data_valid_status
         )
```

The data_valid_status parameter may have the values of VALID or NOT_VALID.

5.5.1.5 PHY_READY.indication

This interface indicates whether the PHY is ready to accept a new MAC frame.

The semantics of the primitive are as follows:

```
PHY_READY.indication (
         ready_status
         )
```

The ready_status parameter may have the values of READY or NOT_READY.

5.6 Management layer service access point

5.6.1 MLSAP

The IEEE 802.17 MAC supports the following Management Layer Service Access Point (MLSAP) primitives:

— MLME_GET.request
— MLME_SET.request

5.6.1.1 MLME_GET.request

This primitive requests the value of an attribute of the MLME MIB from the MAC.

The semantics of the primitive are as follows:

```
MLME_GET.request (
          mib_attribute
          )
```

The mib_attribute parameter contains an attribute of the MLME MIB.

5.6.1.2 MLME_SET.request

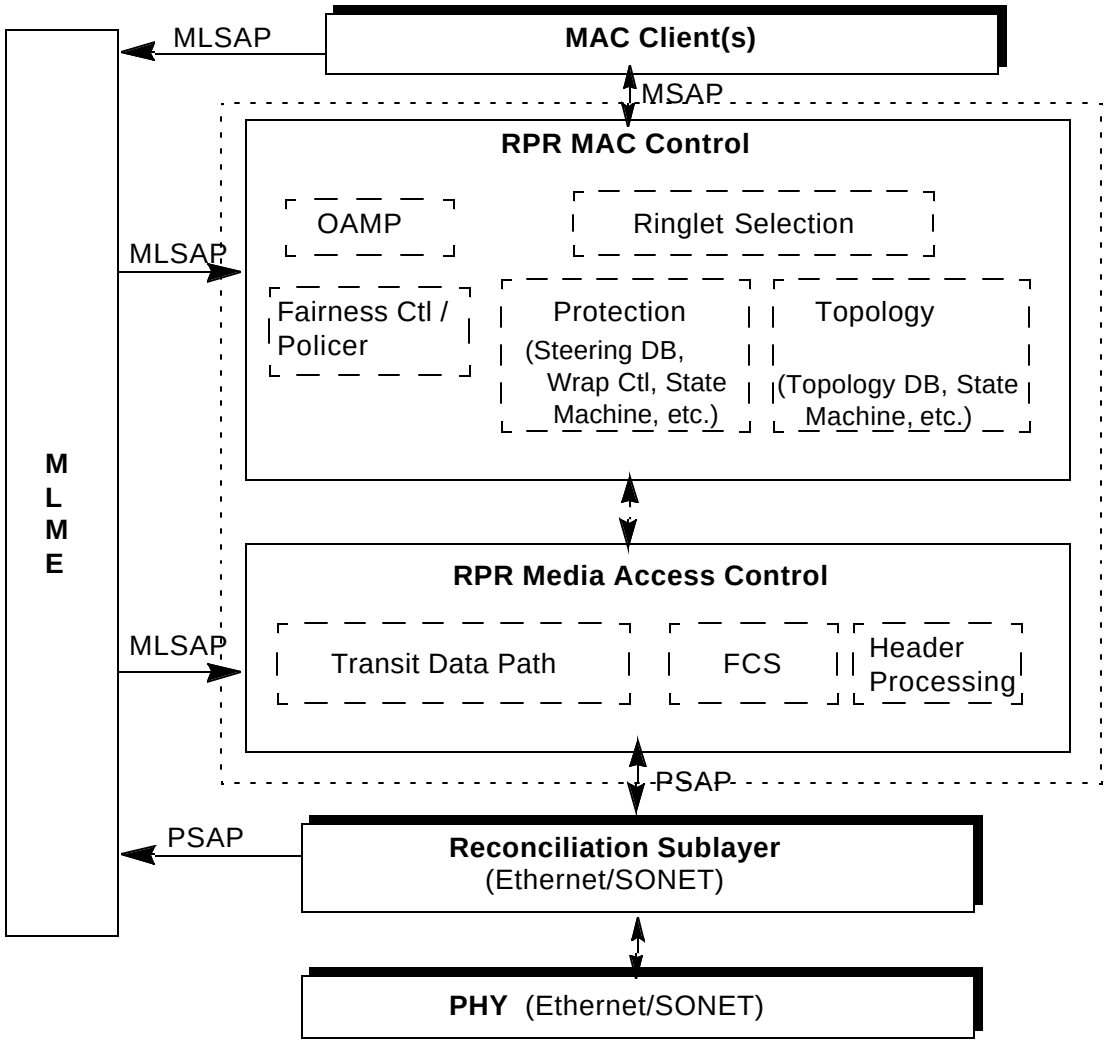
This primitive requests that an attribute of the MLME MIB, residing within the MAC, be set to a given value.

The semantics of the primitive are as follows:

```
MLME_SET.request (
          mib_attribute,
          mib_attribute_value
          )
```

The mib_attribute parameter contains an attribute of the MLME MIB. The mib_attribute_value contains the requested value for mib_attribute. If the given mib_attribute is associated with a specific action, then the action shall be performed by the MAC.

5.7 MAC Reference Model



5.7.1 PHY

The 1Gb/s and 10 Gb/s Ethernet Physical Layers and GFP SONET Physical are referenced, Byte Synchronous HDLC is described in relationship to the referenced standard. It is understood that different Physical Layers will provide different services. The differences are reconciled by each PHY specific Reconciliation Sublayer.

5.7.2 Reconciliation Sublayer

The Reconciliation Sublayer is part of the Physical Layer, and provides a uniform, reconciled service interface to the MAC Layer. There is one Reconciliation Sublayer Entity for each Physical Layer interface. In addition to providing a common interface for MAC/PHY data, the Reconciliation Sublayer also presents a common interface for management and control of the PHY that is needed by the MAC.

5.7.3 RPR Media Access Control

The RPR Media Access Control sublayer of the MAC Layer provides the access control for the Physical Layer. It also provides the control of the transit path through the MAC.

5.7.4 RPR MAC Control

The RPR MAC Control sublayer of the MAC Layer provides control of the RPR MAC, including Rate Control, Policing, Protection, Topology, and Link Aggregation.

5.7.4.1 Rate Control/Policer

The Rate Control and Policer function of the MAC Control sublayer provides control of the rate at which frames are transmitted by the MAC, and coordination of this control with the other MAC Control sublayers on the ring.

5.7.4.2 Protection

The Protection function of the MAC Control sublayer provides control of directing frames to the appropriate ringlet based on the protection database. It also provides the protection state machine for the local MAC, and coordination of this control with the other MAC Control sublayers on the ring.

5.7.4.3 Topology

The Topology function of the MAC Control sublayer manages the topology database. It also provides the topology state machine for the local MAC, and coordination of this control with the other MAC Control sublayers on the ring.

5.7.4.4 Ringlet Selection

The ringlet selection is optionally sepecified by trusted client. Otherwise, it is the responsibilty of MAC to decide the ringlet based on the current ring status for a particular destination address.

5.7.5 MAC Layer Management Entity

The MAC Layer Management Entity is an independent entity that resides outside of the MAC Layer in a separate management plane. The MLME contains the Management Information Base for the MAC Layer, and provides Get and Set operations on the MIB to MLME SAP user-entities. The MLME also provides actions upon the MAC Layer as a result of the invocation of Set.request primitives.

6. Media Access Control data path

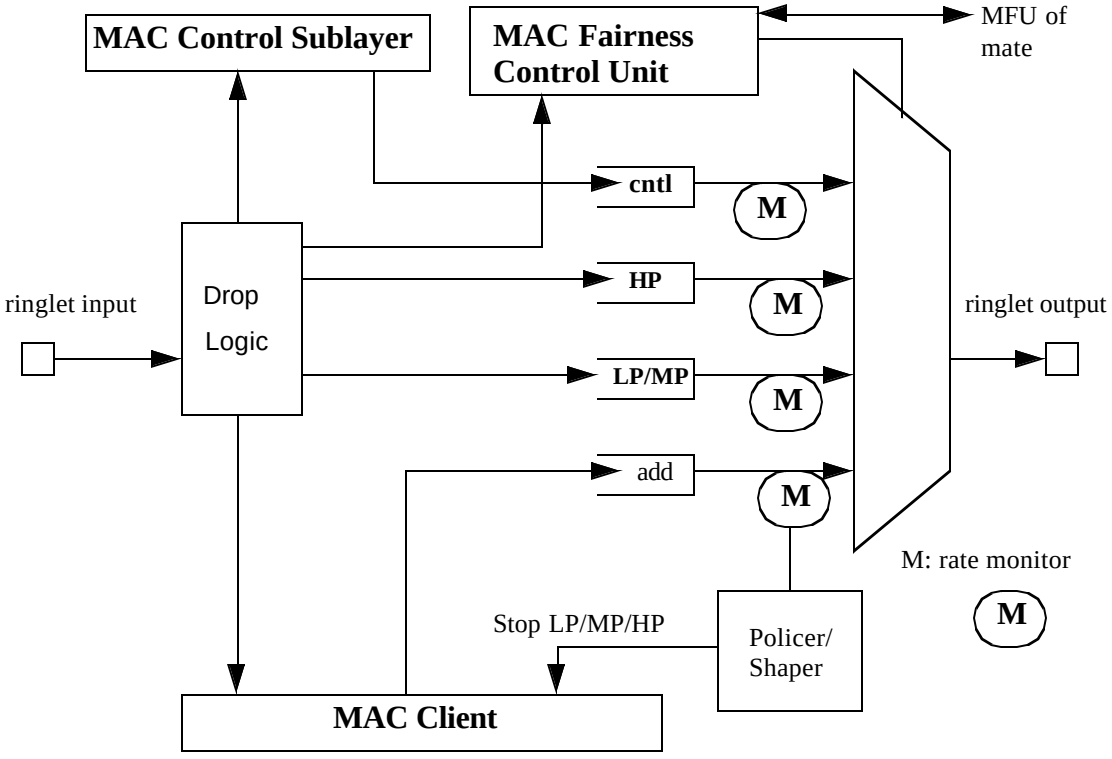


Figure 8—MAC data path

6.1 Transit/Transmit 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 Figure9 below. There are two possible implementations, using either one or two transit buffers. The minimum requirement is the use of the high priority transit buffer, the inclusion of the second transit buffer is optional. In the two transit buffer case, traffic will be separated into three priorities, High, Medium and Low. High priority will be placed into one fifo queue, and Medium and Low priority will utilize another.In the single transit buffer case, only one buffer will be used for all three priorities.

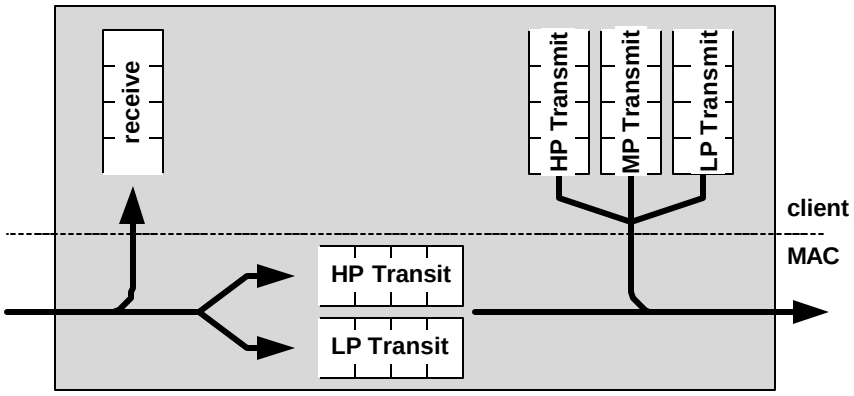


Figure 9—Transit / Transmit Buffer Design

6.2 Transmit and forwarding operation

A RPR MAC can transmit data packets from five possible queues:

1) High priority packets from the high priority transit buffer.
2) Medium or low priority packets from the low priority transit buffer
3) High priority packets from the client Tx high priority fifo.
4) Medium priority packets from the client Tx medium priority fifo.
5) Low priority packets from the client Tx low priority fifo.

Note that Medium priority traffic is assigned a Committed Information Rate (CIR.) Traffic within the CIR is treated as if it is high priority traffic while it is being accepted to the ring. Traffic above the CIR will be treated as low priority, and will be referred to as EIR MP traffic, or eMP. The RPR MAC will decide which traffic to send based on a priority scheme, which will differ between single and dual transit buffer implementations.

6.2.1 Single buffer implementation

In a single transit buffer implementation, transit traffic is always sent first, regardless of priority. High, medium and low priority transmit traffic will then be sent in priority order. All three classes of transmit traffic will be subjected to rate shapers. LP and eMP transmit traffic will also be limited to the fair rate governed by the RPR-fa rules.

6.2.2 Dual buffer implementation

In a dual transit buffer implementation, high priority forwarded data always gets sent first. High priority transmit data may be sent as long as the Low Priority Transit Buffer (LPTB) is not almost full. Medium Priority transmit traffic within CIR is scheduled just after high priority transmit packet, and can be sent as long as the LPTB is not almost full.

EIR medium priority transmit traffic resided in medium priority transmit queue and low priority transmit traffic are can be sent next, assuming their combined rate does not exceed the fair rate governed by the RPR-fa rules, and the LPTB has not exceeded a low priority threshold. There is an dd_rate_ok equation to check

if EIR medium transmit/low priority is allowed to add traffic into ring. If the check is failed, the LPTB will be serve instead to avoid the starvation problem for LPTB traffic.

All three types of transmit traffic are also subjected to rate shapers.

If nothing else can be sent, low priority packets from the low priority transit buffer are sent.This decision tree is shown in Figure10

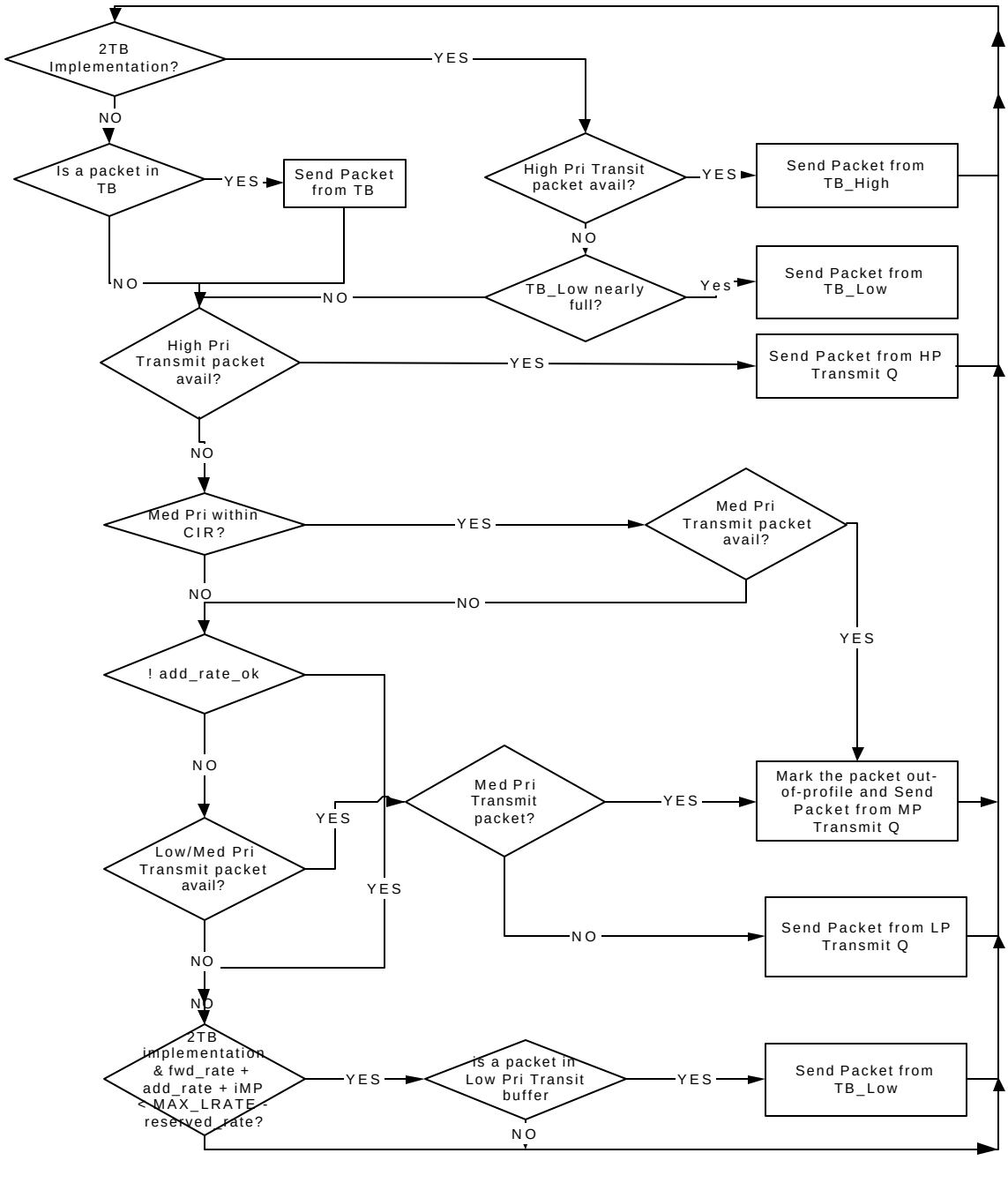


Figure 10—RPR transmit flowchart

6.3 Receive operation

Receive Packets entering a node are copied to the receive buffer if a Destination Address (DA) match is made. 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 placed into

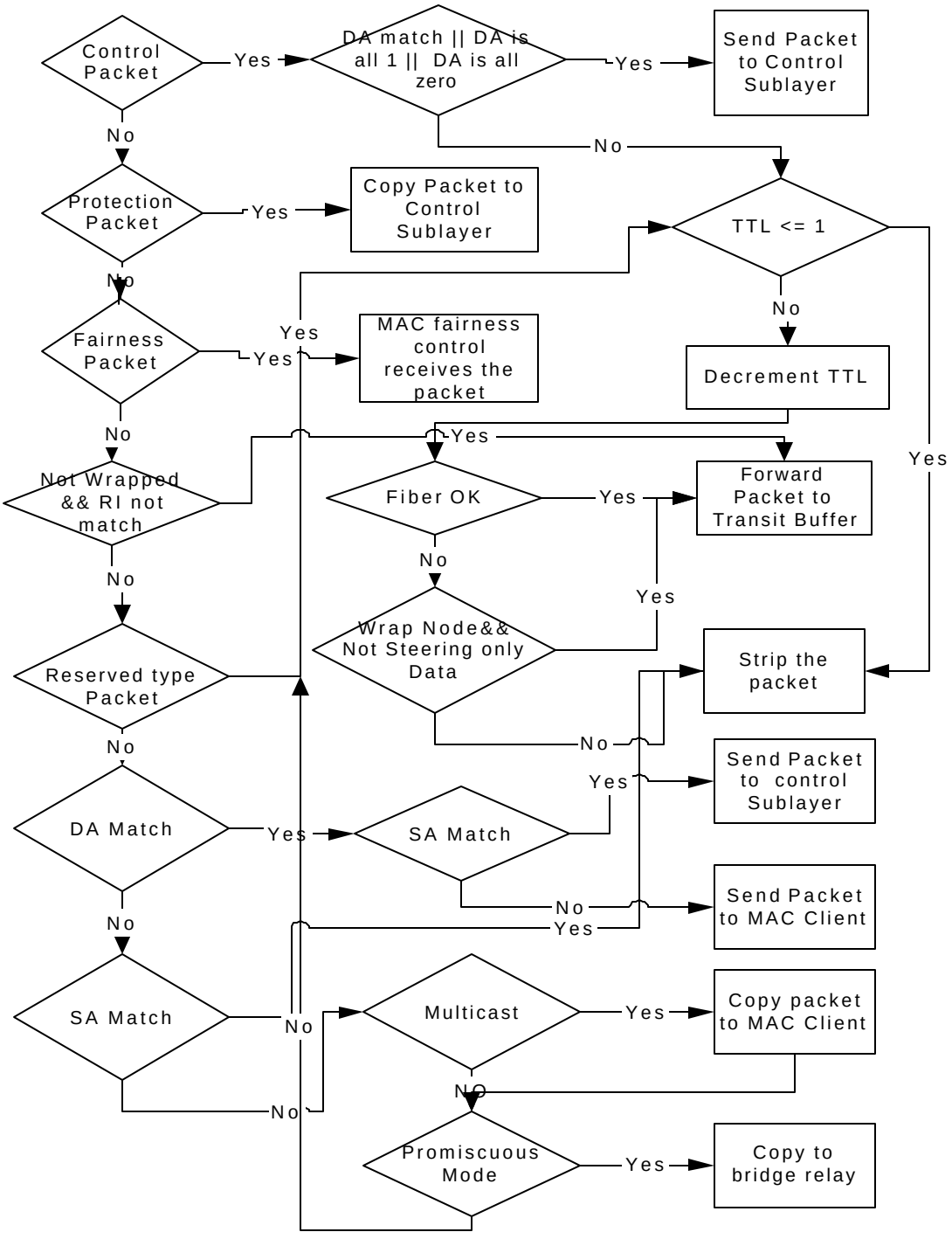


Figure 11—RPR Receive Flowchart

the Transit Buffer (TB) for forwarding to the next node if the packet passes Time To Live and Header Error Check (HEC) tests.

6.4 Transit operation

A series of decisions based on the type of packet (mode), source and destination addresses are made on the MAC incoming packets. Packets can either be control or data packets. Protection messages are broadcast to all nodes on the ring. Control packets may be hop-by-hop, broadcast or unicast destined to a station. Hop-by-hop is encapsulated with a all-zero destination address. Broadcast control packet is with a all-one destination address. Control packets are stripped once the information is extracted. The source and destination addresses are checked in the case of data packets. The rules for reception and stripping are given below as well as in the flow chart in Figure11.

1) Strip unicast packets at the destination station.
2) 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.
3) Do not process packets other than for TTL and forwarding if the type is not supported by the node (e.g. reserved types).
4) Transit packets will be discarded if there is a HEC error.
5) Transit packet will be optionally discarded if there is a FCS error.
6) Packets accepted by the host due to destination address match may be optionally discarded at the MAC if there is an FCS error.
7) Type 4 protection messages are broadcast and should always be copied to the MAC control sublayer.
8) Type 5 control messages are accepted and stripped if there is a DA match, a all-one DA, or a all-zero DA.
9) 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.
10) Steering only data will be stripped in the wrapped node.
11) 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.

Notes:

i) FCS error packets should be passed to the client when there is a DA match. It is the client's responsibility whether to accept or drop the error packet. The CRC will be overwrited with the newly invert CRC to indicate the FCS error.
ii) Conditionals (if statements) in Figure 11 branch to the right if true and branch down if false.

6.4.1 Transit operation in a Bridge (Promiscuous Mode)

When the RPR MAC is part of a bridge, all data packets 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 performed 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. Data should continue to be sorted into high and low priority transit buffers with high priority transit buffers always emptied first. 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 passthru. 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.

7. MAC client interface

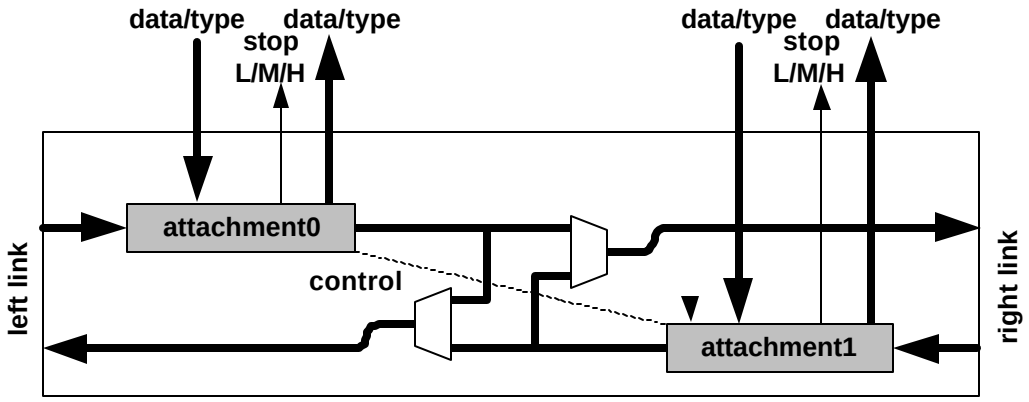


Figure 12—MAC client interface

7.1 Topology awareness

In its canonical form, an 802.17 network consists of dual, counter-rotating ringlets. As such the MAC may provide two different views of the network to the client entity.

The MAC can optionally present two views of the network to the MAC client. A flat view of the network, in which the MAC sublayer hides the dual-ring-based topology from the client, or a topology-aware view in which allows the MAC client to make data and control requests for specific ringlets. Topological information is collected via a MAC sublayer entity process known as topology discovery and is made available to the client via a request to the Layer Management Entity, MAC control indicate.

7.2 Traffic policing

A RPR MAC will have processes to policy access to each service by the MAC client in order to ensure that Media access and bandwidth provisioning rules are obeyed. If the MAC client chooses to disregard feedback from the MAC on service availability and issue a DATA.req, the MAC will accept the request but will return an indication to that effect. The MAC will not transmit the packet until the packet is allowed to be transmitted.

There is one shaper each for HP, MP, and LP traffic. The shapers are simple token buckets, and if a bucket become empty or negative, the RPR MAC communicates with the MAC client on 3 pins: STOP_HIGH, STOP_MED and STOP_LOW. If the client ignores the pins and sends the traffic anyway, the RPR MAC will not schedule the client until the token bucket has a token in it. The detail shaper function will be described in 11.2.

7.3 Multi-Choke client

Additional information on bandwidth utilization for the transit path links through each node on the ring is made available via request to the Layer Management Entity. The information provided in response may be used by the MAC client implementation to detect congestion at a particular node in order to implement a

scheme of Virtual Destination Queueing (VDQ) to avoid head-of-line blocking of PDUs destined to nodes that are not physically situated beyond the point of congestion.

Multi-Choke implementation of RPR-fa uses the fairness algorithm which will be described in Clause 12. The difference is that its client keeps track of up to the number of node congestion locations on the ring(choke points), and uses this information to increase ring utilization and spa tial reuse.

Multi-Choke requires access to topology information as well as per destination queuing in the MAC client. This will allow the MAC client to determine which destinations are located before the first choke point, which are between the first 2 choke points, etc.

In order for client to take advantage of the available bandwidth on the ring, it may keep counters for each destination queue similar to standard RPR-fa counters. When a decision needs to be made for a choke point, the total of the add_rate values for VDQ queues after the choke point are used as a total_add_rate value. This value is normalized and compared to the allow_rate value associated with the choke point, as in the basic RPR-fa algorithm. To determine if a source node can send to a destination node, this calculation must be done (and satisfied) for each choke point until the destination. Also the total rate of all queues must be compared against max_rate. The responsibility of the MAC is to enforce the fairness for the most congested span that the node is contending for. This is done by policing function described in clause 11.1

As an example, imagine that there are 3 choke points at nodes 2, 4, and 6. Node 1 wants to send to node 5. The algorithm will check the total add_rate for destination nodes 3, 4, 5,... against the allow_rate for the choke point at 2. If that test passes, it will check the total rate for destination nodes 5, 6,... against the allow_rate value for the choke point at node 4. Then a final check of the total rate for all nodes against max_rate will be performed. If all tests pass then node 1 is allowed to send to node 5.

The benefit is obvious. Using a single queue in the client may cause head of line blocking and all traffic may be limited to allow_rate. If a node is trying to send to its neighbor, and the congestion point is a few nodes away, this traffic may be penalized without reason. With Multi-Choke implementation of the RPR-fa, though, this traffic may continue, and the link utilization before the choke point may increase

8. MAC Physical Interface

RPR is media independent.RPR Frame will be allowed to send over different physical media.

8.1 SONET/SDH network

RPR may also connect to a SONET/SDH ring network via a tributary connection to a SONET/SDH ADM (Add Drop Multiplexor). The two RPR rings may be mapped into two STS-Nc connections. SONET/SDH networks typically provide fully redundant connections, so RPR mapped into two STS-Nc connections will have two levels of protection. The SONET/SDH network provides layer 1 protection, and RPR provides layer 2 protection. In this case it is recommended to hold off the RPR Signal Fail protection message triggers (which correspond to failures which can be protected by SONET/SDH) for about a programable hold-off timer in order to allow the SONET/SDH network to protect. Only if a failure persists for over the hold-off timer(indicating SONET/SDH protection failure) should the protection switch protocol take place.

Since multiple protection levels over the same physical infrastructure are not very desirable, an alternate way of connecting RPR over a SONET/SDH network is configuring SONET/SDH without protection. Since the connection is unprotected at layer 1, RPR would be the sole protection mechanism.

Hybrid RPR rings may also be built where some parts of the ring traverse over a SONET/SDH network while other parts do not.

Connections to a SONET/SDH network would have to be synchronized to network timing by some means. This can be accomplished by locking the transmit connection to the frequency of the receive connection (called loop timing) or via an external synchronization technique.

Connections made via dark fiber or over a WDM optical network should utilize internal timing as clock synchronization is not necessary in this case.

8.1.1 Byte Synchronous HDLC framing

Flag delimiting on SONET/SDH uses the octet stuffing method defined for Byte Synchronous HDLC. The packet delimiter flags (0x7E) are required for SONET/SDH links but may not be necessary for RPR on other media types. An End-of-Packet is delineated by a flag, which might also be the next packet's starting flag. If the data appears to be a flag (0x7E) or an escape character (0x7D) anywhere inside of a packet, the data must be marked with an escape character.

SONET/SDH framing plus Byte Synchronous HDLC packet delimiting allows RPR to be used directly over fiber or through an optical network (including WDM equipment).

8.1.2 GFP framing

Generic Framing Procedure (GFP) is a standard method of delineating octets of variable length payload into octet-synchronous payload envelope as defined in ANSI T1.105.02 and ITU-T G.707, G.709. GFP defines a frame format for protocol data units (PDUs) transferred between GFP initiation and termination points, as well as the mapping procedure for client signals into GFP.

The GFP delineation scheme uses a Header Error Check (HEC) polynomial. The latest GFP standard can be found in ITU specification G.7041.

The preferred method for GFP to encapsulate 802.17 frames is to use the Null header definition as defined

by the Extension Header Identifier (EXI) with a User Payload Identifier (UPI) for RPR payload.

The SONET/SDH Reconciliation Sublayer (RS) and the MAC sublayer provide an optional capability to propagate the PDU length value to optimize the forwarding of PDU from one SAP to another.

GFP allows RPR packets to be transported over SONET/SDH and OTN networks.

8.2 Ethernet

RPR frames can be sent over Ethernet physical media.

9. PHY RS

10. Frame formats

10.1 Overview

This clause defines in detail the frame structure for data communication systems using the RPR MAC. It defines the syntax and semantics of the various components of the MAC frame.

Four frame formats are specified in this clause:

a) A data RPR frame format,
b) A control RPR frame format, and
c) An extension of the basic RPR frame format for Tagged RPR frames, i.e., frames that carry QTag
d) An extension of the basic RPR frame format for CID RPR frames, i.e., frames that carry CID

This section describes the frame formats used by RPR. Packets can be sent over any point to point link layer (e.g. SONET/SDH, point to point ETHERNET connections). The maximum transfer length (MTU) is 9216 octets.

These limits include everything listed in Figure13 but are exclusive of the frame delineation (e.g. for RPR over SONET/SDH, the flags used for frame delineation are not included in the size limits).

The following frame format does not include any layer 1 frame delineation. For RPR over POS, there will be an additional flag that delineates start and end of frame.

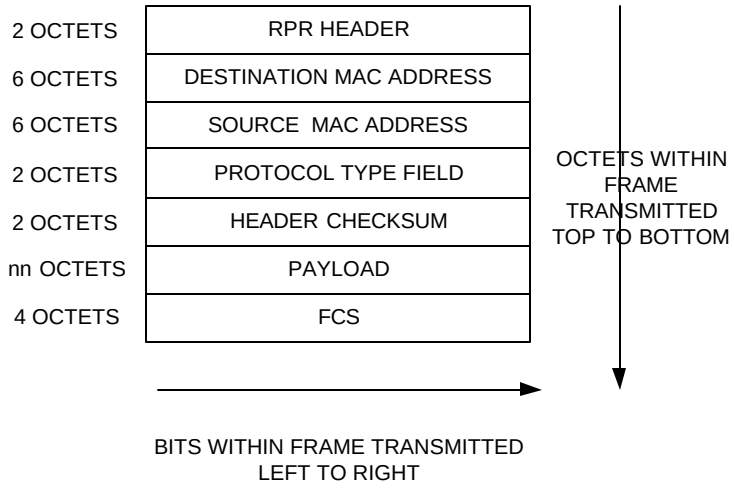


Figure 13—RPR Frame Format

10.2 RPR packet header format

Each packet has a fixed-sized header. The packet header format is shown in Figure14

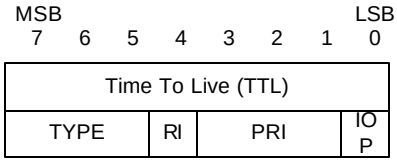


Figure 14—RPR Packet Header Format

The fields are described below:

10.2.1 Time To Live (TTL)

This 8 bit field is a hop-count that must be decremented every time a node forwards a packet. If the TTL reaches zero it is stripped off the ring. While the packet is traveled in wrapped span due to failure protection, the TTL will not be decremented hop by hop. This allows for a total node space of 256 nodes on a ring.

10.2.2 Type field

This three bit field is used to identify the mode of the packet. The following modes are defined in Table3:.

Table 3—Type Values

Value (bin)	Description
000	Reserved
001	Reserved
010	Reserved
011	Steering only data
100	Protection Control packet
101	Control packet
110	Fairness packet
111	Data packet

These modes will be further explained in later sections.

10.2.3 Ring Identifier

This bit indicates the ringlet onto which the frame was originally transmit.

Table 4—Ring Identifier

	Ring Identifier
inner ringlet	0
outer ringlet	1

10.2.4 Priority field (PRI)

This three bit field indicates the priority level of the RPR packet (0 through 7). The higher the value the higher the priority. Since there are only two queues in the transit buffer (HPTB and LPTB) a packet is treated as either low or high priority once it is on the ring. Each node determines the threshold value for determining what is considered a high priority packet and what is considered a low priority packet. In order to be consistent between nodes, only priority 7 packet as default will be queued in high priority transit buffer. The rest packets will be queued in low priority transit buffer. However, the full 8 levels of priority in the RPR header can be used prior to transmission onto the ring (transmit queues) as well as after reception from the ring (receive queues).

10.2.5 IOP

This bit could be used to mark if the packet is in or out of profiles. This bit could be used for medium priority traffic to classify if the packet is within Committed Information Rate(CIR) or not. If the bit is not set for a medium priority packet, that means the packet is with Committed Access Rate or in profile (0). Otherwise, that means the packet is excess Medium priority traffic(eMP) or out of profile(1). For those excess Medium priority traffic or out of profile(1), it will be treated as low priority traffic in RPR fairness algorithm.

Table 5—IOP value

IOP bit	Value
in profile	0
out of profile	1

10.3 Overall packet format

The overall packet format is show in Figure 13:

10.3.1 IEEE 802.17 address field

a) Each address field shall be 48 bits in length.
b) The least significant bit of the most significant octet is used to identify the destination address either as an individual or as a group address. If this bit is 0, it shall indicate that address field contains an individual address. If the bit is 1, it shall indicate that the address field contains a group address. In the source address field, the first bit is reserved and set to 0.

A MAC sublayer address is one o two types:

a) Individual Address. The address associated with a particular station on the ringlet.

b) Group Address. A multidestination address, associated with oneor more stations on the ringlet..

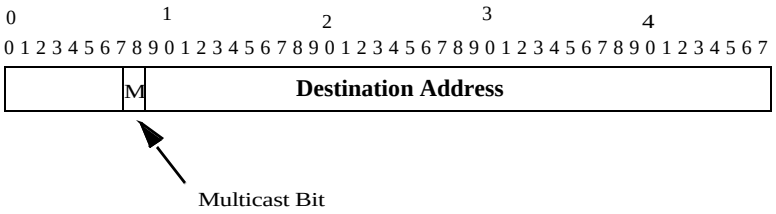


Figure 15—Group address bit position

1) Mutlticast-Group Address. An address associated by higher-level convention with a group of logically related station.
2) Broadcast Address. A distinguished, predefined multicast address that always denots the set of all stations on a given ringlet. All 1's in the destination address shall be predefined to be the Broadcast Address.

10.3.2 Destination address

The destination address is a globally unique 48 bit IEEE 802.17 address.

10.3.3 Source address

The source address is a globally unique 48 bit IEEE 802.17 address.

10.3.4 Protocol type/Length field

When the value of this field is greater than or equal to 1536 decimal (equal to 0600 hexadecimal) the Type Field indicates the nature of the MAC client protocol (Type interpretation). The length and type interpretations of this field are mutually exclusive.

Therefore when the value of the two octet field is equal to or greater than 0600 hex then it is a Type Field and the value of the Type Field is obtained from the IEEE Type Field Registrar. Some additional value will be requested for RPR frame as in Table6

Table 6: Defined protocol type

Value	Protocol Type
TBD	RPR Control
TBD	Payload with Customer Seperation ID

10.3.5 HEC field

This is a 16 bit HEC. The generator polynomial is:

HEC-16 = $x^{16} + x^{12} + x^5 + 1$

The HEC is computed over the RPR header, destination address, source address and protocol type. Single 1 bit error correction is optional.

10.3.6 FCS

The frame check sequence (FCS) is a 32-bit cyclic redundancy check (CRC) as specified in RFC-1662 and is the same CRC as used in Packet Over SONET (POS - specified in RFC-2615). The generator polynomial is:

$$CRC\text{-}32 = x^{32} + x^{26} + x^{23} + x^{22} + x^{16} + x^{12} + x^{11} + x^{10} + x^8 + x^7 + x^5 + x^4 + x^2 + x^1 + 1$$

The FCS is calculated starting from the octet following the HEC to the end of packet. The inital value for CRC calculation an all-one value.

10.4 RPR control packet format

If the type field is set to 101 then this indicates a control message. RPR control packet could be a hop-by-hop ,broadcast or point-to-point unicast message. If the control packet is a hop-by-hop message, the destination address field for control packets should be set to 0's. The TTL is not relevant but it preferably should be set to one. If the control packet is broadcast, the TTL shall be set as the number of node on the ring for broadcast control packet. The destination address shall be set as broadcast address which all is 1. The source address field for a control packet should be set to the source address of the transmitting node.

The control packet format is shown in Figure16.

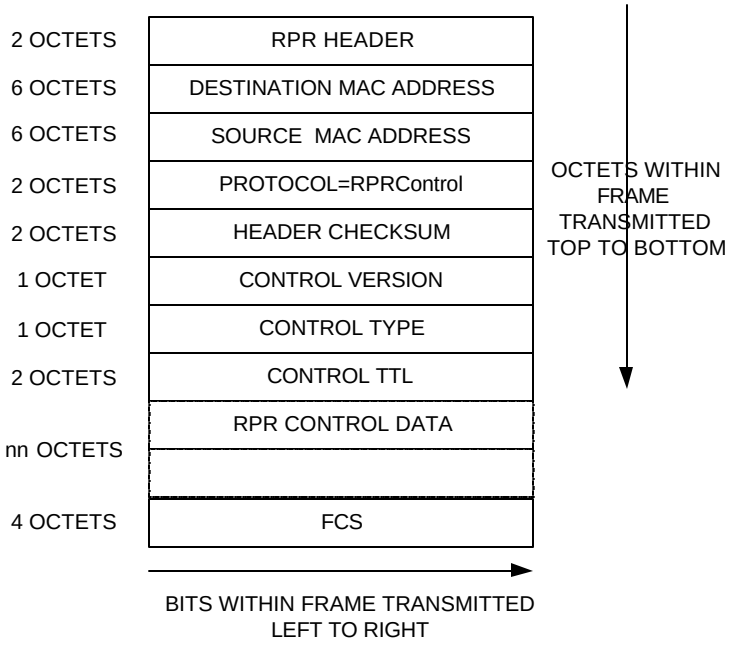


Figure 16—Control Packet Format

The priority (PRI) value in RPR header of control packet shall be set to 0x7 (all one's) except for some OAM control packet which might be set to any PRI value as indicated by the control command.

10.4.1 Control ver

This one octet field is the version number associated with the control type field. Initially, all control types will be version 0.

10.4.2 Control type

This one octet field represents the control message type. Table7 contains the currently defined control types.

Table 7—Control types

Control Type	Description
0x01	Topology Discovery
0x02	Protection message
0x03	OAM control packet
0x04 - 0xFF	Reserved

10.4.3 Control TTL

The Control TTL is a control layer hop-count that must be decremented every time a node forwards a hop-by-hop control packet. If a node receives a control packet with a control TTL <= 1, then it should accept the packet but not forward it.

10.4.4 Payload

The payload is a variable length field dependent on the control type.

10.5 RPR Fairness Frame Format

RPR Fairness Frame as shown in Figure 17 is sent to MAC neighbors to convey fairness algorithm which

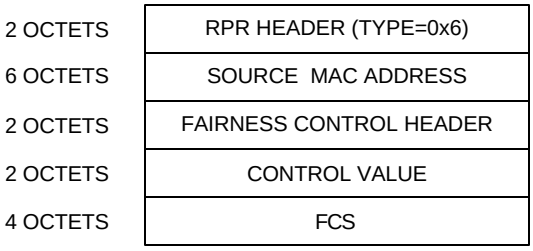


Figure 17—RPR Fairness Frame Format

will be specified in Clause 12. The IOP bit in Fairness frame will serve as a parity bit for the header.

10.6 Order of bit transmission

Each octet of RPR frame, with the exception of the FCS, is transmitted high-order bit first.

10.7 Invalid RPR frame

An invalid RPR frame shall be defined as one that meets at least one of the following conditions

a) HEC does not match with the frame is received.
b) (Optional) FCS does not match with the frame is received.

The contents of invalid RPR frames shall not be passed to the LLC or MAC control sublayers. The occurrence of invalid MAC frames may be communicated to network management.

10.8 Elements of tagged RPR frame

Tagged RPR frame format is shown as in Figure18. This format is an extension of the RPR frame format.

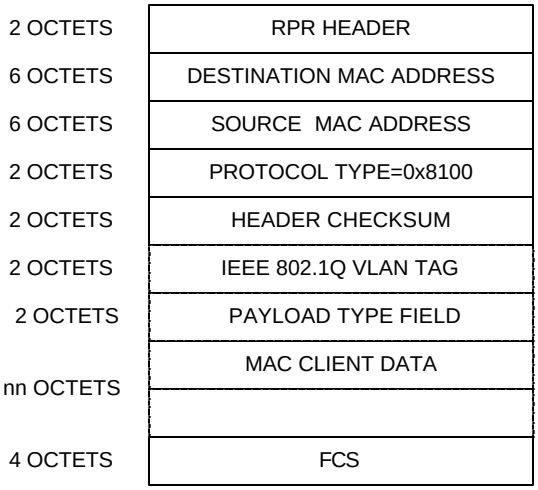


Figure 18—Tagged RPR Frame Format

10.8.1 Protocol Type/Length field

The Protocol Type/Length field of a tagged RPR frame always uses the Type interpretation, and contains the 802.1Q Tag Protocol Type: a constant equal to 0x8100.

10.8.2 Tag control information field (informative)

The Tag Control Information field is subdivided as follows in Figure19

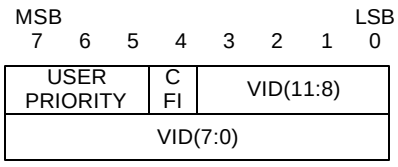


Figure 19—Tag control information field

a) A 3-bit User Priority field
b) A Canonical Format Indicator (CFI), and
c) A 12-bit VLAN Identifier

The structure and semantics within the Tag Control Information field are defined in IEEE 802.1Q.

10.8.3 Payload Type

The Payload type field followed by Tag control information field contains the original protocol type from RPR frame prior to add the QTag Prefix. The value for this field is the same as Ethernet type defined in 802.3.

10.9 Elements of RPR frame with customer seperation ID

The payload of RPR frame might have a customer separation ID (CID) as shown in Figure20.

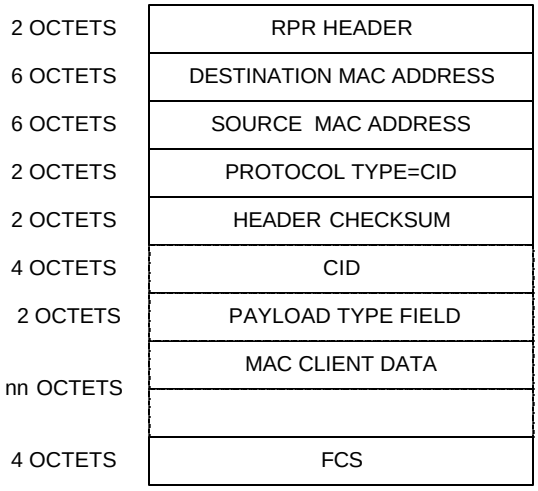


Figure 20—RPR with CID

10.9.1 Protocol Type/Length field

The Protocol Type/Length field of a CID RPR frame always uses the type interpretation, and contains a TBD value assigned by IEEE Registration Authority.

10.9.2 Customer separation ID field

This 32 bit field might be used to as an identifier of customers to separate customer traffic.

10.9.3 Payload type

This field shall be used as type interpretation for the payload. All Ethernet type value might be applied here besides some additional type value might be requested for RPR frame.

Table 8:

Value	Protocol Type
0x8100	Vlan Tagged Frame

10.9.3.1 Tagged within CID RPR frame

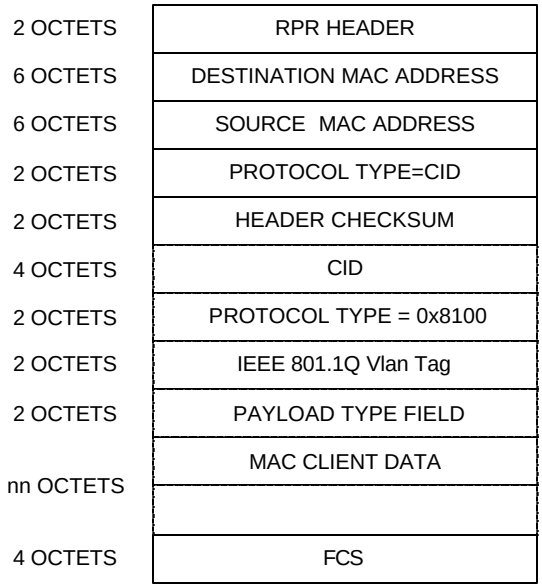


Figure 21—Tagged frame within CID RPR frame

A IEEE 802.1Q Vlan tag might be inserted after customer seperation ID as the frame in Figure21.

11. Media Access Control

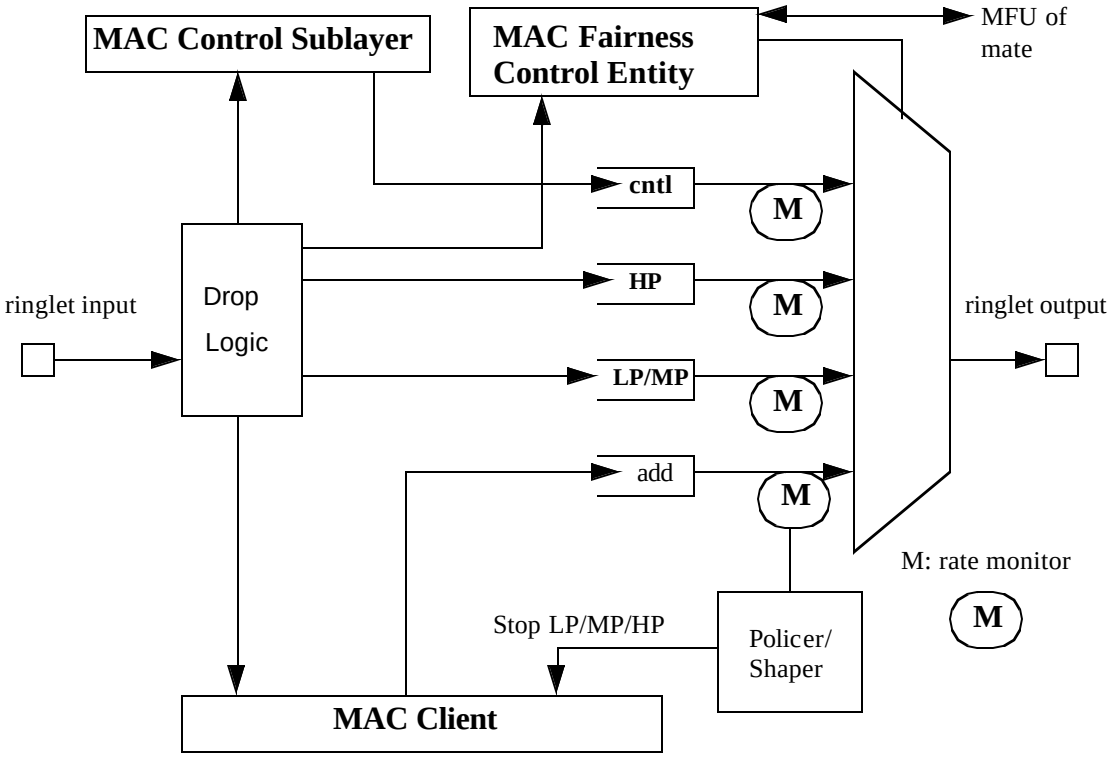


Figure 22—MAC Reference Model

11.1 Traffic policing function

RPR-fa utilizes the allow_rate_congestion and TTL_to_congestion registers and add_rate_congestion accumulator to police the add traffic to the ring. A fairness message from a choke point is always transmitted with an initial TTL of 255. Thus every station receiving the message can determine exactly the number of hops to the choke point. Allow_rate_congestion and TTL_to_congestion registers keep the values of fair rate and the distance to the choke point of the most recently received fairness packet, respectively. add_rate_congestion accumulates the number of LP and eMP bytes sent beyond the congestion point on the ring (i.e., only packets with a TTL > distance_to_choke_point and with TTL != number of nodes on the ring contribute to add_rate_congestion). Note that the client has to use a TTL value greater or equal to the number of nodes between source and the destination otherwise its packets will not reach to the destination. Hence, the policing function can not be defeated.

11.2 Dynamic traffic shaping

RPR-fa controls ring access rate within allow_rate_congestion. Due to RPR-fa dynamic control nature, the allow_rate_congestion can vary in a wide range. When access traffic contains transmission bursts, ring access rate can vary between bursty low rate and high rate as limited by allow_rate_congestion. As a result, traffic on the ring can be bursty. To achieve better jitter performance, ring access rate can be dynamically shaped to a low-pass filtered value of allow_rate_congestion, which allows ring access rate conform to a fair rate as governed by RPR-fa while reduces extra bursty transmission.

RPR-fa dynamic shaper consists of 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 bursty transmission rate. The token octets are generated at a rate equal to the low-pass filtered value of allow_rate_congestion. For every octet that is transmitted, the token octets will be deducted by one. When a packet is waiting for accessing the ring at the head of its media access queue, it will be granted ring access if its rate conforms to a RPR-fa fair rate and there is at least one octet token in the leaky bucket. Therefore, the token could become negative after subtracting the octets of transmit packet. When the number of token reaches the maximum token octets, token bucket shall saturate.

11.3 Pre-Provision bandwidth for high priority traffic

For some application, it will be required to reserve certain bandwidth for high priority traffic. The reserved high priority bandwidth can be provisioned as reserved_rate. In RPR scheduling block, it will check if (add_rate + forward_rate) is less than (max_rate - reserved_rate). LP, eMP transmit packet, LP transit packet will be served only if the check is true. Otherwise, the congestion flow control will be triggered to request upstream nodes to constrain their traffic.

11.4 RPR ring access operation

The RPR-fa governs access to the ring. The RPR-fa only applies to Low priority and excess Medium priority (eMP) traffic. High priority and “within CIR” Medium priority traffic does not follow RPR-fa rules and may be transmitted at any time as long as traffic shaping/policing allow it, LP transit buffer is not almost full and HP transit buffer does not have a packet.

The RPR-fa requires few counters/registers which control the traffic forwarded and sourced on the RPR ring. Those counters are add_rate (tracks the amount of LP and eMP traffic sourced on the ring), add_rate_congestion (tracks the amount of LP and eMP traffic sourced on the ring and destined beyond the congestion point) and forward_rate (amount of LP and eMP traffic transitted through the station from the LP transit buffer), allow_rate_congestion (the current maximum LP+eMP transmit rate for that node beyond the congestion point) and max_rate (the current maximum LP+eMP transmit rate for that node).

Traffic policer shall not allow add_rate_congestion and add_rate to pass allow_rate_congestion and max_rate, respectively. It is accomplished by generating stop_low signal to the client when add_rate_congestion and add_rate exceed allow_rate_congestion and max_rate, respectively.

With no congestion, all nodes increment allow_rate_congestion periodically. The maximum value for allow_rate_congestion is max_rate. Max_rate is a per node parameter that limits the maximum amount of LP/eMP traffic that a node can send.

When a node sees congestion it starts to advertise a normalized add_rate value to upstream nodes. The value (nlp_add_rate) is obtained by passing add_rate through a low pass filter and then dividing by its weight. In this way, the fair rates passed on the links are always normalized to a weight of 1.0. Congestion detection is described in clause 12.2

When a node determines that it is a congestion point, it will send a fairness message. The TTL of the message should be set to 255. This allows any node receiving the message will be able to determine the number of hops to the congestion point (255 - TTL of received packet). A node that receives a non-null fairness message (rcvd_advertised_rate) will set its allow_rate_congestion and TTL_to_congestion to the rcvd_advertised_rate value multiplied by its weight and 255 - received TTL value, respectively. This allows a node with a weight of N to utilize N times as much bandwidth as a node with a weight of 1.0. If the source of the rcvd_advertised_rate is the same node that received it then the rcvd_advertised_rate shall be treated as a null value. When comparing the rcvd_advertised_rate source address the ring identifier of the fairness

packet must match the receiver's ring identifier in order to qualify as a valid compare. The exception is if the receive node is in the wrap state in which case the fairness packet's ring identifier is ignored.

Nodes that are not congested and that receive a non-null rcvd_advertised_rate generally propagate rcvd_advertised_rate to their upstream neighbor else propagate a null value of fair rate (all 1's). An exception occurs when an opportunity for local reuse is detected. The node compares its forward_rate (low pass filtered) to allow_rate_congestion divided by its weight. If the forward_rate is less than the normalized allow_rate_congestion, then a null value is propagated to the upstream neighbor instead of the rcvd_advertised_rate.

Nodes that are congested propagate the smaller of normalized nlp_add_rate and rcvd_advertised_rate.

Convergence is dependent upon number of nodes and distance. Simulation has shown convergence within 100 msec for rings of several hundred miles.

12. MAC fairness

12.1 Overview

RPR-fa is a Local Fairness algorithm that provides a fair access for all stations on the ring. The fairness algorithm consists of three components:

1) determine congestion threshold crossed or congestion alleviated
2) determine advertised rate
3) determine station allowed_rate

The RPR-fa is a mechanism that enforces fairness among the stations on the ring. It applies to only to LP and excess Medium priority (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.

The RPR-fa shall be implemented completely in the MAC Fairness control Entity (MFE). The MFE does not need to understand the ring topology, however it utilizes the implicit topology information (i.e. TTL value) passed by the MAC client to perform fairness and policing functions.

In RPR-fa, if a node experiences congestion, it will advertise a fair rate (e.g. add_rate) to upstream nodes via the opposite ring. 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 so as not to exceed the advertised value (adjusted by their weights). Generally, 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.

If the client passes the minimum possible value for the TTL, then it can take advantage of the available bandwidth on the ring otherwise (if it uses a larger than necessary TTL value) it will at least receive its fair share of the bandwidth from the most congested span that it contends for. 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 RPR-fa. VDQ combined with RPR-fa 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. RPR-fa 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 of the RPR-fa, 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 is declared by using 3 criteria depending on the type of transit path design.

If the transit path has only a single buffer, then congestion is detected when:

1) the outgoing link rate passes a configured congestion threshold. The outgoing link rate is measured using a byte counter which is passed through a low pass filter before the comparison. The node will remain congested until it's outgoing link utilization gets lower than a configured lower threshold. i.e. if (add_rate + forward_rate) is more than (max_rate - reserved_rate) then it is congested, or
2) access delay timer expires for LP and eMP packet

If the transit path has 2 buffers, then congestion is detected when:

1) the outgoing link rate passes a configured congestion threshold, or
2) access delay timer expires for LP and eMP packet, or
3) the depth of the low priority transit buffer reaches a congestion threshold.

12.3 Inter operability between single/dual transit buffer MACs

If a ring consists of mixed RPR MACs (single and dual transit buffer nodes), their fairness scheme will need to interact without disadvantaging any nodes on the ring. This will require a unique congestion message format and unique fairness algorithm.

The interaction between single transit and dual transit buffer nodes is the same as the interaction of same kind nodes. Upon receiving a congestion message the RPR MAC shall reduce its allowed rate to the fair rate received in the fairness message, and then forward the message upstream with the minimum of its own rate and the received fair rate.

The RPR MACs should rate shape their transmit traffic based on the dynamic traffic shaping algorithm described in section 11.2. Also, in mixed node ring configurations, dual transit buffer MACs should employ a traffic shaper to limit their outgoing LP traffic to a certain preset percentage (default value is 95%) of the line rate to allow downstream single transit buffer MACs to ensure their HP delay jitter.

12.4 Threshold settings

The high priority transit buffer needs to hold 2 to 3 MTUs or about 30KB.

The adequate sizing of the low priority transit buffer and associated high and low threshold values (TB_HI_THRESHOLD, TB_LO_THRESHOLD) depends on the ring size and traffic profile of the ring. Recommendation 10 G rings and ring diameter of 100km 256KB is adequate. For 10G 1000km rings 512KB and for 3000km rings 1MB of low priority transit buffer are recommended.

The goal of setting the appropriate threshold values is to deliver best possible end-to-end delay for the low priority traffic without penalizing the high priority traffic.

The following guidelines can be used to determine the proper threshold values:

TB_LO_THRESHOLD should be set to about 25% of the total buffer available. Lower values will result in higher end-to-end delays for low priority data packets. If either low or high priority data traffic is extremely burst, then a lower threshold value should be considered.

TB_HI_THRESHOLD should be set to about (total buffer size - 1MTU).

If the high priority data traffic has a burst nature a more conservative (lower) value is recommended to avoid overflow of the low priority transit buffer.

12.5 RPR fairness packet format

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

If a receive interface has not seen a fairness packet within the keepalive time-out interval it will trigger an L2 keepalive time-out interrupt/event. The protection software will subsequently mark that interface as faulty and initiate a protection switch around that interface. The keepalive time-out interval should be set to 16 times the RPR fairness packet transmission interval.

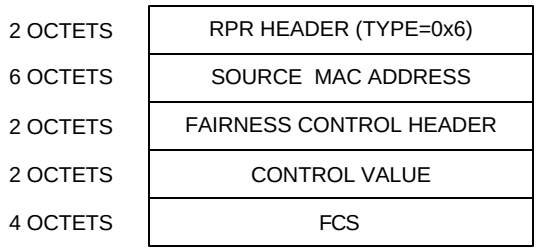
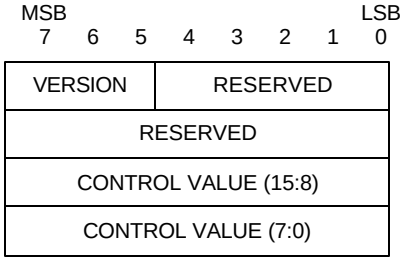


Figure 23—Fairness Packet Format

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



12.5.1 When generated

RPR-fa Fairness messages are generated periodically. They are also act as keepalives informing the upstream station that a valid data link exists for Protection entity.

12.5.2 Version field (3 bits)

This field is to specify the version number of fairness packet. Table 9 shows the fairness message version values. Type 1 fairness message is used to implement basic RPR-fa. 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 MFE and information contained is passed to the MAC clients whereas type 2 messages are not processed by MFE and passed to the MAC clients as well.

Table 9—Version values

Value (binary)	Type of fairness packet	How is it used
000	Type 1 fairness packet	Follows RPR-fa fair rate rules, generated in every fair rate interval, SA is the MAC address of the most congested node in the fairness domain
001	Type 2 fairness packet	Generated in every 10 fair rate intervals by every MAC with its SA and broadcast - fair rate is passed to each client on the way
010 to 111	Reserved	Future use

12.5.3 Reserved field (12 bits)

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

12.5.4 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.

12.5.5 Control value (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 a 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.

13. Topology discovery

Each node performs a topology discovery by sending out topology discovery packets on one or both rings. The node originating a topology packet marks the packet with the egressing ring identifier, appends the node's mac binding to the packet and sets the length field in the packet before sending. This packet is a point-to-point packet which hops around the ring from node to node. Each node appends its mac address binding, updates the length field and sends it to the next hop on the ring. If there is a wrap on the ring, the wrapped node will indicate a wrap when appending its mac binding and then wrap the packet. When the topology packets travel on the wrapped section with the ring identifier being different from that of the topology packet itself, the mac address bindings are not added to the packet.

Eventually the node that generated the topology discovery packet gets back the packet. The node makes sure that the packet has the same ingress and egress ring identifier before accepting the packet. A topology map is changed only after receiving two topology packets which indicate the same new topology (to prevent topology changes on transient conditions).

Besides periodical topology discovery, the topology could be updated accordingly whenever an protection switch request message is received or a fiber failure is detected by local node.

Note that the topology map only contains the reachable nodes. It does not correspond to the failure-free ring in case of wraps and ring segmentations.

Note that the Source address should be set to the source address of the TRANSMITTING node (which is not necessarily the ORIGINATING node).

2 OCTETS	RPR HEADER(TYPE=0x5)
6 OCTETS	DESTINATION MAC ADDRESS
6 OCTETS	SOURCE MAC ADDRESS
2 OCTETS	PROTOCOL=RPR Control
2 OCTETS	HEADER CHECKSUM
1 OCTET	CONTROL VERSION(0x0)
1 OCTET	CONTROL TYPE(0x1)
2 OCTETS	CONTROL TTL
2 OCTETS	TOPOLOGY LENGTH
6 OCTETS	ORIGINATOR's MAC ADDRESS
2 OCTETS	MAC TYPE
6 OCTETS	MAC ADDRESS
nn OCTETS	OTHER MAC BINDINGS
4 OCTETS	FCS

Figure 24—Topology Packet Format

13.1 Topology discovery packet format

13.1.1 Topology length

This two octet field represents the length of the topology message in octets starting with the first MAC Type/ MAC Address binding.

13.1.2 Topology originator

A topology discovery packet is determined to have been originated by a node if the originator's globally unique MAC address of the packet is that node's globally unique MAC address (assigned by the IEEE).

Because the mac addresses could be changed at a node, the IEEE MAC address ensures that a unique identifier is used to determine that the topology packet has gone around the ring and is to be consumed.

13.1.3 MAC bindings

Each MAC binding shall consist of a MAC Type field followed by the node's 48 bit MAC address. The first MAC binding shall be the MAC binding of the originator. Usually the originator's MAC address will be it's globally unique MAC Address but some implementations may allow this value to be overridden by the network administrator.

13.1.4 MAC type format

The MAC type is used to indicate the characteristic of a node. This 2-octets field is encoded as follows:

Table 10: MAC type format

Bit	Value
0	Single transit buffer(0)/Dual transit buffer
1	Ring identifier (1 or 0)
2	Wrapped node (1) / Unwrapped node (0)
3	Wrap protection capable(1)
4-6	Fairness message version
7-13	Weight
14-15	Reserved

Determination of whether a packet's egress and ingress ring identifier's are a match should be done by using the ring identifier found in the MAC Type field of the last MAC binding as the ingress ring identifier.

The topology information is not required for the protection mechanism. This information can be used to calculate the number of nodes in the ring as well as to calculate hop distances to nodes to determine the shortest path to a node (since there are two counter-rotating rings).

The implementation of the topology discovery mechanism could be a periodic activity or on "a need to discover" basis. In the periodic implementation, each node generates the topology packet periodically and uses

the cached topology map until it gets a new one. In the need to discover implementation, each node generates a topology discovery packet whenever they need one e.g., on first entering a ring or detecting a wrap.

13.2 Topology discovery state transition

13.2.1 Constants

— Topology_Query_Timeout_Time
The number of seconds to wait for triggering topology discovery process.

13.3 Variables

— Topology_Query_Ctrl
Topology discovery control packet
— Local_MAC_Add
MAC address of local MAC
— MAC_Ring_ID
ring identifier of local MAC
— Local_Fiber_Failure
detect a local fiber failure
— Protection_Message
protection message
— MAC_Type
The attribute of local MAC
— Ring_ID
ring identifier of received topology discovery packet
— Originated_MAC_Add
The originated MAC of received topology discovery packet
— CONT_TTL
The control ttl of received topology discovery packet
— BEGIN
A Boolean variable that is set to TRUE when the System is initialized or reinitialized, and is set to FALSE when (re-)initialization has completed.
Value: Boolean

13.3.1 Timers

— Topology_Query_Timer

This Timer is used to trigger topology discovery periodically. The initial value is Topology_Query_Timeout_Time

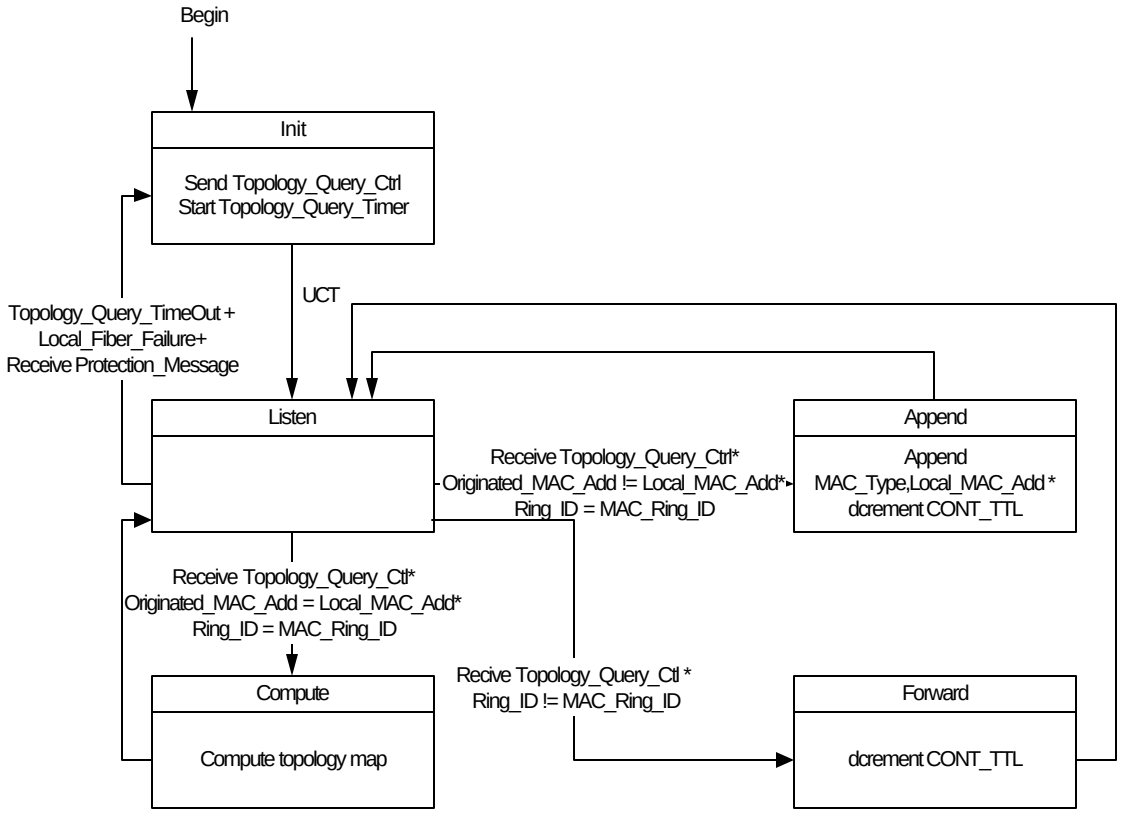


Figure 25—Topology discovery state transition diagram

14. Protection

Resiliency is an important RPR objective. That requires to provide protection within 50ms in case of ring or node failure. There are two known protection mechanisms: wrapping and steering. Steering is mandatory protection mechanism and wrapping is optional one for RPR nodes. However, all nodes within the same RPR ring shall choose the same protection mechanism.

In the topology discovery, every RPR node shall indicate if it supports wrapping protection or not. If all nodes are able to support wrapping protection, protection mechanism shall be based on provisioning to pick wrapping or steering. Otherwise, steering shall be selected as the protection scheme within RPR ring.

A protection message shall be proposed to signal the failure between node within the same RPR ring. That message shall accommodate the necessary information for RPR to do wrap or steer protection.

14.1 Wrap protection

A RPR Ring is composed of two counter-rotating, single fiber rings. If an equipment or fiber facility failure is detected, traffic going towards and from the failure direction is wrapped (looped) back to go in the opposite direction on the other ring (subject to the protection hierarchy). Wrapping takes place on the nodes adjacent to the failure, under control of the protection switch protocol. The wrap re-routes the traffic away from the failed span.

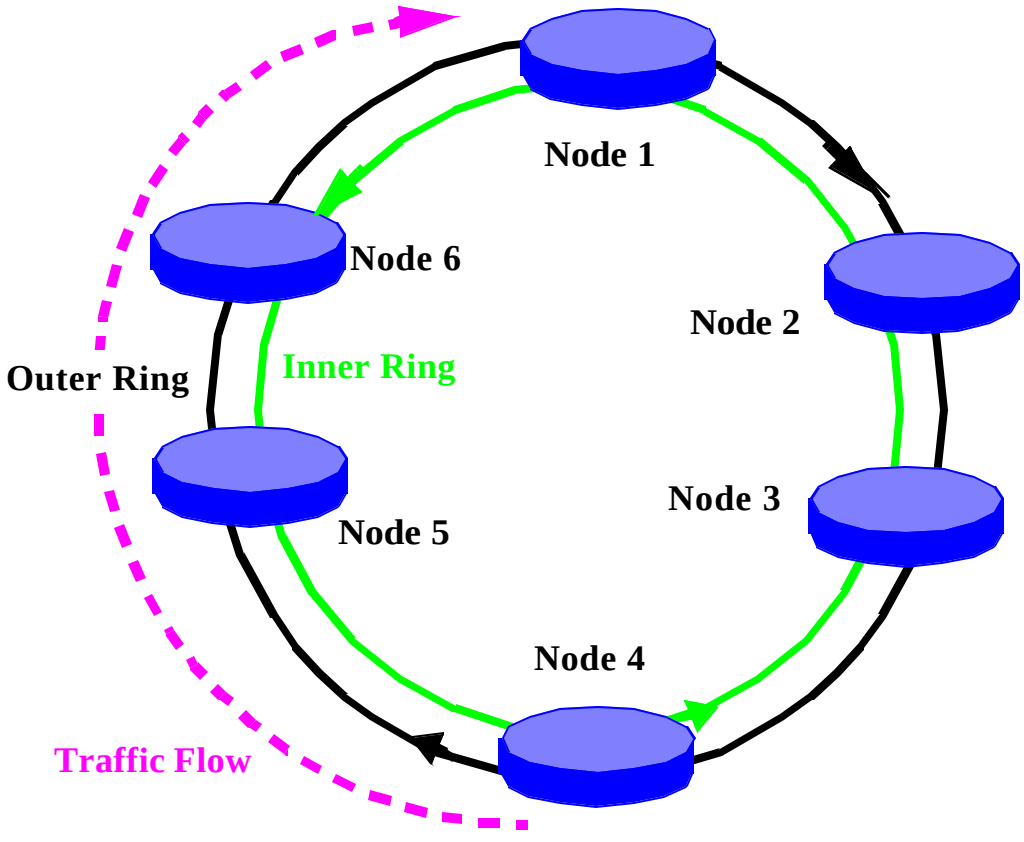


Figure 26—Data flow before fiber cut.

An example of the data paths taken before and after a wrap are shown in Figure 26 and Figure27. Before the fiber cut, N4 sends to N1 via the path N4->N5->N6->N1.

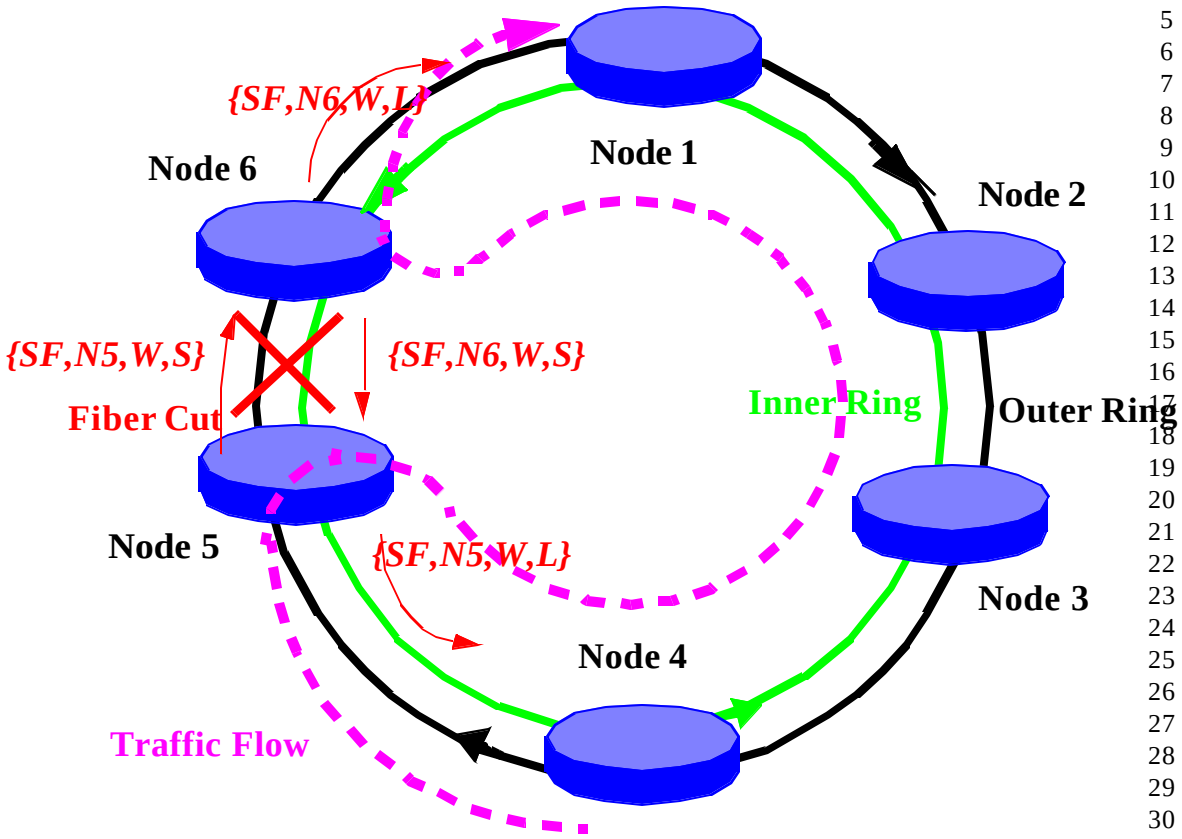


Figure 27—Data Path after Wrap

If there is a fiber cut between N5 and N6, N5 and N6 will wrap the inner ring traffic to the outer ring. After the wraps have been set up, traffic from N4 to N1 initially goes through the non-optimal path N4->N5->N4->N3->N2->N1->N6->N1.

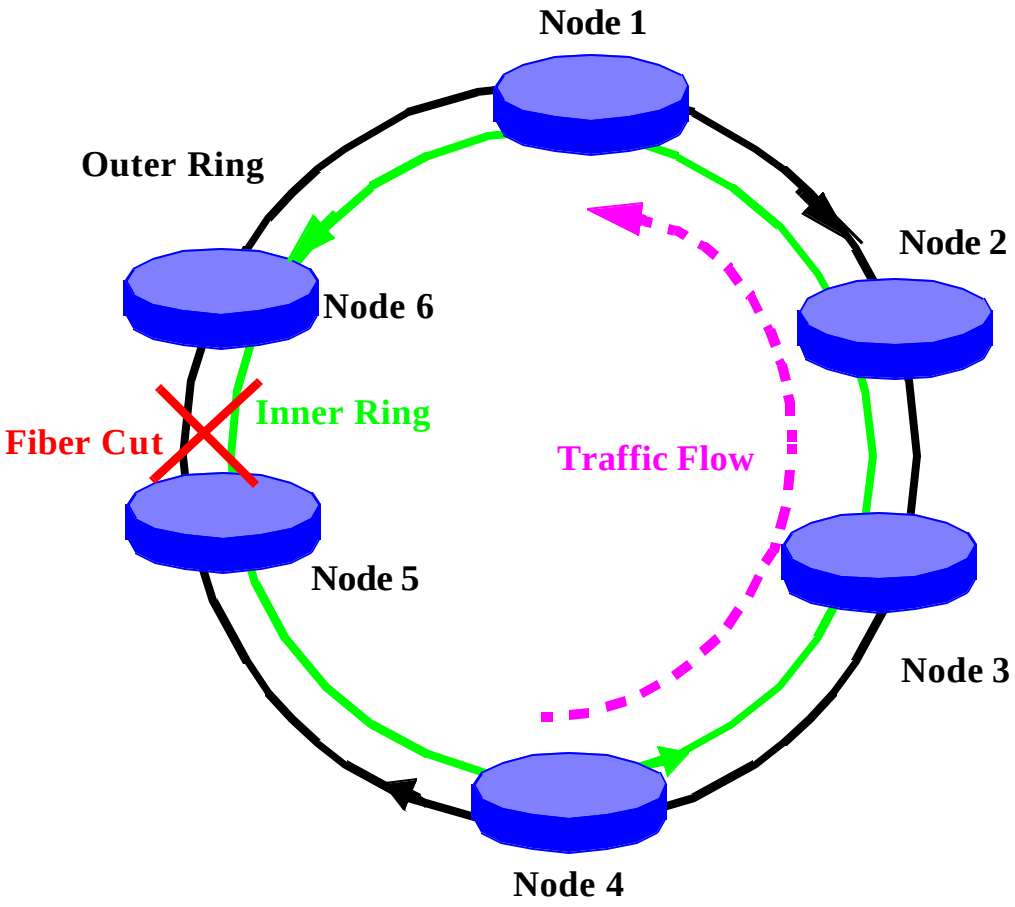


Figure 28—Data Path after new Topology Discovery

Subsequently a new ring topology is discovered and a new optimal path N4->N3->N2-N1 is used, as shown in Figure28. Note that the topology discovery and the subsequent optimal path selection are not part of the protection switch protocol.

The ring wrap is controlled through SONET BLSR [3][4] style protection switch signaling. It is an objective to perform the wrapping as fast as in the SONET equipment or faster.

14.2 .Steering protection

For steering protection, a node shall not wrap the failed span while a failure is detected. A protection request message will be sent to every node to indicate there is a fiber cut just like in the wrap protection scheme. When nodes receive the protection request message indicating the failure, the topology will be updated accordingly. It will be the responsibility of source traffic node to direct the traffic inner or outer ring to avoid the failed span.

Packets that have been transmitted onto the ring that are destined to a node beyond the point of failure before the topology is updated at the source node will be dropped at the failure point since there is no delivery mechanism available.

14.3 Multicast consideration

For the case of a steering ring, when the ring fails, the packet must be sent in both direction with the TTL set to expire after one trip and not allowing duplicate packets to exist when the station is restored. For the case of a wrapping ring, the multicast packet can be sent in single direction with the TTL set to be the number of nodes on the ring.

14.4 Protection hierarchy

The protection switch protocol processes the following request types (in the order of priority, from highest to lowest):

1) Forced Switch (FS): operator originated, performs a protection switch on a requested span (wraps at both ends of the span)
2) Signal Fail (SF): automatic, caused by a media Signal Failure or RPR keep-alive failure - performs a protection switch on a requested span
3) Signal Degrade (SD): automatic, caused by a media Signal Degrade (e.g. excessive Bit Error Rate) - performs a protection switch on a requested span
4) Manual Switch (MS): operator originated, like Forced Switched but of a lower priority
5) Wait to Restore (WTR): automatic, entered after the working channel meets the restoration criteria after SF or SD condition disappears. protection switch waits WTR period before restoring traffic in order to prevent protection switch oscillations

As the protection requests travel around the ring, the protection hierarchy is applied. If the requested protection switch is of the highest priority (e.g. Signal Fail request is of higher priority than the Signal Degrade) then this protection switch takes place and the lower priority switches elsewhere in the ring are taken down, as appropriate. When a lower priority request is presented, it is not allowed if a higher priority request is present in the ring. The only exception is multiple SF and FS switches, which can coexist in the ring.

All protection switches are performed bidirectionally (protect at both ends of a span for both transmit and receive directions, even if a failure is only unidirectional).

14.5 Protection message packet format

Protection switch protocol is a method for automatically recovering from various ring failures and line degradation scenarios. The protection switch message packet format is outlined in Figure29 below.

2 OCTETS	RPR HEADER(TYPE=0x4)
6 OCTETS	DESTINATION MAC ADDRESS
6 OCTETS	SOURCE MAC ADDRESS
2 OCTETS	PROTOCOL=RPR Control
2 OCTETS	HEADER CHECKSUM
1 OCTET	CONTROL VERSION(0x0)
1 OCTET	CONTROL TYPE(0x2)
1 OCTET	Protection Message Octet
1 OCTET	Reserved
4 OCTETS	FCS

Figure 29—Protection Switch Packet Format

The protection switch specific fields are detailed below.

14.5.1 Destination MAC address

The Destination MAC address is a pre-registered multicast address for protection switch packets. Therefore the transmission delay for protection switch packet can be minimized.

14.5.2 Source MAC address

This is the MAC address of the originator of the protection message.

14.5.3 Protection message octet

The protection message octet contains specific protection information. The format of the protection message octet is as follows:

The currently defined request types with values, hierarchy and interpretation are as used in SONET BLSR [3], [4], except as noted.

14.5.4 The Protection message request types

The following is a list of the request types, from the highest to the lowest priority. All requests are signaled using protection control messages.

1) Forced Switch (FS - operator originated)
   This command performs the ring switch from the working channel to the protection, wrapping

Table 11: Protection message octet format

Bit	Value
0-3	**Protection Message Request Type** 1101 - Forced Switch (FS) 1011 - Signal Fail (SF) 1000 - Signal Degrade (SD) 0110 - Manual Switch (MS) 0101 - Wait to Restore (WTR) 0000 - No Request (IDLE)
4	**Path Indicator** 0 - Short (S) 1 - Long (L)
5-7	**Status Code** 010 - Protection Switch Completed - Traffic Wrapped (W) 000 - Idle

the traffic on the node at which the command is issued and at the adjacent node to which the command is destined. Used for example to add another node to the ring in a controlled fashion.

2) Signal Fail (SF - automatic)
   Protection caused by a media “hard failure” or RPR keep- alive failure. SONET examples of SF triggers are: Loss of Signal (LOS), Loss of Frame (LOF), Line Bit Error Rate (BER) above a preselected SF threshold, Line Alarm Indication Signal (AIS). Note that the RPR keep-alive failure provides end-to-end coverage and as a result SONET Path triggers are not necessary.
3) Signal Degrade (SD - automatic)
   Protection caused by a media “soft failure”. SONET example of a SD is Line BER or Path BER above a preselected SD threshold.
4) Manual Switch (MS - operator originated)
   Like the FS, but of lower priority. Can be used for example to take down the WTR.
5) Wait to Restore (WTR - automatic)
   Entered after the working channel meets the restoration threshold after an SD or SF condition disappears. Protection switch protocol waits WTR time-out before restoring traffic in order to prevent protection switch oscillations.

14.5.5 The Protection message path indicator

There are two types of protection messages, long and short. Short messages are sent to the other side of failed span through the opposite ring. They indicate a failure on the other ring before the source address of the protection request packet. Long messages, on the other hand, indicate there is a failure after the source address of the protection request packet.

The protection control messages are shown in this document as:

{REQUEST_TYPE, SOURCE_ADDRESS, WRAP_STATUS, PATH_INDICATOR}

14.6 RPR protection protocol states

Each node in the protection protocol is in one of the following states for each of the rings:

14.6.1 Idle

In this mode the node is ready to perform the protection switches and it sends to both neighboring nodes "idle" protection messages, which include "self" in the source address field {IDLE, SELF, I, S}

14.6.2 Wrapping/Steering

Node participates in a protection switch. This state is entered based on a protection request issued locally or based on received protection messages.

14.7 Protection protocol rules

14.7.1 RPR protection packet transfer mechanism

R T.1:

Protection packets are transferred in a broadcast packet format between nodes on the ring. A received packet (payload portion) is passed to MAC control section.

R T.2:

All protection messages are triggered by self-detect or user request. The message is sent while the state is changed.

14.7.2 RPR protection signaling and wrapping mechanism

R S.1:

Protection switch signaling is performed using protection control packets as defined in Figure29 "Protection Switch Packet Format".

R S.2:

Node executing a local request signals the protection request on both short (across the failed span) and long (around the ring) paths after detecting a fiber failure. The node will repeat the protection message every T1 timer until it receive its protection message. Default T1 timer is one second.

R S.3:

Protection protection packets are never wrapped.

R S.4:

If the protocol calls for sending both short and long path requests on the same span (for example if a node has all fibers disconnected), only the short path request should be sent.

R S.5:

A node wraps and unwraps only on a local request or on a short path request. A node never wraps or unwraps as a result of a long path request. Long path requests are used only to maintain the protection hierarchy. (Since the long path requests do not trigger protection, there is no need for destination addresses and no need for topology maps).

14.7.3 Example

In Figure30, Node A detects SF (local request/ self-detected request) on the span between Node A and Node B and starts sourcing {SF, A, W, S} on the outer ring and {SF, A, W, L} on the inner ring. Node B receives the protection request from Node A (short path request) and starts sourcing {IDLE, B, W, L} periodically.

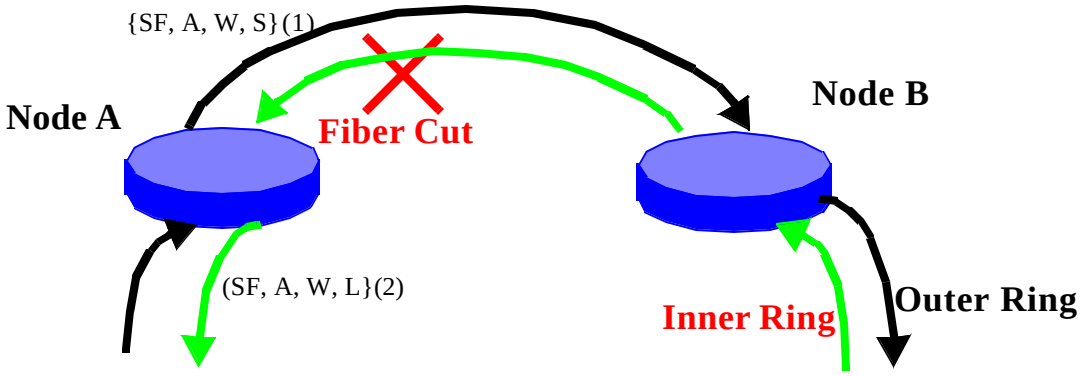


Figure 30—RPR Protection Switch Signaling

14.8 RPR protection protocol rules

R P.1:

Protection Request Hierarchy is as follows (Highest priority to the lowest priority). In general a higher priority request preempts a lower priority request within the ring with exceptions noted as rules. The 4 bit values below correspond to the REQUEST_TYPE field in the protection packet.

Table 12: Protection request type

Value
Protection request type 1101 - Forced Switch (FS) 1011 - Signal Fail (SF) 1000 - Signal Degrade (SD) 0110 - Manual Switch (MS) 0101 - Wait to Restore (WTR) 0000 - No Request (IDLE)

R P.2:

Requests >= SF can coexist. All requests above SF need to be cleared before the state is transferred into idle state.

R P.3:

Requests < SF can not coexist with other requests. A higher priority request will preempt a lower priority request.

R P.4:

A node always honors the highest of {short path request, self detected request} if there is no higher long path message passing through the node.

R P.5:

When there are more requests of priority < SF, the first request to complete long path signaling will take priority. However, a higher request can preempt the request as long as its long path signal is completed.

R P.6:

A node will strip an protection packet which was originally generated by the node itself (it has the node's source address).

R P.7:

When a node receives a long path request and the request is >= to the highest of {short path request, self detected request}, the node checks the message to determine if the message is coming from its neighbor on the short path. If that is the case then it strips the message.

R P.8:

When a node receives a long path request, it strips (terminates) the request if it is a wrapping/steering node with a request >= than that in the request; otherwise it passes it through.

R P.9:

Each node keeps track of the addresses of the immediate neighbors through topology discovery.

R P.10:

When a node (which initially detected the failure) discovers disappearance of the failure, it enters WTR (user-configured WTR time-period). WTR can be configured in the 10-600 sec. range with a default value of 60 sec.

R P.11:

When a node is in WTR mode, and detects that the new neighbor (as identified from the received short path protection message) is not the same as the old neighbor (stored at the time of wrap initiation), the node drops the WTR.

R P.12:

When a node is in WTR mode and the source of long path request is not equal to its neighbor on the opposite side (as stored at the time of wrap initiation), the node drops the WTR. This is the case when a new neighbor add to the ring.

R P.13:

When a node receives a local protection request of type SD or SF and it cannot be executed (according to protocol rules) it keeps the request pending. (The request can be kept pending outside of the protection protocol implementation).

R P.14:

If a local non-failure request (WTR, MS, FS) clears and if there are no other requests pending, the node enters idle state.

R P.15:

If there are two failures and two resulting WTR conditions on a single span, the second WTR to time out brings both the wraps down (after the WTR time expires a node does not unwrap automatically but waits till it receives idle messages from its neighbor on the previously failed span)

R P.16:

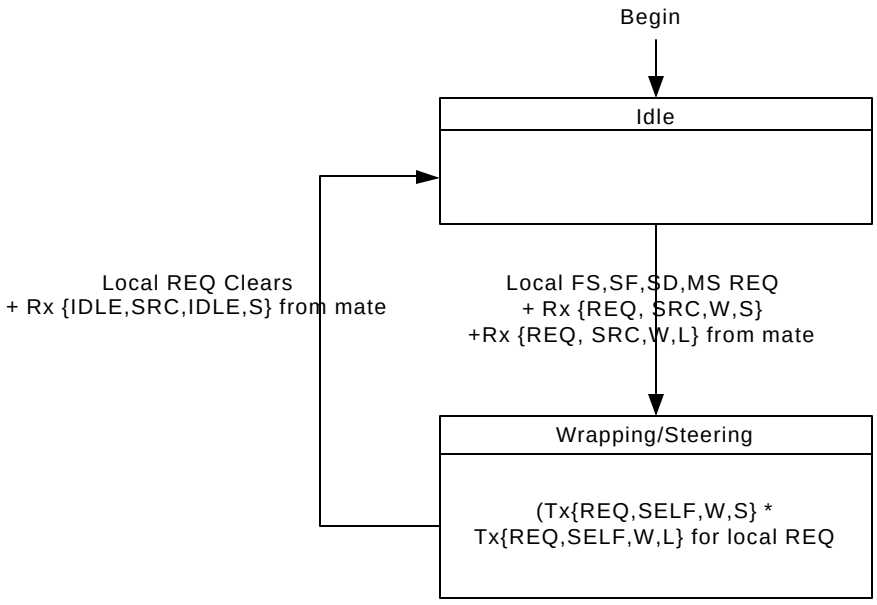
If a short path FS request is present on a given side and a SF/SD condition takes place on the same side, accept and process the SF/SD condition ignoring the FS. Without this rule a single ended wrap condition could take place. (Wrap on one end of a span only).

R.P.17:

If a node receives a protection message, it should update its topology accordingly and trigger a topology discovery process.

14.9 Protection state transition

Figure31 shows the simplified state transition diagram for the protection protocol:



Lengend:
mate = node on the other end of the affected span
REQ is any othercontext = FS,SF,SD,MA

Figure 31—Simplified Protection State Transitions Diagram

14.10 Failure examples

14.10.1 Signal failure - single fiber cut scenario

Sample scenario in a ring of four nodes A, B, C and D, with unidirectional failure on a fiber from A to B, detected on B. Ring is in the Idle state (all nodes are Idle) prior to failure.

14.10.1.1 Signal fail scenario

1) Node B detects SF on the outer ring, transitions to Wrapped state (performs a wrap), Tx towards A on the inner ring/short path: {SF, B, W, S} and on the outer ring/long path: Tx {SF, B, W, L}
2) Node A receives protection request on the short path, transitions to Wrapped state.
3) Steady state is reached

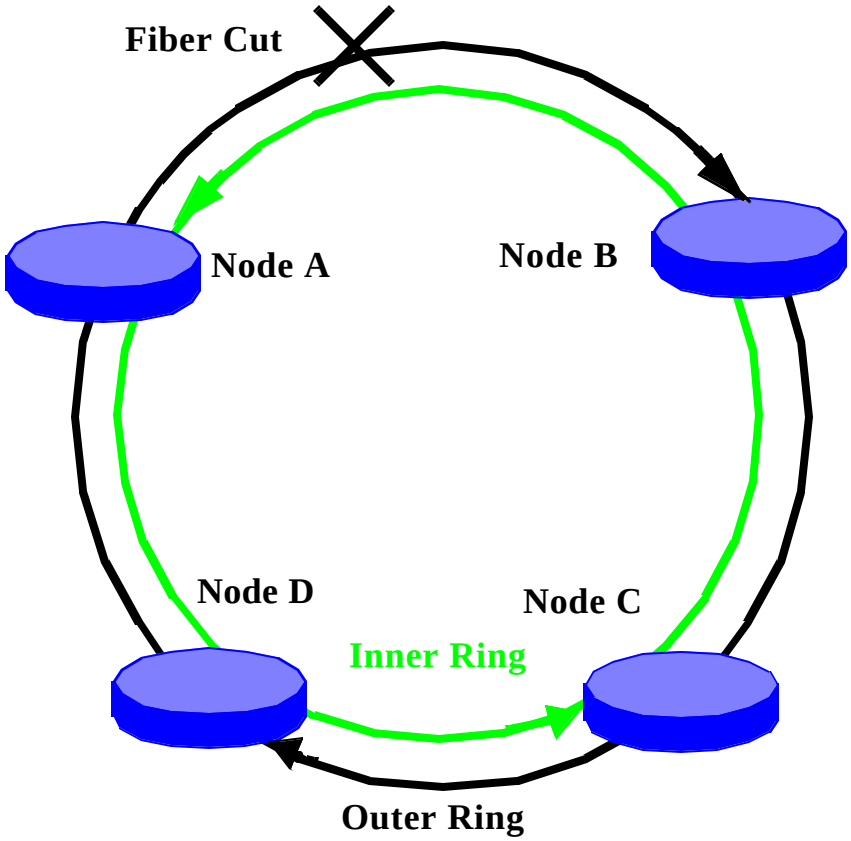


Figure 32—An RPR Ring with a fiber cut in outer ring

14.10.1.2 Signal fail clears

1) SF on Node B clears, Node B does not unwrap if it is in wrap state, sets WTR timer, Tx {WTR, B, W, S} on inner and Tx {WTR, B, W,L} on outer ring.
2) Node A receives WTR request on the short path, does not unwrap.
3) Steady state is reached
4) WTR times out on B. B transitions to idle state (unwraps) Tx {IDLE, B, I, S} on inner ring and Tx{IDLE,B,I.L} on outer ring.
5) Node A receives Rx {IDLE, B, I, L} and transitions to Idle
6) Steady state it reached

14.10.2 Signal failure - bidirectional fiber cut scenario

Sample scenario in a ring of four nodes A, B, C and D, with a bidirectional failure between A and B. Ring is in the Idle state (all nodes are Idle) prior to failure.

14.10.2.1 Signal fail scenario

1) Node A detects SF on the inner ring, transitions to Wrapped state (performs a wrap), Tx towards B on the outer ring/short path: {SF, A, W, S} and on the inner ring/long path: Tx {SF, A, W, L}
2) Node B detects SF on the outer ring, transitions to Wrapped state (performs a wrap), Tx towards A on the inner ring/short path: {SF, B, W, S} and on the outer ring/long path: Tx {SF, B, W, L}

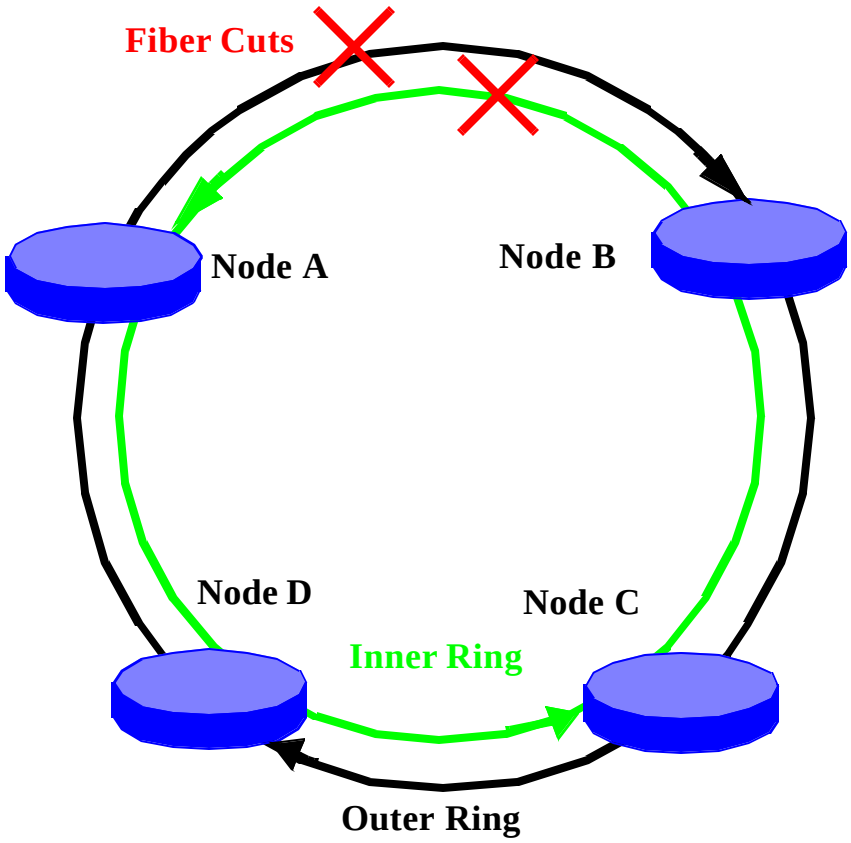


Figure 33—An RPR Ring with bidirectional fiber cut

3) Steady state is reached

14.10.2.2 Signal fail clears

1) SF on A clears, A does not unwrap, sets WTR timer, Tx {WTR, A, W, S} through outer ring towards B and Tx {WTR, A, W, L} on the long path through inner ring.
2) SF on B clears, B does not unwrap. Since it now has a short path WTR request present from A it acts upon this request. It keeps the wrap, Tx {IDLE, B, W, S} towards A and Tx {WTR, B, W, L} on the long path
3) Nodes C and D relay long path messages without changing the protection switch octet
4) Steady state is reached
5) WTR times out on A. A enters the idle state (drops wraps) and starts transmitting idle in both rings
6) B sees idle request on short path and enters idle state
7) Steady state is reached

14.10.3 Failed node scenario

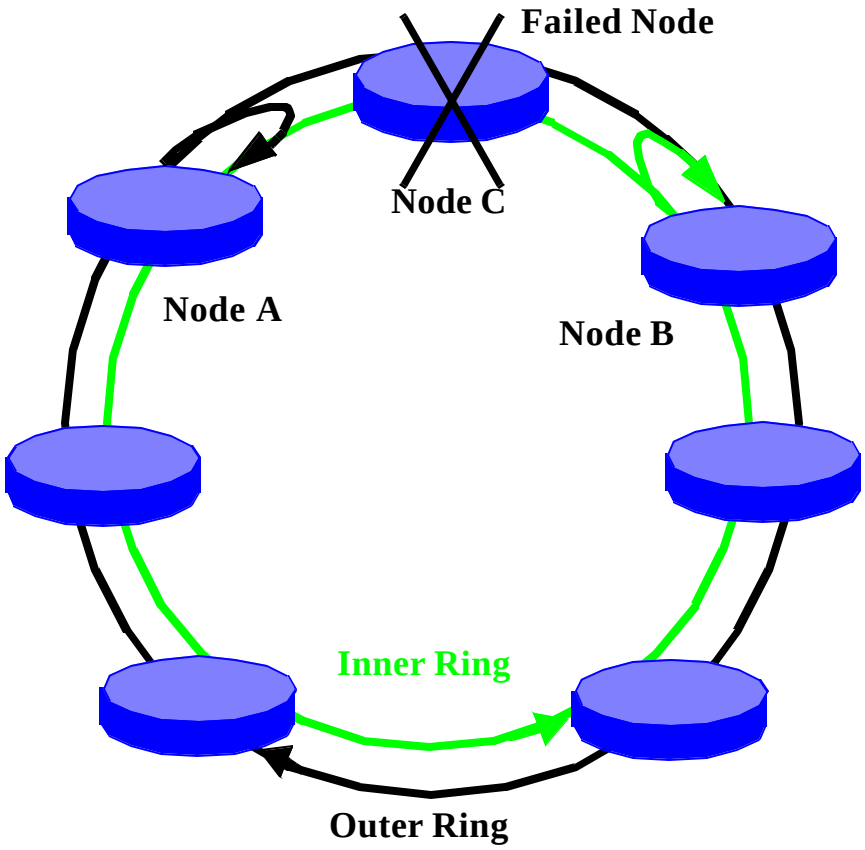


Figure 34—An RPR Ring with a failed node

Sample scenario in a ring where node C fails. Ring is in the Idle state (all nodes are Idle) prior to failure.

14.10.3.1 Node failure (or fiber cuts on both sides of the node)

1) B detects SF on the outer ring, transitions to Wrapped state (performs a wrap), Tx towards C on the inner ring/short path: {SF, B, W, S} and on the outer ring/long path: Tx {SF, B, W, L}
2) A detects SF on the inner ring, transitions to Wrapped state (performs a wrap), Tx towards C on the outer ring/short path: {SF, A, W, S} and on the inner ring/long path: Tx {SF, A, W, L }
3) Steady state is reached

14.10.3.2 Failed node and one span return to service

Note: Practically the node will always return to service with one span coming after the other (with the time delta potentially close to 0). Here, a node is powered up with the fibers connected and fault free.

1) Node C and a span between A and C return to service (SF between A and C disappears)
2) Node C, not seeing any faults starts to source idle messages {IDLE, C, I, S} in both directions.
3) Fault disappears on A and A enters a WTR (briefly)

4) Node A receives idle message from node C. Because the long path protection request {SF, B, W, L} received over the long span is not originating from the short path neighbor (C), node A drops the WTR.
5) Steady state is reached

14.10.3.3 Second span returns to service

The scenario is like the Bidirectional Fiber Cut fault clearing scenario.

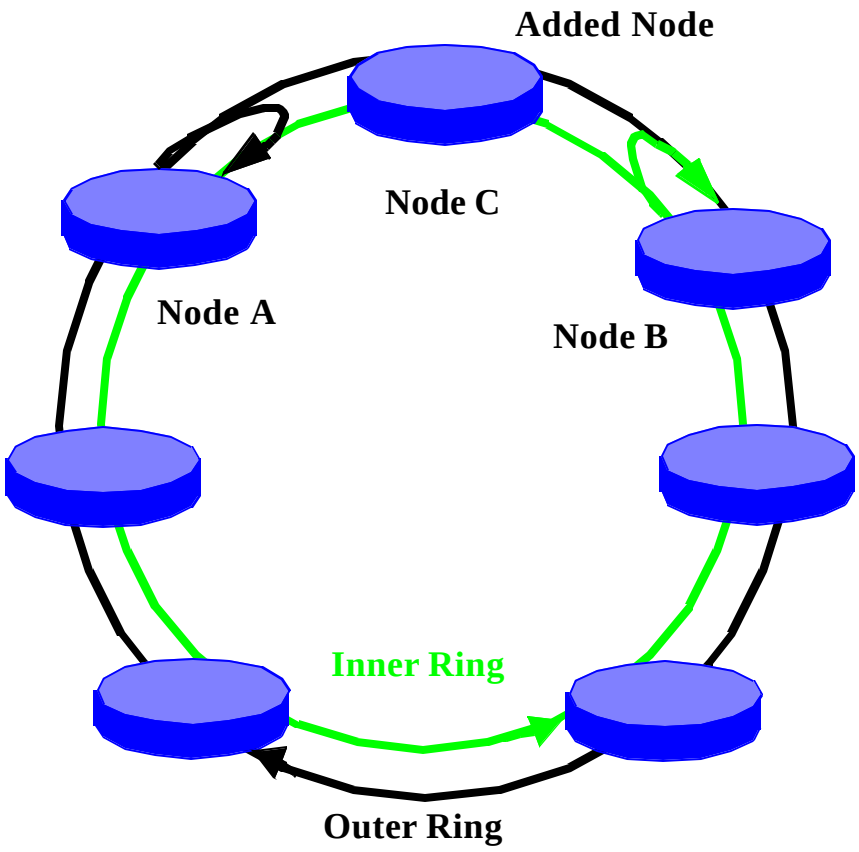


Figure 35—An RPR Ring with a failed node

Sample scenario in a ring where initially nodes A and B are connected. Subsequently fibers between the nodes A and B are disconnected and a new node C is inserted.

14.10.3.4 Bidirectional fiber cut

1) Fibers are removed between nodes A and B
2) B detects SF on the outer ring, transitions to Wrapped state (performs a wrap), Tx towards A on the inner ring/short path: {SF, B, W, S} and on the outer ring/long path: Tx {SF, B, W, L}
3) A detects SF on the inner ring, transitions to Wrapped state (performs a wrap), Tx towards B on the inner ring/short path: {SF, A, W, S} and on the outer ring/long path: Tx {SF, A, W, L}

4) As the nodes on the long path between A and B receive a SF request, they enter a pass-through mode (in each direction), stop sourcing the Idle messages and start passing the messages between A an B
5) Steady state is reached

14.10.3.5 Node C is powered up and fibers between nodes A and C are reconnected

This scenario is identical to the returning a Failed Node to Service scenario.

14.10.3.6 Second span put into service

Nodes C and B are connected. The scenario is identical to Bidirectional Fiber Cut fault clearing scenario

15. Ringlet selection

16. Operation administration and maintenance (OAM)

16.1 Overview

This section provides OAM&P (Operation, Administration, Maintenance and Provisioning) functionalists supported by RPR equipment.

Management Functional Areas pertinent to RPR are:

- Configuration management
- Fault management
- Performance management
- Security management (not addressed in this draft)

Configuration Management exercises control over, identifies, collects data from, and provides data to Network Elements (NEs) and the connections between NEs. Configuration Management is responsible for the installation of NEs, their interconnection into a network (configuration) and provisioning. [Note that it does not check and report configuration anomalies, this is fault management. Depending on the system used to configure the NE, it may not allow the user to enter an invalid value, but it cannot be used for checking configuration errors.]

Fault (or Maintenance) Management enables the detection, isolation, and correction of abnormal operation of the NEs and its network. It is responsible to detect and process any faults as well as to report it to the management system.

Performance Management evaluates and reports upon the behavior of NEs and the effectiveness of the network and NEs for the support of services. Performance Management provides mechanisms to measure service quality, by monitoring the system performance. It also reports statistics information to the management system

In order to improve fault and performance management capability, e.g. to allow fault detection, some in-band OAM functions are envisaged.

OAM functions in a network are performed on hierarchical levels. Not all protocols support OAM, for example a physical layer based on SONET/SDH includes extensive OAM functions, while a physical layer based on Ethernet phys lacks OAM functions completely.

SONET/SDH has in-band OAM functionality specified in ITU G.783 [B3]. Based on G.783 specification, it is possible to detect Signal Fail (SF) in each span (SONET SPE or SDH VC). SONET/SDH interfaces used in RPR systems should support the OAM requirements as specified in Recommendation G.783 [B3].

The IEEE 802.3 standards for 1 GbE (802.3ac) and 10 GbE (802.3ae) LAN PHY interfaces have no in-band OAM functionality defined. The only information that is known is if the physical interface is available or not available. The unavailability of the physical interface directly translates into a Signal Fail (SF) condition for the MAC layer.

In-band OAM functionality for IEEE 802.3 standard 10 GbE (802.3ae) WAN PHY interface is TBD.

IEEE 802.17 RPR re-uses existing in-band OAM mechanisms provided by the PHY layers.

In-band OAM functionality used in upper layers is outside the scope of the IEEE 802.17 Standard.

16.1.1 OAM functions of the RPR layer

The OAM function in RPR is based on special frames sent between Stations. These frames correspond to flows. A flow is defined by the SA and DA.

Fault Management frames are used to indicate components, Stations and ring failures, loss of continuity between ring Stations, and to perform Loopback operations.

Activation/Deactivation frames are used to Activate or Deactivate the transmission of Continuity Check frames. These frames allow coordinating the transmission and reception of continuity frames to avoid the generation of undesirable alarm indication.

[Editor's note - The need for the activation/deactivation flow is to be discussed.]

The OAM frame types of the RPR layer are:

RDI - For reporting defect indications in the backward direction on a flow level

CC - For monitoring continuity of flows

LB - For on demand connectivity monitoring and fault localization on flows

Activation/Deactivation - For Activating/Deactivating CC.

16.2 Fault management

Fault management includes alarm surveillance, fault localization, fault correction and testing. Alarm Surveillance provides the capability to monitor failures detected in NEs. In support of alarm surveillance RPR NEs should perform checks on hardware and software in order to detect failures, and generate alarms for such failures. Upon detecting a failure, in addition to generating and sending alarms to systems, NEs should also send RDI in the backward direction, in order to notify the peer node that a failure has occurred (and some action is required).

Loss Of Continuity (LOC) is the only defect defined at this stage that the NE has to detect. This is addressed by the use of a continuity check (CC) mechanism. CC also assists in fault localization, since it is possible to identify between which NEs the flow is interrupted. Another type of failure that RPR NEs may identify is software misconfiguration/failure. Such failure/misconfiguration can lead to invalid/unrecognizable header field value when the RPR frame is generated. Software checks can be performed on the RPR header to check for invalid/unrecognizable field value.

Fault Localization determines the root cause of a failure. In addition to the initial failure information, it may use failure information from other entities in order to correlate and localize the fault.

Fault Correction is responsible for the repair of a fault and for the control of procedures that use redundant resources to replace equipment or facilities that have failed. For RPR, in case of fiber cut or node failure, a protection switching is used to restore service.

Testing performs repair functions using some testing and diagnostic routines. Testing is characterized as the application of signals/messages and their measurement. Loopback is one example of a testing routine and can be activated upon request.

The fault management mechanisms are useful to check the reachability at the MAC layer between two RPR nodes on the ring, especially when there are some failures (e.g. one station on the ring steals packets addressed to some other stations) that are not detected at layer 1.

This section defines two Fault Management mechanisms.

The first mechanism is an on-demand in-service loopback mechanism (see section 1.2.3) used for trouble-shooting the RPR network (reactive mechanism).

The second mechanism is a Continuity Check (CC) that is used for fault detection (proactive mechanism) by continuously running and warning the operator when there are some problems (see section 1.2.1).

As part of the Fault Management, a Remote Defect Indication (RDI) mechanism is also defined. It is used together with the CC mechanism to inform the source node of a flow that the destination node has detected a failure on that flow (see section 1.2.2). Currently, the only failure that can be notified is the Loss of Continuity (LOC).

16.2.1 RPR continuity check (CC)

Continuity Check allows detecting per flow failures, such as one Station "stealing" the frames from another Station. It can also be used to verify connectivity in the protection path.

Continuity Check frames transmission and reception can be activated either using Activation/Deactivation frames or by configuration. In order to avoid meaningless alarms, it is required that the CC frame insertion is activated in the source station before activating the CC frame detection in the destination station.

CC frames are sent with a periodicity of nominally 1 frame per second. CC frames use the highest priority Class Of Service.

When the sink Station does not receive any CC frame within a time interval of 3.5±0.5 seconds, it will declare a Loss Of Continuity (LOC) defect. LOC shall be removed when a CC frame is detected.

Each side of the Station can have separate CC capability (ringlet specific CC). In this case the source station always sends the CC frames on the specified ringlet. The destination station will declare LOC on per ringlet basis.

It is also possible to activate CC without specifying the side (ringlet agnostic CC). In this case the source station sends the CC frames on the shortest path and steer the flow during ring failures. The destination station will declare LOC on per station basis.

It is recommended to activate CC bi-directional.

16.2.2 RPR remote defect indication (RDI)

The Station detecting a Loss Of Continuity (LOC) defect shall generate and send back to the source station of the failed flow, a RDI frame. The RDI frame shall be transmitted through the working path. The SA should be the detecting Station MAC address and the DA should be the MAC address of the Station sourcing the failed flow, the ringlet in which the LOC was detected shall be indicated in the Defect Location field.

The RDI frame shall be generated and transmitted as soon as possible after detection of the LOC defect, and shall be periodically transmitted during the defect condition. The generation frequency of the RDI frame shall be one frame per second.

The RDI frame generation shall be stopped as soon as the defect indication is removed.

The RDI frames shall be detected at the respective flow sourcing Station. The RDI state shall be declared at the RDI frame detecting Station as soon as a RDI frame is received. The RDI state is released when RDI frames are absent for 2.5±0.5 seconds.

16.2.3 RPR loopback capability

The IEEE 802.17 allows the management system to request an loopback operation to a specified destination in order to check the reachability of an RPR station.

The RPR Loopback capability allows for a frame to be inserted at one Station in the ring, and returned back by another Station through the same or opposite ringlet, without impairing the normal flow operation. Loopback frames can be activated for each Class Of Service.

The Loopback command can be sent through the shortest path (using the topology discovery), the clockwise or the counterclockwise ringlet. By default, if nothing is specified, the Loopback command is sent on the shortest path.

The Loopback source Station shall set the DA to the Loopback target MAC address and the SA to its own MAC address, and it shall set the function type to Loopback command.

The target Station shall perform the following operations:

— Set the function type to Loopback response
— Change the SA to its MAC address
— Set the DA to the original Loopback frame SA
— Change the ring ID
— Copy all other received bytes to the transmit frame
— Loopback the resulting frame according to the request filed.

The Loopback command can require the target station to reply either on the shortest path, or on the same ringlet it received the command, or on the opposite ringlet, or on the counterclockwise or on the clockwise ringlet. By default, if nothing is specified, the target station is required to reply on the shortest path.

The waiting time between the transmissions of consecutive Loopback frames on a flow shall be 5 seconds. The Loopback shall be considered unsuccessful if the Loopback frame is not returned to the source Station within 5 seconds.

16.3 RPR Activation/Deactivation of OAM

The CC mechanism is an optional mechanism and should be activated in both the source and the destination station. The activation/deactivation of the CC mechanism on both stations can be done by the management system either on both stations or on only one of them. In the former case, the coordination between the beginning of the transmission and reception of CC frames is under the responsibility of the management system. In the latter case, an Activation/Deactivation procedure is defined to coordinate the activation between the source and destination stations.

Specifically, this initialization procedure may serve the following purposes:

— To coordinate the beginning or end of the transmission and reception of CC
— To establish the type of procedure
— To specify relevant parameters (if required)

The initialization procedure may be performed either via configuration, or using the Activation/Deactivation frames.

If no response is received for an Activation/Deactivation frame within 5 seconds, the frame shall be resent. If no response is received after 3 attempts the operation shall be declared as failed.

A Station that does not support Continuity Check may respond to the relevant Activation messages with Activation Request Denied, or silently discard the Activation/Deactivation frame.

[Editor's note - The need to have the Activation/Deactivation OAM mechanism is still an open issue to discuss.]

16.4 OAM frame handling during failures

Two protection schemes are used to protect the rings: Steer and Wrap.

16.4.1 Steer protection

Ringlet specific CC frames are always sent on the required ringlet irrespective of the failure conditions (they are never steered).

During a single failure of the ring all the affected flows that are using the ringlet specific CC will stop receiving the CC frames and LOC will be declared. The Station may mask LOC since a ring failure is declared in the path used by the CC flow.

If the ringlet specific CC was not activated in the protection path the steered path will not be protected by CC.

Ringlet agnostic CC frames are always sent on the shortest path and steered during ring failures.

During a single failure of the ring, all the affected flows that are using the ringlet agnostic CC will continue receiving the CC frames from the other ringlet and LOC will not be declared.

RDI frames are always sent on the shortest path and steered during ring failures.

Loopback frames (either commands or responses) can be sent on the shortest path or on a specified ringlet. When they have to be sent on a specified ringlet, they are always sent on such a ringlet. During ring failures that affect them, they are not steered and then lost. When they have to be sent on the shortest path, they are steered during ring failures.

16.4.2 Wrap protection

During a single failure of the ring the OAM frames are wrapped, so they will reach their original destination. No LOC will be declared and the flow will remain protected by CC.

16.5 OAM frame

The OAM frame includes a common part and a function specific part. Figure36 shows the general frame format.

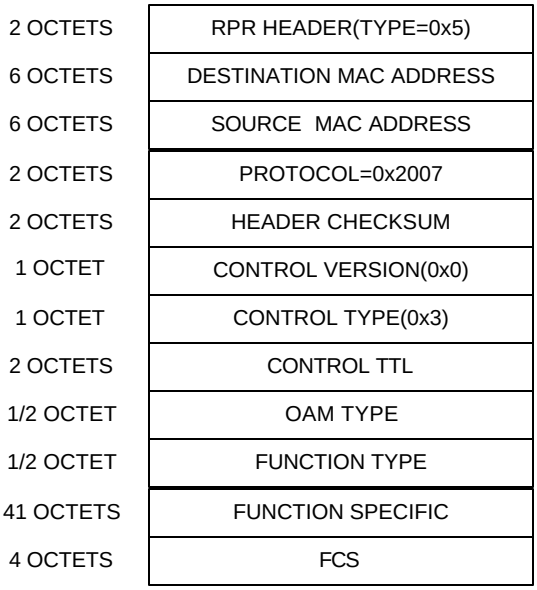


Figure 36—OAM frame format

The OAM frames length is fixed. Padding is added to provide a minimum packet length of 42 bytes.

16.5.1 OAM Class Of Service

OAM CoS is indicated in the PRI field, and depends on the OAM frame type.

The following OAM frames use the highest priority class:

— RDI
— CC
— Activation/Deactivation

Loopback frames use the CoS defined by the operation.

16.5.2 OAM type

The OAM type identifies the OAM group of the OAM frame. Table13shows the possible values of the OAM type field.

16.5.3 Function type

This field indicates the actual function performed by this frame within the group indicated by the OAM Type. Table13 shows the possible values of the Function Type field.

Table 13—OAM type field values

OAM type	Coding	Function Type	Code
Fault Management	0001	RDI	0001
		CC	0100
		LB Command	1000
		LB Response	1001
Activation/Deactivation	1000	CC	0001

16.5.4 Specific fields for OAM frames

The definition of the specific fields for the different OAM frames are provided in the sub clauses that follow.

16.5.4.1 Continuity check fault management frame

No fields are specified for the Continuity Check Fault Management frame

16.5.4.2 RDI fault management frame

The function specific fields for RDI fault management frames are illustrated in Figure 37

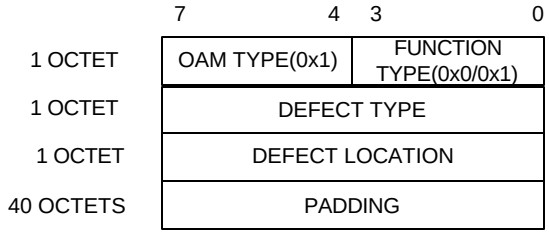


Figure 37—RDI fault management frame

16.5.4.2.1 Defect type

Optional field used to provide further information about the nature of the failure. Examples of this information are specified in Table14

Table 14—Defect Types

Defect Type	Coding
Defect not specified	11111111
Loss of Continuity (LOC) in the RPR layer.	00000000

16.5.4.2.2 Defect location

This field identifies the ringlet in which the LOC was detected. Examples of this information are specified in Table15

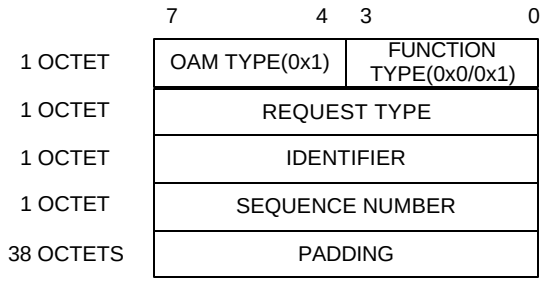
Table 15—Defect Location

Defect Location	Coding
Ringlet 0	00000000
Ringlet 1	00000001
Unspecified	11111111

16.5.4.3 Loopback frame

The function specific fields for Loopback frames are illustrated in Figure38

Figure 38—Loopback frame



16.5.4.3.1 Request type

The Request type field should be interpreted by the Loopback Station to decide through which interface the Loopback frame should be transmitted back to the source Station. Table16 shows the possible values of the Loopback type and the required action.

Table 16—Loopback Request Type values

Request Type	Action
0x00	Reply through shortest path
0x01	Reply through Inner ring
0x02	Reply through Outer ring

Table 16—Loopback Request Type values

Request Type	Action
0x03	Reply on same ring
0x04	Reply on opposite ring

16.5.4.3.2 Identifier and sequence number

An Identifier and a Sequence number are generated for each Loopback process so Stations can correlate Loopback commands with Loopback responses. The value of these fields in the looped back frame must match the value in the associated received frame. Consecutively generated Identifiers and/or Sequence numbers should be different, in order to correctly correlate commands with responses.

16.5.5 Activation/Deactivation frame

The Function Type field for the Activation/Deactivation frame will be used to identify the functions. Only Continuity Check activation/deactivation is defined, other functions may be defined in the future.

The function specific fields for the Activation/Deactivation frame are illustrated in Figure 39

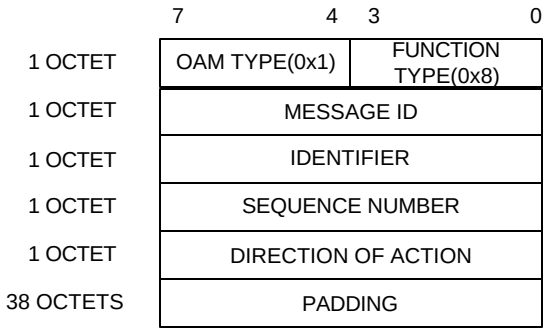


Figure 39—Activation/Deactivation frame

16.5.5.1 Message ID

This field indicates the Message ID for activating or deactivating specific OAM functions. Code values for this field are shown in Table17.

Table 17— Message ID values

Message	Command/Response	Coding
Active	Command	00000001
Activation Confirmed	Response	00000010
Activation Request Denied	Response	00000011
Deactivate	Command	00000101
Deactivation Confirmed	Response	00000110

16.5.5.2 Identifier and sequence number

An Identifier and a Sequence number are generated for each Activation/Deactivation command so Stations can correlate commands with responses. The value of these fields in the response frame must match the value in the associated command frame. Consecutively generated Identifiers and/or Sequence numbers should be different, in order to correctly correlate commands with responses.

16.5.5.3 Direction of action

This field identifies the direction, or directions, of transmission to activate/deactivate OAM functions. The East (Inner ring sink, Outer ring source) and West (Inner ring source, Outer ring sink) notation is used to differentiate between the directions of transmission. This field is used as a parameter for the Activate and Deactivate messages. This field shall be encoded as shown in Table18

Table 18—Direction values

Direction	Coding
East	00000010
West	00000001
Both	00000011
Unspecified	11111111
Not applicable	00000000

16.6 OAM frame detection procedure

OAM frames are detected through the following procedure (no specific ordering is implied):

- Check RPR header to determine if it is a Control frame of type OAM, and if it is for this Station
- Check the OAM type and Function type values according to Table 1.1 to determine the type of OAM frame received
- Silently discard OAM frames with unsupported type or Function

16.7 OAM frames support

All RPR compliant Stations should support Loopback OAM frames.

CC and RDI support is optional.

Activation/Deactivation support is optional

17. Layer Management Entity Interface

17.1 Overview of the management model

Both RPR MAC and all applicable PHY layers conceptually include management entities, called MAC sub layer management and PHY layer management entities (MLME and PLME, respectively). These entities provide the layer management service interfaces through which layer management functions may be invoked.

In order to provide correct RPR MAC operation, a station management entity (SME) must be present. The SME is a layer-independent entity that may be viewed as residing in a separate management plane. The exact functions of the SME are not specified in this standard, but in general this entity may be viewed as being responsible for such functions as the gathering of layer-dependent status from the various layer management entities, and similarly setting the value of layer-specific parameters. SME would typically perform such functions on behalf of general system management entities and would implement standard management protocols. Figure 17.1 depicts the relationship among management entities.

The management SAPs within this model are the following:

— SME-MLME SAP
— SME-PLME SAP

In this fashion, the model reflects what is anticipated to be a stackable implementation approach in which PLME functions are controlled by SME. In particular, different PHY implementations are required to have separate interfaces with the SME. The interfaces of the SME with the different PHYs are not part of this standard and are specified in the respective standard documents that specify the management primitives and MIBs for the different PHYs.

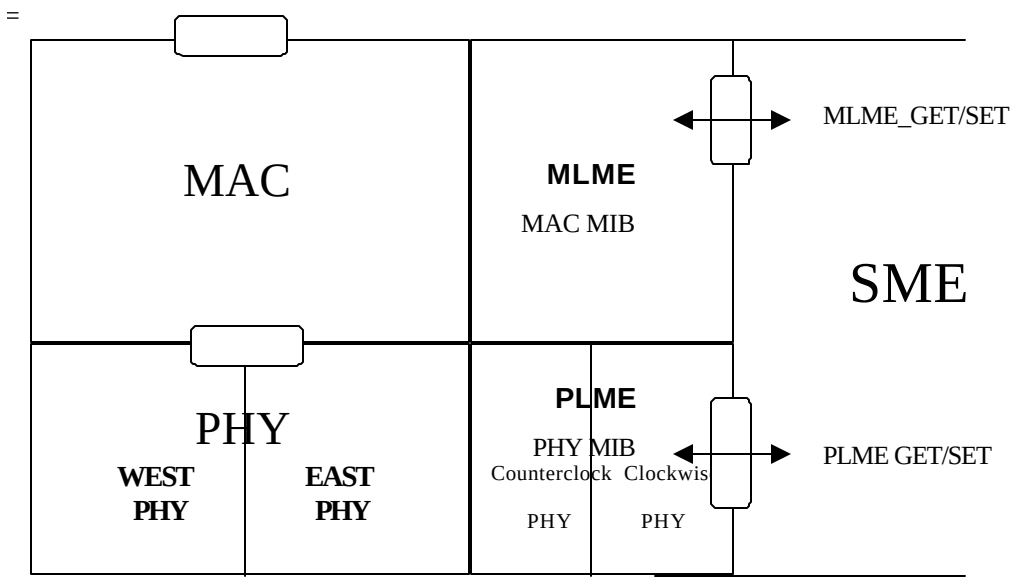


Figure 40—RPR MAC Management SAPs

17.2 Generic management primitives

The management information specific to each layer is represented as a management information base (MIB) for that layer. The MAC and PHY layer management entities are viewed as "containing" the MIB for that layer. The generic model of MIB-related management primitives exchanged across the management SAPs is to allow the SAP user-entity to either GET the value of a MIB attribute, or to SET the value of a MIB attribute. The invocation of a SET.request primitive may require that the layer entity perform certain defined actions.

The GET and SET primitives are represented as REQUESTs with associated CONFIRM primitives. These primitives are prefixed by MLME or PLME depending upon whether the MAC or PHY layer management SAP is involved.

In the following, XX denotes MLME or PLME:

```
XX-GET.request (
   MIBattribute
   )
```

Requests the value of the given MIBattribute.

```
XX-GET.confirm (
   status,
   MIBattribute,
   MIBattributevalue
   )
```

Returns the appropriate MIB attribute value if status = “success,” otherwise returns an error indication in the Status field. Possible error status values include “invalid MIB attribute”.

```
XX-SET.request (
   MIBattribute,
   MIBattributevalue
   )
```

Requests that the indicated MIB attribute be set to the given value. If this MIBattribute implies a specific action, then this requests that the action be performed.

```
XX-SET.confirm (
   status,
   MIBattribute
   )
```

If status = “success,” this confirms that the indicated MIB attribute was set to the requested value, otherwise it returns an error condition in status field. If this MIBattribute implies a specific action, then this confirms that the action was performed. Possible error status values include “invalid MIBattribute” and “attempt to set read-only MIB attribute.”

Additionally, there are certain requests (with associated confirms) that may be invoked across a given SAP that do not involve the setting or getting of a specific MIB attribute. Each SAP supports one of these, as follows:

```
XX-RESET.request: where XX is MLME or PLME as appropriate
XX-RESET.confirm
```

This service is used to initialize the management entities, the MIBs, and the data path entities. It may include a list of attributes for items to be initialized to non-default values. The corresponding.confirm indicates success or failure of the request.

Other SAP-specific primitives are in the following sections.

17.3 MLME SAP interface

The services provided by the MLME to the SME are specified in this section. These services are described in an abstract way and do not imply any particular implementation or exposed interface. MLME SAP primitives are of the general form ACTION.request followed by ACTION.confirm. The SME uses the services provided by the MLME through the MLME SAP.

According to the IETF layering principles, the RPR interface should be stacked over the two west and east span interfaces.

17.3.1 RPR interface configuration

The RPR ring interface can be activated/deactivated for administrative purposes. Its activation/deactivation allows/forbids the upper layers to send packets on the ring. It can be activated if and only if at least one of the underlying span interfaces is activated.

The RPR interface has its own operational state that goes down when both the span interfaces go down.

The RPR MAC entity requires the MAC address. This is fixed by the vendor and can be only read for maintenance purposes.

17.3.2 Topology discovery monitoring

The configuration management should allow monitoring, for maintenance purposes, the state of the auto-configuration and topology discovery protocols. It should also allow disabling the support of some features, even if supported by all the stations on the ring.

The detailed configuration requirements for the auto-configuration and topology discovery protocols depend from the actual mechanism that will be used and are now for further study.

If some misconfiguration condition is detected, a notification is sent to the management system for maintenance purposes. For example, the topology discovery mechanism can discover that two stations on the ring have the same MAC address.

17.3.3 Protection switching

The protection management mechanisms that will be defined in this standard are based on information known to the MAC entity. PHY layer protection mechanisms are independent and are addressed in the appropriate PHY specifications.

The configuration management should allow monitoring, for maintenance purposes, the state of the protection switching.

It should also allow activating/deactivating the RPR protection as well as activating/deactivating the usage of the Signal Degrade condition as switching criteria.

When the RPR protection is deactivated, the RPR MAC never switches under failure conditions. When the usage of the SD as a switching criteria is disabled, the RPR MAC never switches because of a signal degrade condition detected on the physical interface.

In order to support multi-layer protection mechanisms, a hold-off timer should be configured for each span interface, because it highly depends on the kind of physical network you have between two adjacent RPR nodes, if any.

In order to ensure robustness against unstable alarms, a wait time to restore (WTR) should be configured for each node, because there is no need to have a granular configuration of it.

The usage of the hold-off timer and of the wait to restore is specified in the protection switching section.

It should be possible to force a switching event for operational purposes.

This mechanism deals with the modes under which an RPR MAC switches to protection.

17.3.3.1 MLME-SWITCHTOPROTECTION.request

This primitive requests a change in the protection mode of an RPR ring.

The primitive parameters are as follows:

```
MLME-SWITCHTOPROTECTION.request (
         ReversionMode,
         Side
         )
```

This primitive is generated by the SME to implement a user request for a node to switch to protection at the east or the west side.

Name	Type	Valid range	Description
ReversionMode	**Enumeration**	**FORCED, AUTOMATIC-REVERTIVE, AUTOMATIC-NONREVERTIVE**	**An enumerated type that describes the desired way to revert from protection**
Side	**Enumeration**	**WEST, EAST**	**An enumerated type that describes the side of the ring relative to the station that will switch**

17.3.3.1.1 Effect of receipt

This request sets the reversion mode of the protection switch that occurs at the specified side. The MLME subsequently issues a MLME- SWITCHTOPROTECTION.confirm that reflects the results of the protection switch request.

17.3.3.2 MLME- SWITCHTOPROTECTION.confirm

This primitive confirms the change in protection switching mode.

The primitive parameters are as follows:

```
MLME- SWITCHTOPROTECTION.confirm (
         ResultCode
         )
```

Name	Type	Valid range	Description
ResultCode	Enumeration	SUCCESS, INVALID_PARAMETERS, NOT_SUPPORTED	Indicates the result of the MLME- SWITCHTOPROTEC-TION.request

This primitive is generated by the MLME as a result of an MLME- SWITCHTOPROTECTION.request by the SME. It is not generated until the switch is completed.

17.3.3.2.1 Effect of receipt

The SME is notified of the protection switch.

17.3.4 Performance and Accounting Measurements

The following statistics should be kept for the RPR interface.

1) Fairness message performance monitoring.
2) How many frames/octets have been inserted on the ring (by the upper layer)
3) How many frames/octets have been delivered to the upper layer
4) How many frames, received from the west interface, have been stripped because originated by the node itself
5) How many frames, received from the east interface, have been stripped because originated by the node itself
6) How many transit frames, received from the west interface, have been discarded because of the TTL expiration
7) How many transit frames, received from the east interface, have been discarded because of the TTL expiration
8) How many frames, received from the west interface, have been discarded because of a bad FCS (this counter is fixed to 0 when the cut-through method is implemented)
9) How many frames, received from the east interface, have been discarded because of a bad FCS (this counter is fixed to 0 when the cut-through method is implemented)
10) How many frames addressed to the node have been discarded because of a bad FCS (this counter is fixed to 0 when the store and forward method is implemented)
11) How many frames, addressed to the node, have been discarded because of an unknown or unsupported protocol.
12) How many frames addressed to the node have been discarded even if no error has been detected (e.g. because of buffer congestion). This counter has an implementation specific meaning and in some implementations it may be always equal to 0.
13) How many frames originated by the node have been discarded (e.g. because of buffer congestion). This counter has an implementation specific meaning and in some implementations it may be always equal to 0.
14) How many frames, received from the west interface, correctly passed-through the MAC entity.
15) How many frames, received from the east interface, correctly passed-through the MAC entity.
16) How many transit frames, received from the west interface, have been discarded even if no error has been detected (e.g. because of buffer congestion). This counter has an implementation specific meaning and in some implementations it may be always equal to 0.
17) How many transit frames, received from the east interface, have been discarded even if no error has been detected (e.g. because of buffer congestion). This counter has an implementation specific meaning and in some implementations it may be always equal to 0.

For the counters defined in points 1, 2, 9, 11, 12, 13, 14, 15 and 16, there is one counter for each QoS class and a global counter.

17.3.4.1 Topology Discovery statistics

The exact requirements depend on the mechanism that is used and then are for further study.

17.3.5 Notifications and Fault Management

TBD

17.3.6 RPR Ping Management

17.3.6.1 MLME-PING.request

This primitive requests the station to loopback another station on the RPR ring.

The primitive parameters are as follows:

```
MLME-PING.request (
         Addressed Station,
         Ringlet,
         Request Type,
         CoS,
         Timer
         )
```

This primitive is generated by the SME to implement a user request for a node to perform a ping operation.

Name	Type	Valid range	Description
Addressed Station	MAC Address	Any valid unicast MAC address	The MAC address of the RPR station to be ping
Ringlet	Enumeration	SHORTEST, COUNTERCLOCKWISE, CLOCKWISE	The ringlet over which the RPR ping request message should be sent (see section 16.2
Request Type	Enumeration	SHORTEST, SAME, OPPOSITE, COUNTERCLOCKWISE, CLOCKWISE	The ringlet over which the addressed station should send the RPR ping reply message (see section 16.2)
CoS	Integer	(0Ö7)	The CoS to be used in the RPR OAM frames carrying the RPR ping request and reply messages.
Timer	Integer	(0..65535)	The number of seconds the source station should wait for the reply before declaring the ping failed.

17.3.6.1.1 Effect of receipt

This requests causes the station to send an OAM RPR ping request message.

17.3.6.2 MLME- PING.confirm

This primitive notifies the success or the failure of a ping operation.

The primitive parameters are as follows:

```
MLME- PING.confirm (
         ResultCode
         )
```

Name	Type	Valid range	Description
ResultCode	Enumeration	SUCCESS, INVALID_PARAMETERS, FAILURE	Indicates the result of the MLME- PING.request

This primitive is generated by the MLME as a result of an MLME- PING.request by the SME. It is not generated until the OAM RPR ping reply frame is received or the timer expires.

17.3.6.2.1 Effect of receipt

The SME is notified of the success or of the failure of the ping procedure.

17.4 Ring Aggregation

Ring aggregation is outside of the scope of this project. If Link Aggregation for RPR Links is supported in the future then Link Aggregation Objects and a Link Aggregation MIB will be defined. For this purpose the IEEE 802.3 link aggregation objects defined in the LAG-MIB (IEEE 802.3 – Annex 30C) might serve as a guide.

Link aggregation can be managed as an interface stacked over multiple RPR interfaces

A request for a new ifType for the aggregation should be forwarded to the IANA

17.5 PLME SAP interface

Each span interface can be activated/deactivated separately for administrative purposes. Its activation/deactivation allows/forbids the MAC layer to send packets on that span.

In order to avoid inconsistent configurations, a span interface can be deactivated only if the RPR interface has been deactivated or the other span interface is still active.

Each span interface has its own operational state that can be read for maintenance purposes.

The management of Sonet/SDH PHY as well as of Ethernet physical interfaces are already defined in the relevant standard recommendations. The IEEE 802.17 will reuse the already defined PLME SAP primitives to manage the Sonet/SDH and Ethernet PHY layers.

The number frames that are received or transmitted on each span interface are counted as span interface statistics representing the number of frames that the span interface has delivered or has received from the upper layer.

17.5.1 The Ethernet PHY

For the Ethernet PHY we will use the 802.3ae LAN and WAN PHY primitives. Also we will use the MIB objects that are defined in the MAU-MIB (RFC 2668) for the Ethernet PHY objects and the updated MIB in preparation (draft-ietf-hubmib-mau-mib-v3-00) for management of 10 Gb/s LAN PHY. In addition we will use the MIB objects defined in (draft-ietf-hubmib-wis-mib-00) to manage the10 Gb/s WAN PHY – called ETHER-WIS-MIB.

The Ethernet Reconciliation Sublayer management should be defined in the IEEE 802.17 specification.

According to the IETF layering principles, the RPR interface, when working over Ethernet interfaces, should be stacked over the west and east “Ethernet” interfaces.

An Ethernet span interface can be activated at any time, because it is the lowest level interface on the system. Its operational state goes down when the media is unavailable.

17.5.2 The SONET PHY

Sonet/SDH interfaces are layered interfaces and are managed as a set of stacked interfaces. The physical medium, the section and the line layers are managed as a single layered interface. The recommended IANA ifType is sonet (39). The Path layer is managed as a stacked interface over the Medium/Section/Line interface. The recommended IANA ifType is sonetPath (50). For the SONET PHY the SONET/SDH PHY objects are defined in the SONET-MIB (RFC 2558).

17.5.2.1 The GFP Adaptation Layer

An additional GFP interface might be stacked over the Path interface to represent the GFP adaptation layer.

The GFP management is outside the scope of IEEE 802.17.

According to the IETF layering principles, the RPR interface, when working over Sonet/SDH interfaces, should be stacked over the west and east GFP interfaces.

A GFP span interface can be activated only if the underlying Sonet/SDH interface is activated. Its operational state goes down when a signal fail condition is detected on the Sonet/SDH path or if there is a payload mismatch (the value in the received C2 byte is different than the GFP code, i.e. [TBD] value) or there is a loss of frame alignment.

[Editor’s note – The signal label value to be put in the C2 byte of Sonet/SDH interfaces has not yet been assigned by ITU-T.]

Annex ABibliography

[B1] IEEE 802.3 - 2000 Edition Carrier sense multiple access with collision detection (CSMA/CD) MAC and physical layer specification

[B2] ITU-T G.707 Network Node Interface for the Synchronous Digital Hierarchy (SDH).

[B3] ITU-T G.774 Series SDH Management Information Model for the Network Element View

[B4] ITU-T G.783 Characteristics of Synchronous Digital Hierarchy (SDH) Equipment Functional Blocks

[B5] ITU-T G.7041 Generic Framing Procedure

[B6] RFC 2615 PPP over SONET/SDH

[B7] RFC 1471 The Definitions of Managed Objects for the Link Control Protocol of the Point-to-Point Protocol

[B8] RFC 1661 The Point-to-point Protocol (PPP)

[B9] RFC 1662 PPP in HDLC-like Framing

[B10] RFC 2863 The Interfaces Group MIB

Annex BTransmit Clock Sync

B.1 RPR synchronization

Each node operates in “free-run” mode. That is, the receive clock is derived from the incoming receive stream while the transmit clock is derived from a local oscillator. This eliminates the need for expensive clock synchronization as required in existing SONET networks. Differences in clock frequency are accommodated by inserting a small amount of idle bandwidth at each node’s output.

The clock source for the transmit clock shall be selected to deviate by no more than 200 ppm from the center frequency. The overall outgoing rate of the node shall be rate shaped to accommodate the worst case difference between receive and transmit clocks of adjacent nodes. This is accomplished by monitoring the input data rate (from the line and the MAC client), and comparing that to the output data rate. If the rates differ, it can be assumed that there are differences between the clocks, and the output data rate can be adjusted appropriately.

Annex CTen Gigabit Ethernet Resolutions Sub-layer (RS) and Ten Gigabit Media Independent Interface (XGMII)

Note: Text is largely borrowed from P802.3ae/D3.1, and needs to be cleaned up. Need to insert figures. XGMII remains unchanged, RS has been modified for RPR.

C.1 Overview

This clause defines the logical and electrical characteristics for the Reconciliation Sublayer (RS) and 10 Gigabit Media Independent Interface (XGMII) between the RPR media access control (MAC) layer and various 10 Gigabit Ethernet PHYs. Figure 2 shows the relationship of the RS and XGMII to the ISO (IEEE) OSI reference model.

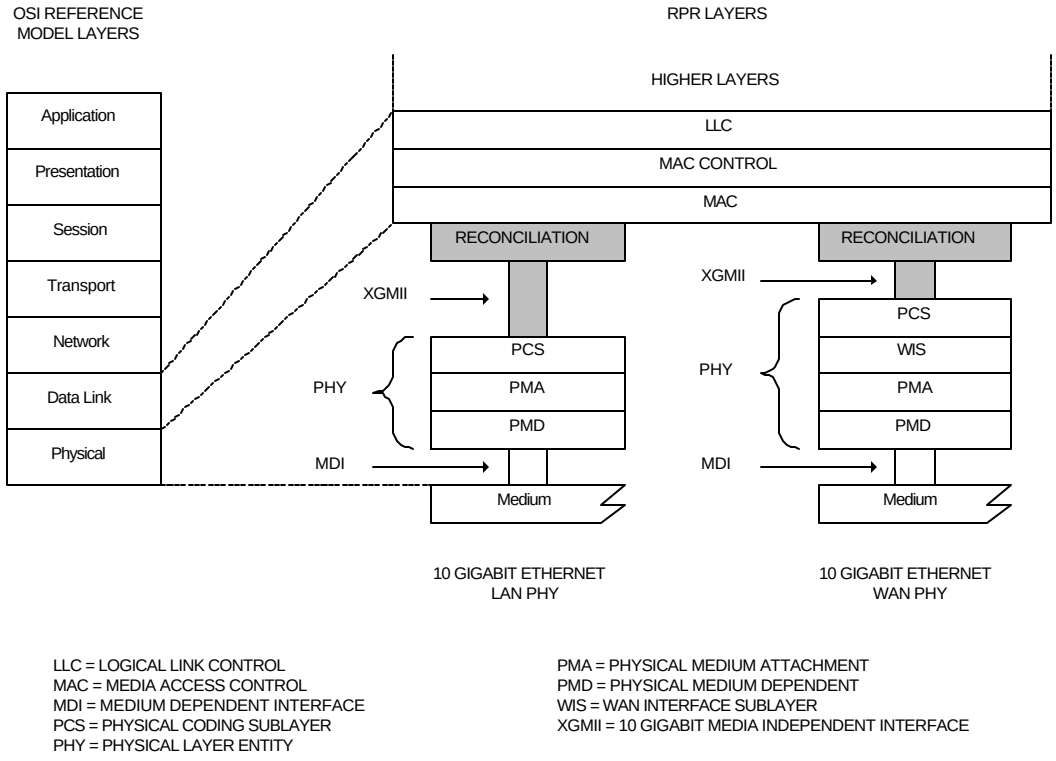


Figure A.1—RS and XGMII location in the OSI protocol stack

The purpose of the RS is to convert the logical P-SAP interface of the MAC to the XGMII interface that is specified by P802.3ae as an interface to 10 Gigabit PHYs. Though the XGMII is an optional interface, it is used extensively in this standard as a basis for specification. The 10 Gb/s Physical Coding Sublayer (PCS) is specified to the XGMII interface, so if not implemented, a conforming implementation shall behave functionally as if the RS and XGMII were implemented.

The purpose of the 10 Gigabit Media Independent Interface (XGMII) is to provide a simple, inexpensive, and easy-to-implement interconnection between the Media Access Control (MAC) sublayer and the Physical layer (PHY). The 10 Gigabit Attachment Unit Interface (XAUI) may optionally be used to extend the operational distance of the XGMII with reduced pin count.

The RS adapts the logical P-SAP interface of the MAC to the parallel encoding of 10 Gb/s PHYs. Though the XGMII is an optional interface, it is used extensively in this standard as a basis for specification. The 10 Gb/s Physical Coding Sublayer (PCS) is specified to the XGMII interface, so if not implemented, a conforming implementation shall behave functionally as if the RS and XGMII were implemented.

The XGMII has the following characteristics:

a) It is capable of supporting 10 Gb/s operation;
b) Data and delimiter are synchronous to clock reference;
c) It provides independent 32-bit-wide transmit and receive data paths;
d) It uses signal levels compatible with common digital ASIC processes;
e) It provides for full-duplex operation only

C.1.1 Summary of major concepts

a) The RS converts the P-SAP service primitives to the parallel data paths of the XGMII;
b) The RS maps the signal set provided at the XGMII to the P-SAP service primitives provided at the MAC;
c) Each direction of data transfer is independent and serviced by data, control, and clock signals;
d) The RS generates continuous data or control characters on the transmit path and expects continuous data or control characters on the receive path;
e) The RS participates in link fault detection and reporting by monitoring the receive path for status reports that indicate an unreliable link;
f) When the XGMII is optionally extended with XAUI, two XGMII interfaces logically exist (see Figure 46–1). When extended, signals on the transmit path are from the RS to an XGXS of the XAUI interface on one XGMII and from the XGXS to the PHY on the other XGMII. When extended, signals on the receive path are from the PHY to an XGXS of the XAUI interface on one XGMII and from the XGXS to the RS on the other XGMII. The descriptions of the XGMII as between the RS and the PHY are therefore equally applicable between the RS and the XGXS or the XGXS and the PHY.

C.1.2 Application

This clause applies to the interface between the MAC and PHY. The physical implementation of the interface is primarily intended as a chip-to-chip (integrated circuit to integrated circuit) interface implemented with traces on a printed circuit board. The XGMII may also be used in other ways, for example, as a logical interface between ASIC logic modules within an integrated circuit.

This interface is used to provide media independence so that an identical media access controller may be used with all 10GBASE PHY types using either serial or wavelength division multiplexed optics.

C.1.3 Rate of operation

The XGMII supports only the 10 Gb/s data rate as defined within this clause.

C.1.4 Delay Constraints

Predictable operation of the MAC Control PAUSE operation requires that there be an upper bound on the propagation delays through the network. This implies that MAC, MAC Control sub-layer, and PHY implementors must conform to certain delay maxima, and that network planners and administrators conform to constraints regarding the cable topology and concatenation of devices. Table 46–1 contains the values of

maximum cumulative MAC Control, MAC and RS round-trip (sum of transmit and receive) delay in bit time as specified in 1.4 and pause quanta as specified in 31B.2.

C.1.5 Allocation of functions

The allocation of functions at the XGMII balances the need for media independence with the need for a simple and cost-effective interface. The bus width and signaling rate are applicable to short distance chip-to-chip interconnect with printed circuit board trace lengths electrically limited to approximately 7 cm. The XGMII maximizes media independence by cleanly separating the Data Link and Physical Layers of the ISO (IEEE) seven-layer reference model.

C.1.6 XGMII structure

The XGMII is composed of independent transmit and receive paths. Each direction uses 32 data signals (TXD<31:0> and RXD<31:0>), four control signals (TXC<3:0> and RXC<3:0>) and a clock (TX_CLK and RX_CLK). Figure 46–2 depicts a schematic view of the RS inputs and outputs.

The 32 TXD and four TXC signals shall be organized into four data lanes, as are the 32 RXD and four RXC signals (see Table 46–2). The four lanes in each direction share a common clock — TX_CLK for transmit and RX_CLK for receive. The four lanes are used in round-robin sequence to carry an octet stream. On transmit, each eight PHY_DATA.request transactions represent an octet transmitted by the MAC. The first octet is aligned to lane 0, the second to lane 1, the third to lane 2 the fourth to lane 3, then repeating with the fifth to lane 0, etc. Delimiter and interframe idle characters are encoded on the TXD and RXD signals with the control code indicated by assertion of TXC and RXC respectively.

C.1.7 Mapping of XGMII signals to P-SAP service primitives

The Reconciliation Sublayer (RS) shall map the signals provided at the XGMII to the P-SAP service primitives as shown in Figure 2. The following P-SAP service primitives are defined:

— PHY_DATA.request
— PHY_DATA.indicate
— PHY_DATA_VALID.indicate
— PHY_LINK_STATUS.indicate
— PHY_READY.indicate

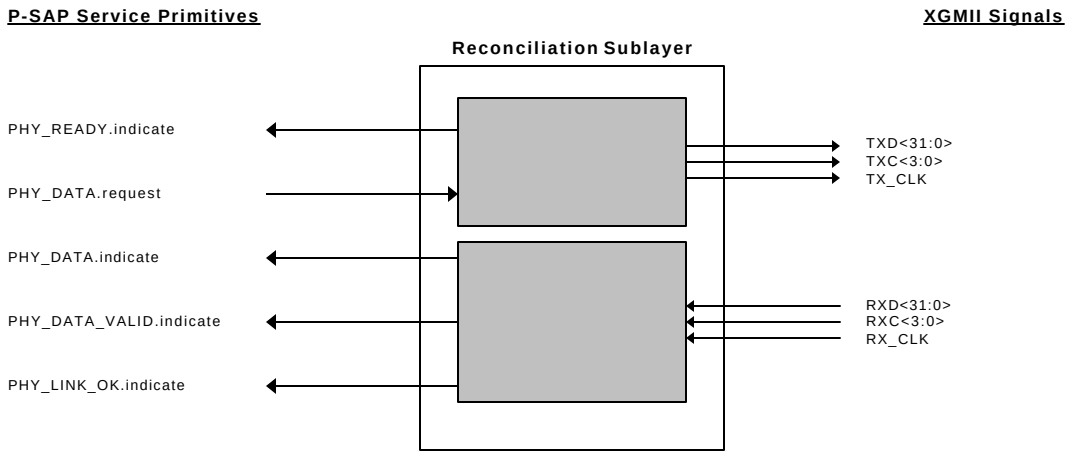


Figure A.2—Reconciliation Sublayer (RS) inputs and outputs

C.1.7.1 Mapping of PHY_DATA.request

C.1.7.1.1 Function

Map the primitive PHY_DATA.request to the XGMII signals TXD<31:0>, TXC<3:0> and TX_CLK.

C.1.7.1.2 Semantics of the service primitive

`PHY_DATA.request (OUTPUT_UNIT)`

The OUTPUT_UNIT parameter either takes the value of an octet of data or DATA_COMPLETE, and represents the transfer of an octet of data from the MAC to the RS. The value DATA_COMPLETE indicates that the MAC has no more data to transfer.

C.1.7.1.3 When generated

This primitive is generated by the MAC sublayer to request the transmission of a data octet on the physical medium, or to indicate that no more data is available for transmission.

C.1.7.1.4 Effect of receipt

The OUTPUT_UNIT values are conveyed to the PHY by the signals TXD<31:0> and TXC<3:0> on each TX_CLK edge. Each PHY_DATA.request transaction shall be mapped to a TXD signal in sequence (TXD<0:7>,... TXD<24:31>, TXD<0:7>) as described in xxx. After four PHY_DATA.request transactions from the MAC sublayer, the RS requests transmission of 32 data bits by the PHY. The first eight octets of the frame shall be converted as follows:

a) The first octet is converted to a Start control character and aligned to lane 0;
b) The following six octets are converted to Preamble characters;
c) The eighth character is converted to a SFD character.

The MAC does not generate interframe spacing—the RS generates IPG between MAC frames. The DATA_COMPLETE value shall initiate IPG generation as follows:

a) The DATA_COMPLETE value is mapped to a Terminate control character;
b) Interframe spacing is generated in accordance with IEEE Std 802.3 Clause 4, and mapped to IPG.

Figure A.3—Transmit and receive lane associations

C.1.7.2 Mapping of PHY_DATA.indicate

C.1.7.2.1 Function

Map the primitive PHY_DATA.indicate to the XGMII signals RXD<31:0>, RXC<3:0> and RX_CLK.

C.1.7.2.2 Semantics of the service primitive

`PHY_DATA.indicate (INPUT_UNIT)`

The INPUT_UNIT parameter takes the value of an octet of data, and defines the transfer of an octet of data from the RS to the MAC.

C.1.7.2.3 When generated

The INPUT_UNIT values are derived from the signals RXC<3:0> and RXD<31:0> received from the PHY on each edge of RX_CLK. Each primitive generated to the MAC sublayer entity corresponds to a PHY_DATA.request issued by the MAC at the remote end of the link connecting two DTEs. For each RXD<31:0> during frame reception, the RS shall generate four PHY_DATA.indicate transactions until the end of frame (Terminate control character), where one, two, three, or four PHY_DATA.indicate transactions will be generated from the RXD<31:0> containing the Terminate. During frame reception, each RXD signal

shall be mapped in sequence into a PHY_DATA.indicate transaction (RXD<0:7>, ... RXD<24:31>, RXD<0:7>) as described in xxx.

The RS shall convert a valid Start control character and the subsequent eight octets of data to null data prior to generation of the associated PHY_DATA.indicate transactions. The RS shall not generate any PHY_DATA.indicate primitives for a Terminate control character. To assure robust operation, the value of the data transferred to the MAC may be changed by the RS as required by XGMII error indications (see 46.3.3). Sequence ordered_sets are not indicated to the MAC (see 46.3.4).

C.1.7.2.4 Effect of receipt

The effect of receipt of this primitive by the MAC sublayer is unspecified.

C.1.7.3 Mapping of PHY_DATA_VALID.indicate

C.1.7.3.1 Function

Map the primitive PHY_DATA_VALID.indicate to the XGMII signals RXC<3:0> and RXD<31:0>.

C.1.7.3.2 Semantics of the service primitive

```
PHY_DATA_VALID.indicate (DATA_VALID_STATUS)
```

The DATA_VALID_STATUS parameter can take one of two values: DATA_VALID or DATA_NOT_VALID. The DATA_VALID value indicates that the INPUT_UNIT parameter of the PHY_DATA.indicate primitive contains a valid data of an incoming frame. The DATA_NOT_VALID value indicates that the INPUT_UNIT parameter of the PHY_DATA.indicate primitive does not contain valid data of an incoming frame.

C.1.7.3.3 When generated

The PHY_DATA_VALID.indicate service primitive shall be generated by the RS whenever the DATA_VALID_STATUS parameter changes from DATA_VALID to DATA_NOT_VALID or vice versa. DATA_VALID_STATUS shall assume the value DATA_VALID when a PHY_DATA.indicate transaction is generated in response to reception of a Start control character on lane 0 if the prior RXC<3:0> and RXD<31:0> contained four Idle characters or a Sequence ordered set. DATA_VALID_STATUS shall assume the value DATA_NOT_VALID when RXC of the current lane in sequence is asserted for anything except an Error control character. In the absence of errors, DATA_NOT_VALID is caused by a Terminate control character. When DATA_VALID_STATUS changes from DATA_VALID to DATA_NOT_VALID because of a control character other than Terminate, the RS shall ensure that the MAC will detect a FrameCheckError prior to indicating DATA_NOT_VALID to the MAC (see xxx).

C.1.7.3.4 Effect of receipt

The effect of receipt of this primitive by the MAC sublayer is unspecified.

C.1.7.4 Mapping of PHY_LINK_STATUS.indicate

C.1.7.4.1 Function

Map the primitive PHY_LINK_STATUS.indicate to the XGMII signals RXD<31:0>, RXC<3:0> and RX_CLK.

C.1.7.4.2 Semantics of the service primitive

```
PHY_LINK_STATUS.indicate (LINK_STATUS)
```

The LINK_STATUS parameter takes the value of OK, FAIL, or DEGRADE, and signifies the status of the PHY as indicated by Link Fault Signaling. OK indicates that no Local Fault signal has been received. FAIL indicates that a Local Fault signal has been received. DEGRADE is not generated by the RS.

C.1.7.4.3 When generated

The LINK_STATUS values are derived from the signals RXC<3:0> and RXD<31:0> received from the PHY on each edge of RX_CLK. Details TBD.

C.1.7.4.4 Effect of receipt

The effect of receipt of this primitive by the MAC sublayer is unspecified.

C.1.7.5 Mapping of PHY_READY.indicate

C.1.7.5.1 Function

Map the primitive PHY_READY.indicate to the XGMII signals RXD<31:0>, RXC<3:0> and RX_CLK.

C.1.7.5.2 Semantics of the service primitive

```
PHY_READY.indicate (READY_STATUS)
```

The READY_STATUS parameter takes the value of READY or NOT_READY, and signifies the status of a transmitted frame. If a frame is being received from the MAC and has not been transmitted, READY_STATUS indicates NOT_READY. If a frame has been received and fully transmitted, READY_STATUS indicates READY signifying that a new frame can be received.

C.1.7.5.3 When generated

The LINK_STATUS values is derived from the RS transmit state machine. Details TBD.

C.1.7.5.4 Effect of receipt

The effect of receipt of this primitive by the MAC sublayer is unspecified.

C.2 XGMII data stream

Data frames transmitted through the XGMII shall be transferred within the XGMII data stream. The data stream is a sequence of bytes, where each byte conveys either a data octet or control character. The parts of the data stream are shown in Figure A.4—.

<inter-frame><preamble><sfd><data><efd>

Figure A.4—XGMII data stream

For the XGMII, transmission and reception of each bit and mapping of data octets to lanes shall be as shown in Figure A.5—.

C.2.1 Inter-frame <inter-frame>

The inter-frame <inter-frame> period on an XGMII transmit or receive path is an interval during which no frame data activity occurs. The interpacket gap is generated by the RS. The <inter-frame> begins with the

Terminate control character, continues with Idle control characters and ends with the Idle control character prior to a Start control character. The length of the interpacket gap may be changed between the transmitting RS and receiving RS by one or more functions (e.g., RS lane alignment, PHY clock rate compensation or 10GBASE-W data rate adaptation functions). The minimum IPG at the XGMII of the receiving RS is five octets.

The signaling of link status information logically occurs in the <inter-frame> period (see xxx).

C.2.2 Preamble <preamble> and start of frame delimiter <sfd>

The preamble <preamble> begins a frame transmission by a MAC as specified in 4.2.5 and when generated by a MAC consists of 7 octets with the following bit values:

```
10101010 10101010 10101010 10101010 10101010 10101010 10101010
```

The Start control character indicates the beginning of MAC data on the XGMII. On transmit, the RS converts the first data octet of preamble transferred from the MAC into a Start control character. On receive, the RS will convert the Start control character into a preamble data octet. The start control character is aligned to lane 0 of the XGMII by the RS on transmit and by the PHY on receive.

The start of frame delimiter <sfd> indicates the start of a frame and immediately follows the preamble. The bit value of <sfd> at the XGMII is unchanged from the Start Frame Delimiter (SFD) specified in 4.2.6 and is the bit sequence:

```
10101011
```

Figure A.5—Relationship of data lanes to MAC serial bit stream

The preamble and SFD are shown above with their bits ordered for serial transmission from left to right. As shown, the left-most bit of each octet is the LSB of the octet and the right-most bit of each octet is the MSB of the octet.

The preamble and SFD are transmitted through the XGMII as octets sequentially ordered on the lanes of the XGMII. The first preamble octet is replaced with a Start control character and it is aligned to lane 0, the second octet on lane 1, the third on lane 2 and the fourth on lane 3, and the four octets are transferred on the next edge of TX_CLK. The fifth octet is assigned to lane 0 with subsequent octets sequentially assigned to the lanes with the SFD assigned to lane 3. The XGMII <preamble> and <sfd> are:

Lane 0	**Lane 1**	**Lane 2**	**Lane 3**
Start	10101010	10101010	10101010
10101010	10101010	10101010	10101011

C.2.3 Data <data>

The data <data> in a well-formed frame shall consist of a set of data octets.

C.2.4 End of frame delimiter <efd>

Assertion of TXC with the appropriate Terminate control character encoding of TXD on a lane constitutes an end of frame delimiter <efd> for the transmit data stream. Similarly, assertion of RXC with the appropriate Terminate control character encoding of RXD constitutes an end of frame delimiter for the receive data stream. The XGMII shall recognize the end of frame delimiter on any of the four lanes of the XGMII.

C.2.5 Definition of Start of Packet and End of Packet Delimiter

For the purposes of Clause 30 layer management the Start of Packet delimiter is defined as the Start control character; and the End of Packet delimiter is defined as the end of the last sequential data octet preceding the Terminate control character or other control character causing a change from DATA_VALID to DATA_NOT_VALID. (See 46.1.7.5.2 and 30.3.2.1.5.)

C.3 XGMII functional specifications

The XGMII is designed to make the differences among the various media and transceiver combinations transparent to the MAC sublayer. The selection of logical control signals and the functional procedures are all designed to this end.

C.3.1 Transmit

C.3.1.1 TX_CLK (10 Gb/s transmit clock)

TX_CLK is a continuous clock used for operation at 10 Gb/s. TX_CLK provides the timing reference for the transfer of the TXC<3:0> and TXD<31:0> signals from the RS to the PHY. The values of TXC<3:0> and TXD<31:0> shall be sampled by the PHY on both the rising edge and falling edge of TX_CLK. TX_CLK is sourced by the RS.

The TX_CLK frequency shall be 156.25MHz ±0.01%, one-sixty-fourth of the MAC transmit data rate.

C.3.1.2 TXC<3:0> (transmit control)

TXC<3:0> indicate that the RS is presenting either data or control characters on the XGMII for transmission. The TXC signal for a lane shall be de-asserted when a data octet is being sent on the corresponding lane and asserted when a control character is being sent. In the absence of errors, the TXC signals are deserted by the RS for each octet of the preamble (except the first octet that is replaced with a Start control character) and remain de-asserted while all octets to be transmitted are presented on the lanes of the XGMII. TXC<3:0> are driven by the RS and shall transition synchronously with respect to both the rising and falling edges of TX_CLK. Table 46–3 specifies the permissible encoding of TXD and TXC for a XGMII transmit lane. Additional requirements apply for proper code sequences and in which lanes particular codes are valid (e.g., Start control character is to be aligned to lane 0).

Table A.1—Permissible encoding of TXC, and TXD

TXC	TXD	Description	PHY_DATA.request parameter
0	00 through FF	Normal data transmission	ZERO, ONE (eight bits)
1	00 through 06	Reserved	--
1	07	Idle	No applicable parameter (Normal inter-frame)
1	08 through 9B	Reserved	--
1	9C	Sequence (only valid in lane 0)	No applicable parameter (Inter-frame status signal)
1	9D through FA	Reserved	--
1	FB	Start (only valid in lane 0)	No applicable parameter, replaces first eight ZERO, ONE of a frame (preamble octet)
1	FC	Reserved	--
1	FD	Terminate	DATA_COMPLETE
1	FE	Transmit error propagation	No applicable parameter
1	FF	Reserved	--
Note—Values in TXD column are in hexadecimal, most significant bit to least significant bit (i.e., <7:0>).			

C.3.1.3 TXD<31:0> (transmit data)

TXD is a bundle of 32 data signals organized into four lanes of eight signals each (TXD<7:0>, TXD<15:8>, TXD<23:16>, and TXD<31:24>) that are driven by the RS. Each lane is associated with a TXC signal as shown in Table 46–2 and shall be encoded as shown in Table 46–3. TXD<31:0> shall transition synchronously with respect to both the rising and falling edges of TX_CLK. For each high or low TX_CLK transition, data and/or control are presented on TXD<31:0> to the PHY for transmission. TXD<0> is the least significant bit of lane 0, TXD<8> the least significant bit of lane 1, TXD<16> the least significant bit of lane 2 and TXD<24> the least significant bit of lane 3.

Assertion on a lane of appropriate TXD values when TXC is asserted will cause the PHY to generate code-groups associated with either Idle, Start, Terminate, Sequence or Error control characters. While the TXC of a lane is de-asserted, TXD of the lane is used to request the PHY to generate code-groups corresponding to the data octet value of TXD. An example of normal frame transmission is illustrated in Figure A.6—.

Figure A.7— shows the behavior of TXD and TXC during an example transmission of a frame propagating an error.

Figure A.6—Normal frame transmission

Figure A.7—Transmit Error Propagation

C.3.1.4 Start control character alignment

On transmit, it may be necessary for the RS to modify the length of the <inter-frame> in order to align the Start control character (first octet of preamble) on lane 0. This shall be accomplished in one of two ways.

— A MAC implementation may incorporate this RS function into its design and always insert additional idle characters to align the start of preamble on a four byte boundary. Note that this will reduce the effective data rate for certain packet sizes separated with minimum inter-frame spacing.
— Alternatively, the RS may maintain the effective data rate by sometimes inserting and sometimes deleting idle characters to align the Start control character. When using this method the RS must maintain a Deficit Idle Count (DIC) that represents the cumulative count of idle characters deleted or inserted. The DIC is increments for each idle character deleted, decremented for each idle character inserted, and the decision of whether to insert or delete idle characters is constrained by bounding the DIC to a minimum value of zero and maximum value of three. Note that this may result in inter-frame spacing observed on the transmit XGMII that is up to three octets shorter than the minimum transmitted preamble specified in Clause 4, however the frequency of shortened inter-frame spacing is constrained by the DIC rules. The Deficit Idle Count is only reset at initialization and is applied regardless of the size of the IPG transmitted by the MAC sublayer. An equivalent technique may be employed to control RS alignment of the Start control character provided that the result is the same as if the RS implemented DIC as described.

C.3.2 Receive

C.3.2.1 RX_CLK (receive clock)

RX_CLK is a continuous clock which provides the timing reference for the transfer of the RXC<3:0> and RXD<31:0> signals from the PHY to the RS. RXC<3:0> and RXD<31:0> shall be sampled by the RS on both the rising and falling edge of RX_CLK. RX_CLK is sourced by the PHY.

The frequency of RX_CLK may be derived from the received data or it may be that of a nominal clock (e.g., TX_CLK). When the received data rate at the PHY is within tolerance, the RX_CLK frequency shall be 156.25MHz ±0.01%, one-sixty-fourth of the MAC receive data rate.

There is no need to transition between the recovered clock reference and a nominal clock reference on a frame-by-frame basis. If loss of received signal from the medium causes a PHY to lose the recovered RX_CLK reference, the PHY shall source the RX_CLK from a nominal clock reference. Transitions from nominal clock to recovered clock or from recovered clock to nominal clock shall not decrease the time between adjacent edges of RX_CLK.

NOTE—This standard neither requires nor assumes a guaranteed phase relationship between the RX_CLK and TX_CLK signals.

C.3.2.2 RXC<3:0> (receive control)

RXC<3:0> indicate that the PHY is presenting either recovered and decoded data or control characters on the XGMII. The RXC signal for a lane shall be de-asserted when a data octet is being received on the corresponding lane and asserted when a control character is being received. In the absence of errors, the RXC signals are de-asserted by the PHY for each octet of the preamble (except the first octet that is replaced with a Start control character) and remain de-asserted while all octets to be received are presented on the lanes of the XGMII. RXC<3:0> are driven by the PHY and shall transition synchronously with respect to both the rising and falling edges of RX_CLK. Table A.2— specifies the permissible encoding of RXD and RXC for a XGMII receive lane. Additional requirements apply for proper code sequences and in which lanes particular codes are valid (e.g., Start control character is to be aligned to lane 0).

Figure 46–7 shows the behavior of RXC<3:0> during an example frame reception with no errors.

Table A.2—Permissible encoding of RXC, and RXD

TXC	TXD	Description	PHY_DATA.request parameter
0	00 through FF	Normal data transmission	ZERO, ONE (eight bits)
1	00 through 06	Reserved	--
1	07	Idle	No applicable parameter (Normal inter-frame)
1	08 through 9B	Reserved	--
1	9C	Sequence (only valid in lane 0)	No applicable parameter (Inter-frame status signal)
1	9D through FA	Reserved	--
1	FB	Start (only valid in lane 0)	No applicable parameter, replaces first eight ZERO, ONE of a frame (preamble octet)
1	FC	Reserved	--
1	FD	Terminate	No applicable parameter (Start of inter-frame)
1	FE	Receive error	No applicable parameter

1	FF	Reserved	--
Note—Values in RXD column are in hexadecimal, most significant bit to least significant bit (i.e., <7:0>).			

Figure A.8—Basic Frame Reception

C.3.2.3 RXD (receive data)

RXD is a bundle of 32 data signals (RXD<31:0>) organized into four lanes of eight signals each (RXD<7:0>, RXD<15:8>, RXD<23:16>, and RXD<31:24>) that are driven by the PHY. Each lane is associated with a RXC signal as shown in Table 46–2 and shall be encoded as shown in Table 46–4. RXD<31:0> shall transition synchronously with respect to both the rising and falling edges of RX_CLK. For each high or low RX_CLK transition, received data and/or control are presented on RXD<31:0> for mapping by the RS. RXD<0> is the least significant bit of lane 0, RXD<8> the least significant bit of lane 1, RXD<16> the least significant bit of lane 2 and RXD<24> the least significant bit of lane 3. Figure 46–7 shows the behavior of RXD<31:0> during frame reception.

While the RXC of a lane is de-asserted, RXD of the lane is used by the RS to generate PHY_DATA.indicate transactions. Assertion on a lane of appropriate RXD values when RXC is asserted indicates to the RS the Start control character, Terminate control character, Sequence control character or Error control character that drive its mapping functions.

RXC of a lane is asserted with the appropriate Error control character encoding on RXD of the lane to indicate an error was detected somewhere in the frame presently being transferred from the PHY to the RS (e.g., a coding error, or any error that the PHY is capable of detecting, and that may otherwise be undetectable at the MAC sublayer).

The effect of an Error control character on the RS is defined in 46.3.3.1.Figure A.9—shows the behavior of RXC and RXD during the reception of an example frame with an error.

Figure A.9—Reception with error

C.3.3 Error and fault handling

C.3.3.1 Response to error indications by the XGMII

If, during frame reception (i.e., when DATA_VALID_STATUS = DATA_VALID), a control character other than a Terminate control character is signaled on a received lane, the RS shall ensure that the MAC will detect a FrameCheckError in that frame. This requirement may be met by incorporating a function in the RS that produces a received frame data sequence delivered to the MAC sublayer that is guaranteed to not yield a valid CRC result, as specified by the frame check sequence algorithm (see 3.2.8). This data sequence may be produced by substituting data delivered to the MAC. The RS generates eight PHY_DATA.indicate primitives for each Error control character received within a frame, and may generate eight PHY_DATA.indicate primitives to ensure FrameCheckError when a control character other than Terminate causes the end of the frame.

Other techniques may be employed to respond to a received Error control character provided that the result is that the MAC sublayer behaves as though a FrameCheckError occurred in the received frame.

C.3.3.2 Conditions for generation of transmit Error control characters

If, during the process of transmitting a frame, it is necessary to request that the PHY deliberately corrupt the contents of the frame in such a manner that a receiver will detect the corruption with the highest degree of probability, then an Error control character may be asserted on a transmit lane by the appropriate encoding of the lane's TXD and TXC signals.

C.3.3.3 Response to indication of invalid frame sequences

The 10 Gb/s PCS is required to align the Start control character to lane 0. The RS shall not indicate DATA_VALID to the MAC for a Start control character received on any other lane. Error free 10 Gb/s operation will not change the SFD alignment in lane 3. A 10 Gb/s MAC/RS implementation is not required to process a packet that has an SFD in a position other than lane 3 of the column following the column containing the Start control character.

C.3.4 Link fault signaling

Sublayers within the PHY are capable of detecting faults that render a link unreliable for communication. Upon recognition of a fault condition a PHY sublayer indicates Local Fault status on the data path. When this Local Fault status reaches an RS, the RS indicates the fault status to the MAC through the PHY_LINK_STATUS.indicate service primitive. When the RS no longer receives fault status messages, it returns to normal operation, it signals normal status to the MAC.

Status is signaled in a four byte Sequence ordered_set as shown in Table A.3—. The PHY indicates Local Fault with a Sequence control character in lane 0 and data characters of 0x00 in lanes 1 and 2 plus a data character of 0x01 in lane 3. Remote Fault is not used. Though most fault detection is on the receive data path of a PHY, in some specific sublayers, faults can be detected on the transmit side of the PHY. This is also indicated by the PHY with a Local Fault status.

Table A.3—Sequence ordered_sets

Lane 0	Lane 1	Lane 2	Lane 3	Description
Sequence	0x00	0x00	0x00	Reserved

Sequence	0x00	0x00	0x01	Local Fault
Sequence	0x00	0x00	0x02	Remote Fault, Not Used
Sequence	≥0x00	≥0x00	≥0x03	Reserved
Note—Values in Lane 1, Lane 2, and Lane 3 columns are in hexadecimal, most significant bit to least significant bit (i.e., <7:0>). The link fault signaling state machine allows future standardization of reserved Sequence ordered sets for functions other than link fault indications.				

The RS reports the fault status of the link. Local Fault indicates a fault detected on the receive data path between the remote RS and the local RS. The RS shall implement the link fault signaling state machine (see Figure 46– 9).

C.3.4.1 Conventions

The notation used in the state diagram follows the conventions of 21.5. The notation ++ after a counter indicates it is to be increments.

C.3.4.2 Variables and counters

The link fault signaling state machine uses the following variables and counters:

col_cnt
: A count of the number of columns received not containing a fault_sequence. This counter increments at RX_CLK rate (on both the rising and falling clock transitions) unless reset.

fault_sequence
: A new column received on RXC<3:0> and RXD<31:0> comprising a Sequence ordered_set of four bytes and consisting of a Sequence control character in lane 0 and a seq_type in lanes 1, 2, and 3 indicating either Local Fault or Remote Fault.

last_seq_type
: The seq_type of the previous Sequence ordered_set received
: Values:Local Fault; 0x00 in lane 1, 0x00 in lane 2, 0x01 in lane 3.
Remote Fault; 0x00 in lane 1, 0x00 in lane 2, 0x02 in lane 3.

link_fault
: An indicator of the fault status.
: Values:OK; No fault.
Local Fault; fault detected by the PHY.
Remote Fault; fault detection signaled by the remote RS.

reset

: Condition that is true until such time as the power supply for the device that contains the RS has

reached the operating region.

: Values: FALSE: The device is completely powered and has not been reset (default).
TRUE: The device has not been completely powered or has been reset.

seq_cnt
: A count of the number of received Sequence ordered_sets of the same type.

seq_type
: The value received in the current Sequence ordered_set
: Values: Local Fault; 0x00 in lane 1, 0x00 in lane 2, 0x01 in lane 3.
Remote Fault; 0x00 in lane 1, 0x00 in lane 2, 0x02 in lane 3.

C.3.4.3 State Diagram

Figure A.10—Link Fault Signaling State Machine

The link fault signaling state machine specifies the RS monitoring of RXC<3:0> and RXD<31:0> for Sequence ordered_sets. The variable link_fault is set to indicate the value of a received Sequence ordered_set when the following conditions have been met:

a) Four fault_sequences containing the same fault value have been received
b) Without receiving any fault_sequence within a period of 128 columns

The variable link_fault is set to OK following any interval of 128 columns not containing a Remote Fault or Local Fault Sequence ordered_set.

The RS output onto TXC<3:0> and TXD<31:0> is controlled by the variable link_fault.

a) link_fault = OK
b) The RS shall send MAC frames as requested through the PHY service interface. In the absence of MAC frames, the RS shall generate Idle control characters.
c) link_fault = Local Fault
d) The RS shall continuously generate Remote Fault Sequence ordered_sets.
e) ink_fault = Remote Fault
f) The RS shall continuously generate Idle control characters.

Annex DSONET Physical Reconciliation Sublayer

D.1 Introduction to SONET/SDH PHY Interface

D.1.1 Overview

This clause couples the IEEE 802.17 (RPR MAC) to a family of SONET/SDH Physical Layers. The relationships among SONET/SDH, the IEEE 802.17 (RPR MAC) and the ISO/IEC Open System Interconnection (OSI) reference model are shown in Figure B.1

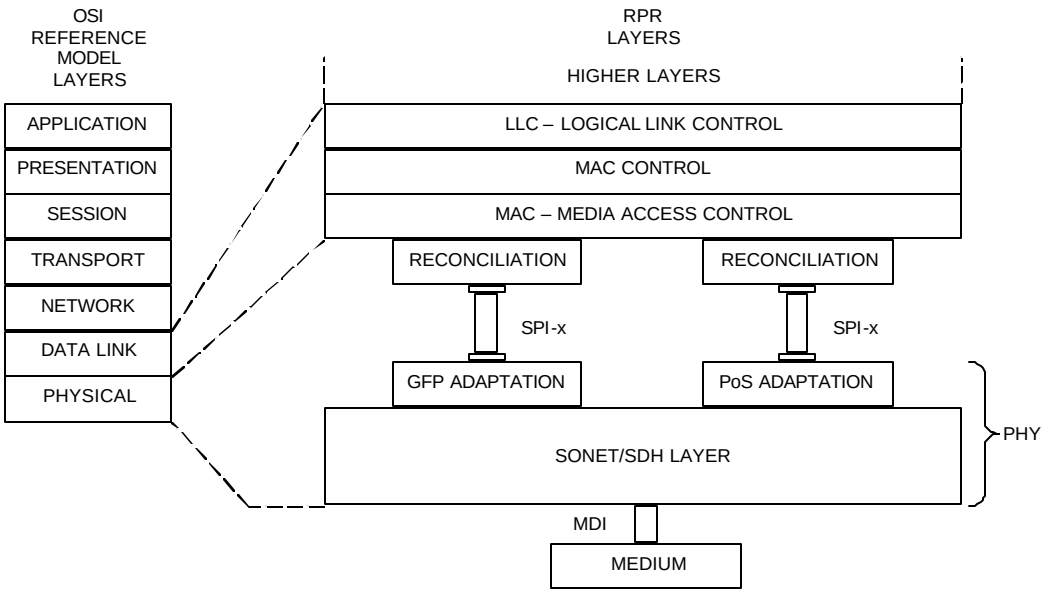


Figure B.1—Architectural positioning of SONET/SDH Physical Interfaces

The SONET/SDH PHY interfaces supports full-duplex operations at all the possible speeds.

The interface between the SONET/SDH layer and the adaptation layer (either the GFP or the PoS) is the standard interface between the SONET/SDH path sublayer and any upper layer as defined in the ITU-T Recommendations [B2] and [B4]

The GFP and PoS adaptation layers are built over the SONET/SDH Path layer and not directly over a SONET/SDH Medium. Because of the multiplexing capacity of SONET/SDH and of the virtual concatenation feature, there is not a one-to-one relationship between the PHY's bit rate seen by the RPR MAC sublayer and the speed of the physical media.

The functionality performed in the SONET/SDH layer are outside the scope of this specification. The implementations must be compliant with the already existing specifications [B2] and [B4]. The higher-layer paths (VC4), all the possible contiguous concatenated paths (VC4-4Nc) and all the possible virtual concatenated paths (VC4-Nv) are supported by the standard IEEE 802.17 RPR MAC.

All the SONET/SDH speeds, ranging from 150 Mb/s (VC-4) up to 10 Gb/s (VC4-64c), are supported by the IEEE 802.17 Standard. All the intermediate speeds that are a 4x upgrades (VC4-4Nc) are supported. In addition, using virtual concatenation all the intermediate speeds at step of 150 Mb/s (VC4-Nv) can also be supported.

[Editor's note – Check if there is the need to mention other specifications, e.g. for timing purposes.]

The IEEE 802.17 RPR MAC supports two kinds of adaptation layers for SONET/SDH interfaces, that differ each other in the way RPR frames are mapped into the SONET/SDH path payload. These two kinds of adaptation layers do not interwork each other so there are two different SONET/SDH PHYs supported by the IEEE 802.17 RPR MAC.

The first PHY type supports the GFP framing (see section D.1.3), as defined in [B5]**.**. In this case a family of Generic Reconciliation Sublayer (GRS) is defined to adapt the MAC P-SAP service interface to the SONET/SDH signal.

The second PHY type supports the PoS framing (see section D.1.4), as defined in [B6]**.**. In this case a family of PoS Reconciliation Sublayer (PRS) is defined to adapt the MAC P-SAP service interface to the SONET/SDH signal.

There are two families of Reconciliation Sublayers that differ on the encapsulation mechanism used to map RPR frames into SONET/SDH paths.

The different versions of the Reconciliation Sublayer in the same family (GRSx or PRSx) differ between each other on the version of the SPI-x interface used.

D.1.1.1 System Packet Interface (SPI)

The System Packet Interface (SPI) is a family of interfaces providing an interconnection between the Media Access Control (MAC) sublayer and the Physical Layer entities (PHY).

Two SPI versions are currently defined and supported by IEEE 802.17 Standard.

1) SPI Level 3 supports operations between 155 Mb/s and 622 Mb/s through eight bits wide transmit and receive data paths. It also supports operations between 155 Mb/s and 2,5 Gb/s through 32-bits wide transmit and receive data paths.
2) SPI Level 4 Phase 1 support operations between 622Mb/s and 10Gb/s through 64-bits wide transmit and receive data paths
3) SPI Level 4 Phase 2 supports operations between 622 Mb/s and 10 Gb/s through 16-bits wide transmit and receive data paths.

Because of the SONET/SDH standard, there is not a one-to-one correspondence between the rate at the SPI interface and the line rate on the physical medium. The former is the link rate seen by the RPR MAC sublayer.

D.1.1.2 Supported Speeds

The standard IEEE 802.17 RPR MAC supports all the allowed speeds defined in [B2]**.** and [B4]**.** for SONET/SDH paths and concatenated (either contiguous or virtual) paths starting from 150 Mb/s up to 9,6 Gb/s.

Table 17.4 summarizes all the supported RPR PHY speeds.

Note that because of the standard SONET/SDH multiplexing and virtual concatenation features, there is not a one-to-one correspondence between the RPR PHY speed and the speed of the SONET/SDH physical media over which the signal is transmitted. The possible mappings are outside the scope of this specification and already defined in the existing ITU-T Recommendations [B2] and [B4]..

Sonet Path sublayer	SDH Path sublayer	Bit-rate
STS-3c-SPE	VC4	150 Mb/s
STS-12c-SPE	VC4-4c	600 Mb/s
STS-48c-SPE	VC4-16c	2,4 Gb/s
STS-192c-SPE	VC4-64c	9,6 Gb/s
STS-3c-Nv SPE (Note 1)	VC4-Nv (Note 1)	Nx150 Nb/s (Note 1)

Table B.1—Sonet/SDH Supported Path layers and speeds

Note 1 – All the integer values of X between 2 and 192 are supported. The virtual concatenated paths can support all the speeds from 300 Mb/s up to 9,6 Gb/s with a granularity of 150 Mb/s.

D.1.1.3 Management

Managed objects, attributes and actions are defined for all the SONET/SDH components in the relevant standard recommendations [B3].

Managed objects, attributes and actions are defined for all the PoS components in the relevant standard specifications [B10],[B7] and [TBD].

[Editor's note – If we need to define an RPRCP we need a MIB module to manage it. IETF has not yet defined it. If we do not use the LCP/NCP negotiations, we need a MIB to manage this "new" standard PoS interface.]

Managed objects, attributes and actions are defined for all the GFP components in the relevant standard recommendations [TBD].

[Editor's note – Neither ITU-T neither IETF has yet addressed the issue of managing the GFP. By the completion of the IEEE 802.17 standard some specification will be available and referred here.]

D.1.2 GFP Framing

The first method to map RPR frames over SONET/SDH path payload is using the GFP mechanisms as defined in [B5]

The standard IEEE 802.17 recommends the usage of the GFP with the null extension header and no GFP FCS. All the required functionality for the GFP framing are divided into two sublayers:

1) The Generic Reconciliation Sublayer (GRS) that provides a mapping between the signals provided at the SPI-x and the MAC/PSAP service definition using GFP.

2) The GFP Adaptation Sublayer that provides a mapping between the signals provided at the SPI-x and the service interface between the GFP adaptation sublayer and the SONET/SDH path sublayer.

D.1.2.1 Generic Reconciliation Sublayer (GRS)

The GRS provides a mapping between the signals provided at the SPI-x and the MAC/PSAP service definition when GFP (refer to [B5]) is used.

There are three GRS versions defined by the IEEE 802.17 Standard.

a) The GRS1 (clause D.2) is used to map the signals provided at the 8-bit SPI-3 and the MAC/PSAP service interface.
b) The GRS2 (clause D.4) is used to map the signals provided at the 32-bit SPI-3 and the MAC/PSAP service interface.
c) The GRS3 (clause D.6) is used to map the signals provided at the SPI-4 Phase 2 and the MAC/PSAP service interface.

The GRS receives an indication of the Signal Fail and/or of the Signal Degrade condition in the SONET/SDH PHY from the Layer Management Entity (LME) and passes this information to the MAC sublayer through the P-SAP service interface.

D.1.2.2 GFP Adaptation Sublayer

All the functions of this sublayer must be compliant to the standard ITU-T Recommendations [B2], [B4] and [B5]**.** The specification of them is outside the scope of this specification. This section only defines which are the GFP functions that are placed in this sublayer.

The GFP Adaptation Sublayer in the transmit direction performs the scrambling of the GFP payload area (refer to [B5]) and inserts the GFP idle frames for rate adaptation purposes. In the receive direction it performs frame delineation, the discarding of the GFP idle frames and the de scrambling of the GFP payload area (refer to [B5]).

This sublayer receives the Trail Signal Fail (TSF) and Trail Signal Degrade (TSD) indications from the SONET/SDH Path service interface that represents respectively a failure condition and a degrade condition in the SONET/SDH layer. More details on how this information is generated and passed to the GFP adaptation sublayer are provided in [B4].

The TSD indication is passed to the Layer Management Entity (LME) that is responsible to rely it to the GRS.

The TSF indication is correlated with the dPLM and dLFD defects, as defined in [B2] and in XXX, to detect a layer 1 failure. This Signal Fail condition is then passed to the Layer Management Entity (LME) that is responsible to rely it to the GRS.

[Editor's note – There is not yet an ITU-T recommendation dealing with the defect correlation in the GFP sublayer. This recommendation should be referred here.]

D.1.3 PoS Framing

The second method to map RPR frames over SONET/SDH path payload is using the PoS mechanisms as defined in **,.**. The PoS entity is fully terminated at each station within the 802.17 network.

The standard IEEE 802.17 recommends the usage of a fixed provisioning of the POS link, with no required link negotiation. Only certain features defined in , are defined for use in 802.17 links. These features will be described in [Clause #ref].

All the required functionality for the PoS framing are divided into two sublayers:

a) The PoS Reconciliation Sublayer (PRS) that provides a mapping between the signals provided at the SPI-x and the MAC/PSAP service definition using PoS.
b) The PoS Adaptation Sublayer that provides a mapping between the signals provided at the SPI-x and the service interface between the PoS adaptation sublayer and the SONET/SDH path sublayer.

D.1.3.1 PoS Reconciliation Sublayer (PRS)

The PRS provides a mapping between the signals provided at the SPI-x and the MAC/PSAP service definition when PoS (refer to [B6]) is used.

There are four PRS versions defined by the IEEE 802.17 Standard.

1) The PRS1 (clause D.3) is used to map the signals provided at the 8-bit SPI-3 and the MAC/PSAP service interface.
2) The PRS2 (clause D.5) is used to map the signals provided at the 32-bit SPI-3 and the MAC/PSAP service interface.
3) The PRS3 (clause D.7) is used to map the signals provided at the SPI-4 Phase 1 and the MAC/PSAP service interface.
4) The PRS3 (clause D.7) is used to map the signals provided at the SPI-4 Phase 2 and the MAC/PSAP service interface.

The PRS receives an indication of the Signal Fail and/or of the Signal Degrade condition in the SONET/SDH PHY from the Layer Management Entity (LME) and passes this information to the MAC sublayer through the P-SAP service interface.

D.1.3.2 POS Adaptation Sublayer

The POS Adaptation sublayer defines the mapping of 802.17 frames into the SONET/SDH path sublayer. The method functionally the same as that defined in PPP over SONET/SDH and PPP in HDLC-like framing , except that:

— No LCP will be necessary for link establishment in 802.17, rather, a set of recommended LCP configuration options are defined which all compliant POS adaptation sublayers must support by default.
— The mapping does not carry PPP frames, it carries 802.17 frames
— PPP specific and legacy options are not required or defined for use.

The following sub clauses will define in more detail the POS Adaptation Sublayer Requirements.

D.1.3.3 RPR in HDLC-Like Framing

Figure B.2—RPR in HDLC Frame Structure

A summary of RPR HDLC-like framing is shown in FigureB.2. The Figure does not include any octets inserted for transparency. Note that the frame structure differs from that in in that certain LCP options are define such that fields from the basic frame are not required. These will be discussed in the clauses that follow.

The POS adaption sublayer defines only the use of octet-stuffed framing on 8-bit asynchronous or octet-synchronous links.

D.1.3.3.1 Flag Sequence

The Flag Sequence indicates the beginning or end of a frame. The octet stream is examined on an octet-by-octet basis for the value 01111110 (hexadecimal 0x7e).

D.1.3.3.2 Payload Transparency

An octet stuffing procedure is used. The Control Escape octet is defined as binary 01111101 (hexadecimal 0x7d), most significant bit first.

As a minimum, sending implementations MUST escape the Flag Sequence and Control Escape octets.

The transmitter examines the entire frame between the two Flag Sequences. Each Flag Sequence, Control Escape octet is replaced by a two octet sequence consisting of the Control Escape octet followed by the original octet exclusive-or'd with hexadecimal 0x20.

This is bit 5 complemented, where the bit positions are numbered 76543210 (the 6th bit as used in ISO numbered 87654321 -- BEWARE when comparing documents).

Receiving implementations MUST correctly process all Control Escape sequences.

On reception, each Control Escape octet is removed, and the following octet is exclusive-or'd with hexadecimal 0x20, unless it is the Flag Sequence (which aborts a frame).

D.1.3.3.3 Invalid Frames

Frames which are too short (less than 2 octets), or which end with a Control Escape octet followed immediately by a closing Flag Sequence, or in which octet-framing is violated, are silently discarded.

D.1.3.4 Time Fill

There is no provision for inter-octet time fill.

The flag sequence MUST be transmitted during interframe time fill.

D.1.3.5 Transmission Considerations

All octets are transmitted Least Significant Bit First.

D.1.3.6 Defined LCP Options

The POS Adaptation sublayer will not required an LCP negotiation to set link parameters. Instead a set of LCP options is statically defined for POS links in an 802.17 network station. These options are as follows:

1) Address and Control Field compression is always used
2) Protocol Field is not used
3) FCS is neither computed nor appended to the Frame
4) The Asynchronous Control Character Map (ACCM) is not used
5) $X^{43}+1$ Payload Scrambling is used

D.1.3.7 Physical Layer Requirements

PPP treats SONET/SDH transport as octet oriented synchronous links.

SONET/SDH links are full-duplex by definition.

D.1.3.7.1 Interface Format

PPP in HDLC-like framing presents an octet interface to the physical layer. There is no provision for sub-octets to be supplied or accepted [B8][B9].

The octet stream is mapped into the SONET STS-SPE/SDH Higher Order VC, with the octet boundaries aligned with the SONET STS-SPE/SDH Higher Order VC octet boundaries.

Scrambling is performed during insertion into the SONET STS- SPE/SDH Higher Order VC to provide adequate transparency and protect against potential security threats (see Section 6). For backwards compatibility with RFC 1619 (STS-3c-SPE/VC-4 only), the scrambler MAY have an on/off capability where the scrambler is bypassed entirely when it is in the off mode. If this capability is provided, the default MUST be set to scrambling enabled.

For RPR over SONET/SDH, the entire SONET/SDH payload (SONET STS- SPE/SDH Higher Order VC minus the path overhead and any fixed stuff) is scrambled using a self-synchronous scrambler of polynomial X**43 + 1. [Clause #REF] for the description of the scrambler.

The proper order of operation is:

When transmitting:

CPDU -> RPR -> Byte stuffing -> Scrambling -> SONET/SDH framing

When receiving:

SONET/SDH framing -> Descrambling -> Byte destuffing -> FCS detection -> RPR -> CPDU

Where CPDU is Client Protocol Data Unit.

The Path Signal Label (C2) indicates the contents of the SONET STS-SPE/SDH Higher Order VC. The value of XX (XX hex) is used to indicate PPP with X^43 + 1 scrambling.

For compatibility with RFC 1619 (STS-3c-SPE/VC-4 only), if scrambling has been configured to be off, then the value XX (XX hex) is used for the Path Signal Label to indicate PPP without scrambling.

The Multiframe Indicator (H4) is unused, and MUST be zero.

[ED: need to get appropriate code-points defined in for RPR over POS - TBD].

D.1.3.7.2 Control Signals

The POS Adaptation layer does not require the use of control signals as discussed in .

D.1.3.8 Link Performance Monitoring

This sublayer receives the Trail Signal Fail (TSF) and Trail Signal Degrade (TSD) indications from the SONET/SDH Path service interface that represents respectively a failure condition and a degrade condition in the SONET/SDH layer. More details on how this information is generated and passed to the POS adaptation sublayer are provided in [B4].

The TSD indication is passed to the Layer Management Entity (LME) that is responsible to rely it to the PRS.

The TSF indication is correlated with the dPLM defect, as defined in [B2], to detect a layer 1 failure. This Signal Fail condition is then passed to the Layer Management Entity (LME) that is responsible to rely it to the PRS.

D.2 Generic Reconciliation Sublayer version 1 (GRS1) and the 8-bit System Packet Interface Level 3 (SPI-3)

D.2.1 Overview

This clause defines the logical and electrical characteristics for the Generic Reconciliation Sublayer version 1 (GRS2) and System Packet Interface Level 3 (SPI-3) using the 8-bit data path, between the RPR media access control and various SONET/SDH PHYs.

FigureB.3 shows the relationship of the Reconciliation sublayer and SPI-3 to the ISO/IEC OSI reference model

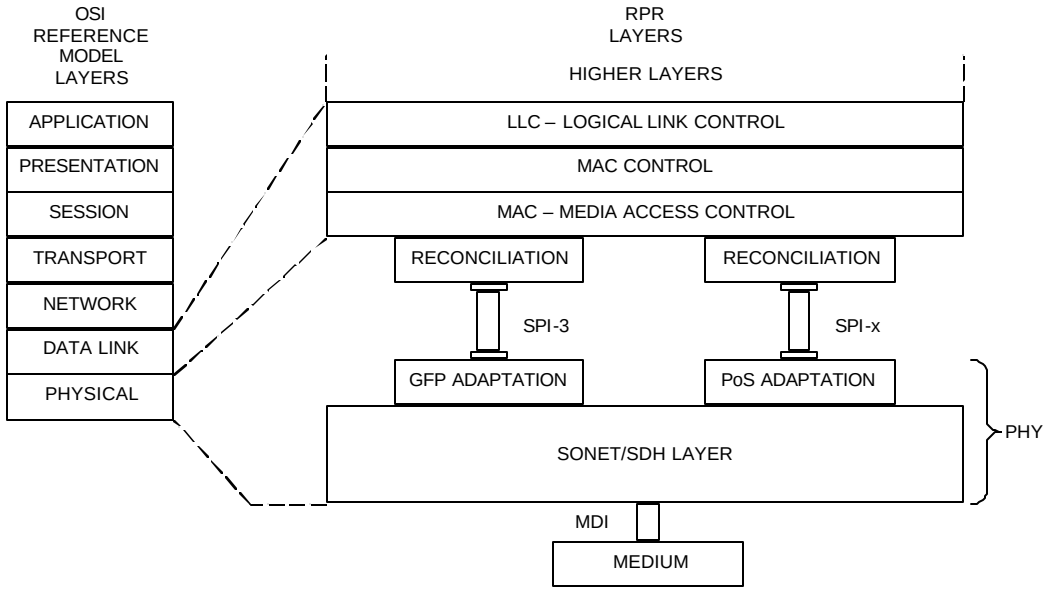


Figure B.3—GRS1 and SPI-3 location in the OSI protocol stack

The purpose of this interface is to provide a simple, inexpensive and easy-to-implement interconnection between Media Access Control (MAC) sublayer and PHYs, and between PHYs and Layer Management Entity (LME).

The interface has the following characteristics:

a) It is capable of supporting operations between 155 Mb/s and 622 Mb/s (with a 155 Mb/s granularity).
b) Data and delimiter are synchronous to clock references.
c) It provides independent 8-bit wide transmit and receive data paths.
d) It provides a simple management interface.
e) It provides for full duplex operations.

D.2.1.1 Summary of major concepts

a) Each direction of data transfer is serviced by Data (a 8-bit bundle), Delimiter, Error and Clock signals.
b) The SPI-3 supports logical channels with individual word-level or packet level flow control.
c) The flow control is done out-of-band.
d) The Reconciliation Sublayer maps the signal set provided at the SPI-3 to the P-SAP service primitives provided to the MAC.

D.2.1.2 Application

This clause applies to the interface between the MAC and PHYs and between PHYs and Layer Management Entity (LME). The implementation of the interface is primarily intended as a chip-to-chip (integrated circuits to integrated circuit) interface implemented with traces on a printed circuit board. SA motherboard-to-daughter board interface between two or more printed circuit boards is not precluded.

This interface is used to provide media independence so that an identical media access controller may be used with any of the supported SONET/SDH PHY types.

D.2.1.3 Rate of operation

The SPI-3 with 8-bit data path supports operation at all the speeds between 155 Mb/s and 622 Mb/s at the granularity of 155 Mb/s. It is defined in **Error! Reference source not found.**

SONET/SDH PHY that provide an SPI-3 with 8-bit data path shall support operations, at the SONET/SDH Path level, at the selected rate on the SPI-3. PHYs must report the rates at which they are operating via the management interface.

D.2.1.4 Allocation of functions

The allocation of functions in the SPI interfaces is such that it readily lends itself to implementation in both PHY and MAC sublayer entities.

D.2.1.5 Mapping of SPI-3 signals to P-SAP service primitive and Station Management

The Generic Reconciliation Sublayer (GRS) shall map the signals provided at the SPI-3 to the P-SAP service primitives as shown in Figure 17.6. The following P-SAP service primitives are defined:

— PHY_DATA.request
— PHY_DATA.indicate
— PHY_DATA_VALID.indicate
— PHY_LINK_OK.indicate
— PHY_READY.indicate

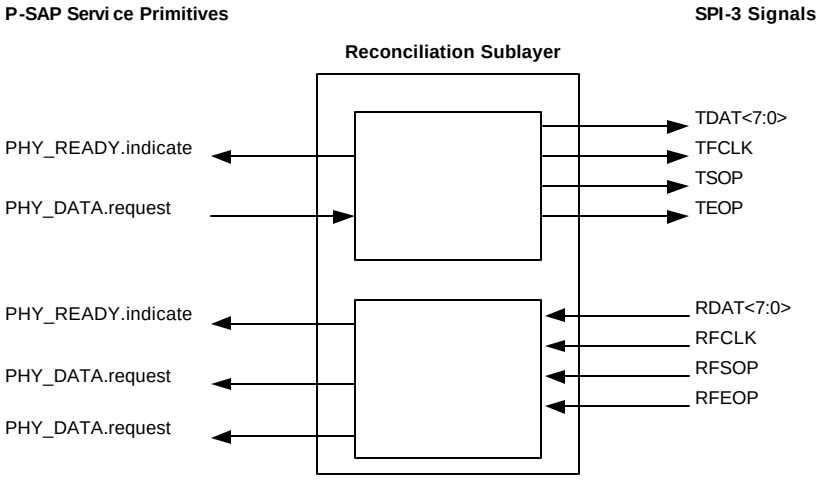


Figure B.4—Generic Reconciliation Sublayer Version 1 inputs and outputs

[Editor's note – The same principles described in section D.4.2.1 applies also here.]

D.2.1.6 SPI-3 functional specifications

The SPI-3 is designed to make the differences among the various SONET/SDH Path sublayers and media combinations transparent to the RPR MAC sublayer.

D.2.1.7 Transmit

The SPI-3 packet interface supports transmit and receive data transfers at clock rates independent of the line bit rate. As a result, PHY layer devices must support packet rate decoupling using FIFOs.

To ease the interface between the Link Layer and PHY layer devices and to support multiple PHY layer interfaces, FIFOs are used. Control signals are provided to both the Link Layer and PHY layer devices to allow either one to exercise flow control. Since the bus interface is point-to-point, the receive interface of the PHY device pushes data to the Link Layer device. For the transmit interface, the packet available status granularity is byte-based.

In the receive direction, when the PHY layer device has stored an end-of packet (a complete small packet or the end of a larger packet) or some predefined number of bytes in its receive FIFO, it sends the in-band address followed by FIFO data to the Link Layer device. The data on the interface bus is marked with the valid signal (RVAL) asserted. A multi-port PHY device with multiple FIFOs would service each port in a round-robin fashion when sufficient data is available in its FIFO. The Link Layer device can pause the data flow by deasserting the enable signal (RENB).

In the transmit direction, when the PHY layer device has space for some predefined number of bytes in its transmit FIFO, it informs the Link Layer device by asserting a transmit packet available (TPA). The Link Layer device can then write the in-band address followed by packet data to the PHY layer device using an enable signal (TENB). The Link Layer device shall monitor TPA for a high to low transition, which would indicate that the transmit FIFO is near full (the number of bytes left in the FIFO can be user selectable, but must be predefined), and suspend data transfer to avoid an overflow. The Link Layer device can pause the data flow by deasserting the enable signal (TENB).

SPI-3 defines both byte-level and packet-level transfer control in the transmit direction. In byte-level transfer, FIFO status information is presented on a cycle-by-cycle basis. With packet-level transfer, the FIFO status information applies to segments of data. When using byte level transfer, direct status indication must be used. In this case, the PHY layer device provides the transmit packet available status of the selected port (STPA) in the PHY device. As well, the PHY layer device may provide direct access to the transmit packet available status of all ports (DTPA[]) in the PHY device if the number of ports is small. With packet level transfer, the Link Layer device is able to do status polling on the transmit direction. The Link Layer device can use the transmit port address TADR[] to poll individual ports of the PHY device, which all respond on a common polled (PTPA) signal.

Since the variable size nature of packets does not allow any guarantee as to the number of bytes available, in both transmit and receive directions, a selected PHY transmit packet available is provided on signal STPA and a receive data valid on signal RVAL. STPA and RVAL always reflect the status of the selected PHY to or from which data is being transferred. RVAL indicates if valid data is available on the receive data bus and is defined such that data transfers can be aligned with packet boundaries.

Physical layer port selection is performed using in-band addressing. In the transmit direction, the Layer device selects a PHY port by sending the address on the TDAT[] bus marked with the TSX signal active and TENB signal inactive. All subsequent TDAT[] bus operations marked with the TSX signal inactive and the TENB active will be packet data for the specified port.

In the receive direction, the PHY device will specify the selected port by sending the address on the RDAT[] bus marked with the RSX signal active and RVAL signal inactive. All subsequent RDAT[] bus operations marked with RSX inactive and RVAL active will be packet data from the specified port.

Both byte-level and packet-level modes are specified in this standard in order to support the current low density multi-port physical layer devices and future higher density multi-port devices. When the number of ports in the physical layer device is limited, byte-level transfer using DTPA[] signals provides a simpler implementation and reduces the need for addressing pins. In this case, direct access will start to become unreasonable as the number of ports increase. Packet-level transfer provides a lower pin count solution using the TADR[] bus when the number of ports is large. In-band addressing ensures the protocol remains consistent between the two approaches. However, the final choice left to the system designers and physical layer device manufacturers to select which approach best suits their desired applications.

Packets shall be written into the transmit FIFO and read from the receive FIFO using a defined data structure. Octets are written in the same order they are to be transmitted or they were received on the SONET line. Within an octet, the MSB (bit 7) is the first bit to be transmitted. The SPI-3 specification does not preclude the transfer of 1-byte packets. In this case, both start of packet and end of packet signals shall be asserted simultaneously.

For packets longer than the PHY device FIFO, the packet must be transferred over the bus interface in sections. The number of bytes of packet data in each section may be fixed or variable depending on the application. In general, the Receive Interface will round-robin between receive FIFOs with fill levels exceeding a programmable high water mark or with at least one end of packet stored in the FIFO. The Receive Interface would end the transfer of data when an end of a packet is transferred or when a programmable number of bytes have been transferred. The Link Layer device may send fixed size sections of packets on the Transmit Interface or use the TPA signal to determine when the FIFO reaches a full level.

The in-band address is specified in a single clock cycle operation marked with the RSX/TSX signals. The port address is specified by the TDAT[7:0]/RDAT[7:0] signals. The address is the numeric value of the TDAT[7:0]/RDAT[7:0] signals where bit 0 is the least significant bit and bit 7 is the most significant bit. Thus, up to 256 ports may be supported by a single interface. With a 32-bit interface, the upper 24 bits shall be ignored.

D.2.1.8 Receive

The receive FIFO shall have a programmable threshold defined in terms of the number of bytes of packet data stored in the FIFO. A multi-port PHY device must service each receive FIFO with sufficient packet data to exceed the threshold or with an end of packet. The PHY should service the required FIFOs in a round-robin fashion. The type of round-robin algorithm will depend on the various data rates supported by the PHY device and is outside this specification.

The amount of packet data transferred, when servicing the receive FIFO, is bounded by the FIFO's programmable threshold. Thus, a transfer is limited to a maximum of 256 bytes of data (64 cycles for a 32-bit interface or 256 cycles for an 8-bit interface) or until an end of packet is transferred to the Layer device. At the end of a transfer, the PHY device will round-robin to the next receive FIFO.

The PHY device should support a programmable minimum pause of 0 or 2 clock cycles between transfers. A pause of 0 clock cycles maximizes the throughput of the interface. A pause of 2 clock cycles allows the Layer device to pause between transfers.

D.2.1.9 SPI-3 Signal timing characteristic

TFCLK Frequency is specified to be 104 MHz. TFCLK duty cycle is min 40% and max 60%.

Setup time relative to TFCLK for TENB, TDAT, TPRTY, TSOP, TEOP, TMOD, TERR, TSX and TADR is 2 ns minimum. Hold time relative to TFCLK for the same set of signals is 0.5 ns minimum.

Validity of DTPA, STPA, PTPA relative to the rising edge of TFCLK is 1.5ns minimum and 6.0 ns maximum.

RFCLK Frequency is specified to be 104 MHz. RFCLK duty cycle is min 40% and max 60%.

Setup time relative to RFCLK for RENB is 2 ns minimum. Hold time relative to RFCLK for RENB is 0.5 ns minimum.

Validity of RDAT, RPRTY, RSOP, REOP, RMOD, RERR, RVAL, and RSX relative to the rising edge of RFCLK is 1.5 ns minimum and 6.0 ns maximum.

D.2.1.10 SPI-3 electrical characteristics

TBD

D.3 POS Reconciliation Sublayer version 1 (PRS1) and the 8-bit System Packet Interface Level 3 (SPI-3)

[Editor's note – The same principles described in section D.5 applies also here.]

TBD

D.4 Generic Reconciliation Sublayer version 2 (GRS2) and the 32-bit System Packet Interface Level 3 (SPI-3)

D.4.1 Overview

This clause defines the logical and electrical characteristics for the Generic Reconciliation Sublayer version 2 (GRS2) and System Packet Interface Level 3 (SPI-3) using the 32-bit data path, between the RPR media access control and various SONET/SDH PHYs.

Figure 17.7 shows the relationship of the Reconciliation sublayer and SPI-3 to the ISO/IEC OSI reference model

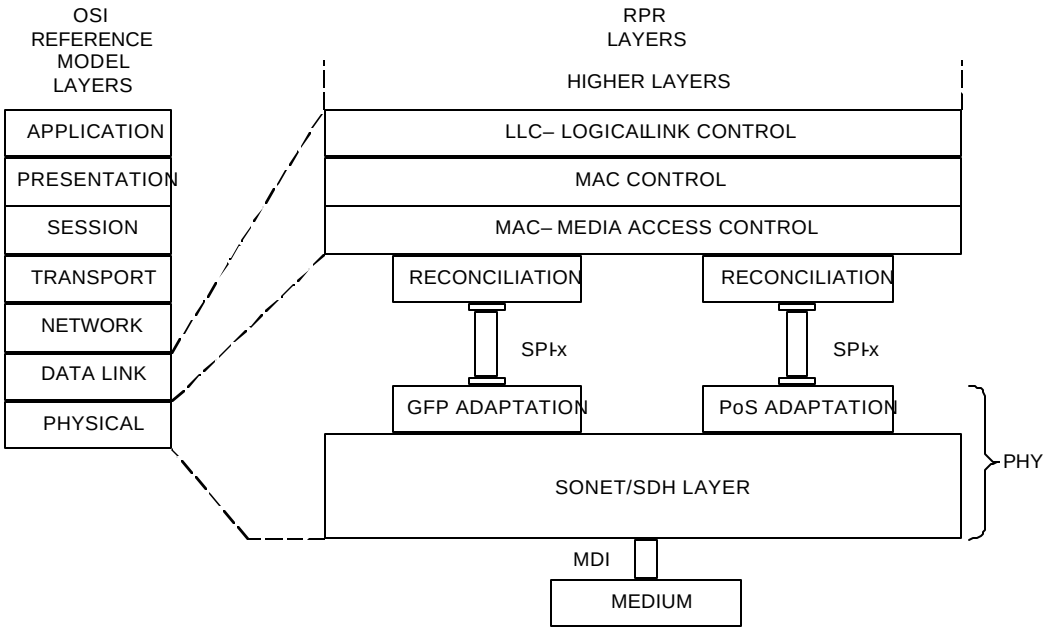


Figure B.5—GRS-3 and SPI-3 location in the OSI protocol stack

The purpose of this interface is to provide a simple, inexpensive and easy-to-implement interconnection between Media Access Control (MAC) sublayer and PHYs, and between PHYs and Layer Management Entity (LME).

The interface has the following characteristics:

a) It is capable of supporting operations between 155 Mb/s and 622 Mb/s (with a 155 Mb/s granularity).
b) Data and delimiter are synchronous to clock references.
c) It provides independent 32-bit wide transmit and receive data paths.
d) It provides a simple management interface.
e) It provides for full duplex operations.

D.4.1.1 Summary of major concepts

a) Each direction of data transfer is serviced by Data (a 32-bit bundle), Delimiter, Error and Clock signals.
b) The SPI-3 supports logical channels with individual word-level or packet level flow control.
a) The flow control is done out-of-band.
a) The Reconciliation Sublayer maps the signal set provided at the SPI-3 to the P-SAP service primitives provided to the MAC.

D.4.1.2 Application

This clause applies to the interface between the MAC and PHYs and between PHYs and Layer Management Entity (LME). The implementation of the interface is primarily intended as a chip-to-chip (integrated circuits to integrated circuit) interface implemented with traces on a printed circuit board. SA motherboard-to-daughterboard interface between two or more printed circuit boards is not precluded.

This interface is used to provide media independence so that an identical media access controller may be used with any of the supported SONET/SDH PHY types.

D.4.1.3 Rate of operation

The SPI-3 with 32-bit data path supports operation at all the speeds between 155 Mb/s and 2.5 Gb/s at the granularity of 155 Mb/s. It is defined in D.8

SONET/SDH PHY that provide an SPI-3 with 32-bit data path shall support operations, at the SONET/SDH Path level, at the selected rate on the SPI-3. PHYs must report the rates at which they are operating via the management interface.

D.4.1.4 Allocation of functions

The allocation of functions in the SPI interfaces is such that it readily lends itself to implementation in both PHY and MAC sublayer entities.

D.4.1.5 Mapping of SPI-3 signals to P-SAP service primitive and Station Management

The Generic Reconciliation Sublayer (GRS) shall map the signals provided at the SPI-3 to the P-SAP service primitives as shown in Figure 17.8. The following P-SAP service primitives are defined:

— PHY_DATA.request
— PHY_DATA.indicate
— PHY_DATA_VALID.indicate
— PHY_LINK_OK.indicate
— PHY_READY.indicate

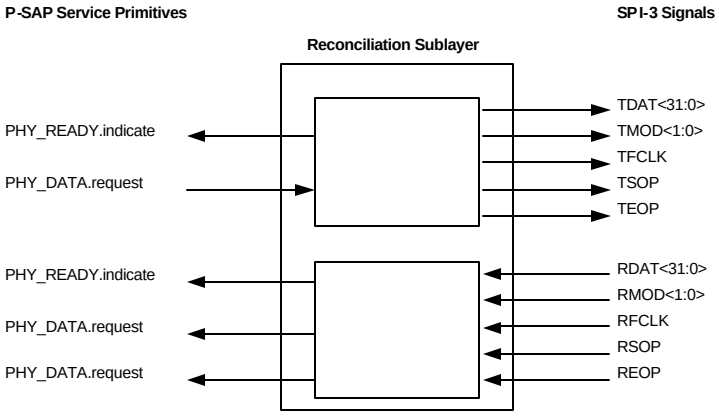


Figure B.6—Generic Reconciliation Sublayer Version 2 inputs and outputs

D.4.1.5.1 Mapping of PHY_DATA.request

D.4.1.5.1.1 Function

Map the primitive PHY_DATA.request to the SPI-3 signals TFCLK, TERR, TENB, TDAT<31:0>,TPRTY, TMOD<1:0>, TSX, TSOP, TEOP.

When transmitting RPR MAC frames the TENB signal is low, so the TSX signal is ignored (see D.8). The usage of the TSX and TENB signals for in-band addressing with multi-port PHY is an implementation choice and out of scope of this specification. If used, this feature should be compliant with the requirements in D.8.

D.4.1.5.1.2 Semantics of the service primitive

`PHY_DATA.request (OUTPUT_UNIT)`

The OUTPUT_UNIT parameter either takes the value of an octet of data or the DATA_COMPLETE, and it represents the transfer of an octet of data from the MAC to the RS. The value DATA_COMPLETE indicates that the MAC has no more data to transfer.

D.4.1.5.1.3 When generated

This primitive is generated by the MAC sublayer to request the transmission of a data octet on the physical medium, or to indicate that no more data is available for transmission.

D.4.1.5.1.4 Effect of receipt

The OUTPUT_UNIT values are conveyed to the PHY by the signals TDAT<31:0> and TMOD<1:0> on each TFCLK rising edge.

Each PHY_DATA.request transaction shall be mapped to a TDAT signal in sequence (TDAT<0:7>, …, TDAT<24:31>, TDAT<0:7>), using the big endian coding, as described in D.2.2.6. After the first five PHY_DATA.request and after each four PHY_DATA.request transactions from the MAC sublayer, the RS requests transmission of 32 data bits by the PHY, together with the proper parity information in the TPRTY signal (see D.8), containing the values of the previous four PHY_DATA.request transactions except the last one. The TMOD<1:0> is always fixed to the “00” value except when the end-of-packet is transmitted.

[Editor’s note – The mapping of a MAC data byte on the SPI interface is left to the open discussion. Do we have to invert the bit sequence, due to the big endian order required at the SPI?]

The first eight octets of the frame shall be converted as follows, in order to build a GFP frame:

a) The first two octets are not changed. They correspond to the PLI field of the GFP core header.
b) The following two octets are filled with the CRC-16 calculated, as described in [B5], on the first two octets. They correspond to the cHEC field of the GFP core header.
c) The following two octets are filled with the value 0x0009, representing the PTI=î000î (user data frame), the PFI=0 (no payload FCS), the EXI=î0000î (null extension header) and the UPI=î0000 1001î (RPR frame). They correspond to the Type field of the GFP payload header.
d) The last two octets are filled with the CRC-16 calculated, as described in [B5]on the previous two octets. They correspond to the tHEC field of the GFP payload header.

[Editor’s note – The UPI value “0000 1001” is assigned to RPR by ITU-T as a provisionally agreed item in the living list.]

When transmitting the first four octets, representing the PLI and cHEC fields of the GFP core header, on TDAT<31:0>, the TSOP signal is also high.

The RPR MAC does not generate inter-frame spacing and the GRS does not generate it because there is no need of IPG between RPR MAC frames when using SONET/SDH PHY layers.

The DATA_COMPLETE value shall trigger the indication of the end-of-packet on the SPI-3 as follows:

a) The DATA_COMPLETE value is not mapped to any character on the TDAT signal.
b) The RS requests transmission of 32 data bits by the PHY, together with the proper parity information in the TPRTY signal, containing the values of the previous PHY_DATA.request transactions not yet transmitted. When transmitting this TDAT<31:0> signal, the TEOP is also high and the TMOD<1:0> represents the number of valid data bytes on the TDAT<31:0> as defined in D.8.

D.4.1.5.2 Mapping of PHY_DATA.indicate

D.4.1.5.2.1 Function

Map the primitive PHY_DATA.indicate to the SPI-3 signals RFCLK, RVAL, RENB, RDAT<31:0>, RPRTY, RMOD<1:0>, RSOP, REOP, RERR, RSX.

When receiving RPR MAC frames the RENB signal is low and the RVAL is high, so the RSX signal is ignored (see D.8). The usage of the RSX and RENB signals for in-band addressing with multi-port PHY is an implementation choice and out of scope of this specification. If used, this feature should be compliant with the requirements in D.8.

D.4.1.5.2.2 Semantics of the service primitive

```
PHY_DATA.indicate (INPUT_UNIT)
```

The INPUT_UNIT parameter takes the value of an octet of data and defines the transfer of an octet of data from the RS to the MAC.

D.4.1.5.2.3 When generated

The INPUT_UNIT values are derived from the signals RMOD<1:0> and RDAT<31:0> received from the PHY on each rising edge of the RFCLK. Each primitive generated to the MAC sublayer entity corresponds to a PHY_DATA.request issued by the MAC at the upstream end of the ringlet connecting two RPR adjacent nodes.

For each RDAT<31:0> during frame reception, the RS shall generate four PHY_DATA.indicate transactions until the end of frame (when the REOP is high), where one, tow, three or four PHY_DATA.indicate transactions will be generated from the RDAT<31:0> according to the value coded in the RMOD<1:0> as defined in D.8.

During frame reception each RDAT signal shall be mapped in sequence into a PHY_DATA.indicate transaction (RDAT<0:7>, …, RDAT<24:31>, RDAT<0:7>) as described in D.2.2.6.

Before starting the frame reception, the first eight octets should be checked to be consistent. The following checks are performed:

a) The GFP tHEC field is checked in accordance with [B5].
b) The GFP Type field, eventually corrected by the tHEC, is checked to be equal to 0x0009, representing the PTI=î000î (user data frame), the PFI=0 (no payload FCS), the EXI=î0000î (null extension header) and the UPI=î0000 1001î (RPR frame). If the Type field is different from the expected one, the GFP frame must be discarded.

If the previous checks have detected that the GFP frame has to be discarded, no PHY_DATA.indicate primitive is generated during the reception of this GFP frame.

If the previous checks succeed, the RS shall convert the first eight octets of data (delineated by the RSOP signal) as follows, prior to generation of the associated PHY_DATA.indicate transactions.

a) The first two octets are not changed, in order to pass to the RPR MAC layer the length information.
b) The following six octets are converted to null data.

[Editor's note – Which is the null data to be placed in these bytes? We need an input from the frame format group.]

D.4.1.5.2.4 Effect of receipt

The effect of receipt of this primitive by the MAC sublayer is unspecified.

D.4.1.5.3 Mapping of PHY_DATA_VALID.indicate

D.4.1.5.3.1 Function

Map the primitive PHY_DATA_VALID.indicate to the SPI-3 signals RFCLK, RVAL, RENB, RDAT<31:0>, RPRTY, RMOD<1:0>, RSOP, REOP, RERR, RSX.

D.4.1.5.3.2 Semantics of the service primitive

```
PHY_DATA_VALID.indicate (DATA_VALID_STATUS)
```

The DATA_VALID_STATUS parameter can take one of two values: DATA_VALID or DATA_NOT_VALID. The DATA_VALID value indicates that the INPUT_UNIT parameter of the PHY_DATA.indicate primitive contains a valid data of an incoming frame. The DATA_NON_VALID value indicates that the INPUT_UNIT parameter of the PHY_DATA.indicate primitive des not contain valid data of an incoming frame.

D.4.1.5.3.3 When generated

The RS shall generate the PHYD_DATA_VALID.indicate service primitive whenever the DATA_VALID_STATUS parameter changes from DATA_VALID to DATA_NOT_VALID or vice versa.

The DATA_VALID_STATUS shall assume the value DATA_VALID when a PHY_DATA.indicate transaction is generated in response to reception of a RSOP signal.

The value DATA_NOT_VALID is assumed after the REOP signal to indicate the end of frame to the RPR MAC.

D.4.1.5.3.4 Effect of receipt

The effect of receipt of this primitive by the MAC sublayer is unspecified.

D.4.1.5.4 Mapping of PHY_LINK_OK.indicate

D.4.1.5.4.1 Function

Map the primitive PHY_LINK_OK.indicate to the STA signals MDSF and MDSD.

D.4.1.5.4.2 Semantics of the service primitive

```
PHY_LINK_OK.indicate (LINK_STATUS)
```

The LINK_STATUS parameter can take the value of OK, FAIL or DEGRADE, and signifies the status of the PHY as indicated by the MDSF (Management Data Signal Fail) and MDSD (Management Data Signal Degrade).

The value OK indicates that there are neither failure nor degrade conditions in the physical layer. The value FAIL indicates that a failure has been detected in the physical layer. The value DEGRADE indicates that a signal degrade condition has been detected in the physical layer.

D.4.1.5.4.3 When generated

The RS shall generate the PHY_LINK_OK.indicate primitives whenever the LINK_STATUS parameter changes from one possible value (OK, FAIL, DEGRADE) to a different possible value.

The LINK_STATUS values are derived from the MDSF and MDSD signals received from the Layer Management asynchronously from the data transfer.

The LINK_STATUS parameter shall assume the value OK when neither MDSF nor MDSD signal has been received.

It shall assume the value FAIL when the MDSF signal is asserted.

The value DEGRADE is assumed when the MDSD signal is asserted and the MDSF has not been received.

D.4.1.5.4.4 Effect of receipt

The effect of receipt of this primitive by the MAC sublayer is unspecified.

D.4.1.5.5 Mapping of PHY_READY.indicate

D.4.1.5.5.1 Function

Map the primitive PHY_READY.indicate to the GRS state machine.

[Editors' note – We have to discuss more in detail this primitive and how it should work.]

D.4.1.5.5.2 Semantics of the service primitive

```
PHY_READY.indicate (READY_STATUS)
```

The READY_STATUS parameter takes the value of READY or NOT_READY, and signifies the status of a transmitted frame. The READY value indicates that a new frame can be received from the RPR MAC sublayer. The NOT_READY value indicates that TBD.

[Editor's note – It is not completely clear to me how the PHY_READY.indicate primitive is used.]

D.4.1.5.5.3 When generated

The RS shall generate the PHY_READY.indicate primitive whenever the READY_STATUS parameter changes from one the READY value to the NOT_READY value and vice versa.

If a frame is being received from the MAC and has not been transmitted, READY_STATUS indicates NOT_READY. If a frame has been received and fully transmitted, READY_STATUS indicates READY, signifying that a new frame can be received.

[Editor's note – It is not completely clear to me how the PHY_READY.indicate primitive is used.]

D.4.1.5.5.4 Effect of receipt

The effect of receipt of this primitive by the MAC sublayer is unspecified.

D.4.2 SPI-3 functional specifications

Refer to section D.2.1.6.

D.5 POS Reconciliation Sublayer version 2 (PRS2) and the 32-bit System Packet Interface Level 3 (SPI-3)

[Editor's note – The same principles of section D.4 applies also here, with some updates due to the usage of POS instead of GFP.]

TBD

D.6 Generic Reconciliation Sublayer version 3 (GRS3) and System Packet Interface Level 4 Phase 2 (SPI-4)

[Editor's note – The same principles of section D.4 applies also here.]

TBD

D.7 POS Reconciliation Sublayer version 3 (PRS3) and System Packet Interface Level 4 Phase 2 (SPI-4)

[Editor's note – The same principles of section D.5 applies also here.]

TBD

D.8 SPI-3 Signaling Function Specifications

All signals are expected to be updated and sampled using the rising edge of the transmit FIFO clock TFCLK.

D.8.1 TFCLK (transmit clock)

TFCLK is a continuous clock used to synchronize data transfer transactions between the LINK Layer device and the PHY layer device. TFCLK may cycle at a rate up to 104 MHz. TFCLK is sourced by the MAC to the PHY.

D.8.2 ERR (transmit error Indicator)

TERR is used to indicate that there is an error in the current packet. TERR should only be asserted when TEOP is asserted; it is considered valid only when TENB is asserted.

D.8.3 TENB (transmit write enable)

TENB used to control the flow of data to the transmit FIFOs. When TENB is high the TDAT, TMOD, TSOP, TEOP and TERR signals are invalid and should be ignored by the PHY. The TSX signal is valid and is processed by the PHY when TENB is high. When TENB is low the TDAT, TMOD, TSOP, TEOP and TERR signals are valid and are processed by the PHY. The TSX signal is ignored by the PHY when TENB is low.

D.8.4 TDAT[31:0] (transmit packet data bus)

TDAT is an 8-bit or 32-bit bus which carries the packet octets that are written to the selected transmit FIFO and the in-band port address to select the desired transmit FIFO. The TDAT bus is considered valid only when TENB is asserted. Data is transmitted in big endian order on TDAT.

D.8.5 TMOD[1:0] (transmit word modulo)

TMOD is required only when TDAT is 32-bits. TMOD indicates the number of valid bytes of data in TDAT. The TMOD bus should always be zero except during the last word transfer of a packet on TDAT. When TEOP is asserted the number of valid packet data bytes on TDAT is decoded from TMOD as:

TMOD[1:0] = ì00îTDAT[31:0] is valid
TMOD[1:0] = ì01îTDAT[31:8] is valid
TMOD[1:0] = ì10îTDAT[31:16] is valid
TMOD[1:0] = ì11îTDAT[31:24] is valid

TMOD is considered valid only when TENB is asserted.

D.8.6 TPRTY (transmit bus parity)

TPRTY indicates the parity calculated over the TDAT bus. TPRTY is considered valid only when TENB or TSX is asserted. The parity calculation is such that odd parity is indicated on this signal.

D.8.7 TSX (transmit start of transfer)

TSX indicates when the in-band port address is present on the TDAT bus. When TSX is high and TENB is high, the value of TDAT is the address of the transmit FIFO to be selected. Subsequent data transfers on the TDAT bus will fill the FIFO specified by this in-band address. For single channel PHY devices the TSX signal is optional. TSX is considered valid only when TENB is not asserted.

D.8.8 TSOP (transmit start of packet)

TSOP is used to delineate the packet boundaries on the TDAT bus. When TSOP is high the start of the packet is present on the TDAT bus. TSOP is required to be present at the beginning of every packet and is considered valid only when TENB is asserted.

D.8.9 TEOP (transmit end of packet)

TEOP is used to delineate the packet boundaries on the TDAT bus. When TEOP is high, the end of the packet is present on the TDAT bus. TEOP is required to be present at the end of every packets and is considered valid only when TENB is asserted.

D.8.10 TADR (transmit PHY Address)

TADR bus is used with the PTPA signal to poll the transmit FIFO's packet available status. When TADR is sampled on the rising edge of TFCLK by the PHY the polled packet available indication PTPA is updated with the status of the channel specified by the TADR address on the following rising edge of TFCLK.

D.8.11 DTPA (direct transmit packet available)

DTPA bus provides direct status indication for the corresponding ports in the PHY device. DTPA transitions high when a predefined minimum number of bytes is available in its transmit FIFO. Once high, the DTPA signal indicates that its corresponding transmit FIFO is not full. When DTPA transitions low it indicates that its transmit FIFO is full or near full. DTPA is updated on the rising edge of TFCLK.

D.8.12 RFCLK (receive FIFO write clock)

RFCLK is a continuous clock used to synchronize data transfer transactions between the MAC and the PHY. RFCLK may cycle at a rate up to 104 MHz.

D.8.13 RVAL (receive data valid)

RVAL indicates the validity of the receive data. RVAL is low between transfers and when RSX is asserted. It is also low when the PHY pauses a transfer due to an empty receive FIFO. When a transfer is paused by holding RENB high RVAL will hold its value unchanged although no new data will be present on RDAT until the transfer resumes. When RVAL is high the RDAT, RMOD, RSOP, REOP and RERR signals are valid. When RVAL is low, the RDAT, RMOD, RSOP, REOP and RERR signals are invalid and must be disregarded. The RSX signal is valid when RVAL is low.

D.8.14 RENB (receive read enable)

The RENB signal is used to control the flow of data from the receive FIFOs. During data transfer, RVAL must be monitored as it will indicate if the RDAT, RPRTY, RMOD, RSOP, REOP, RERR and RSX are valid. The system may deassert RENB at anytime if it is unable to accept data from the PHY device. When RENB is sampled low by the PHY device a read is performed from the receive FIFO and the RDAT, RPRTY, RMOD, RSOP, REOP, RERR, RSX and RVAL signals are updated on the following rising edge of RFCLK. When RENB is sampled high by the PHY device a read is not performed and the RDAT, RPRTY, RMOD, RSOP, REOP, RERR, RSX and RVAL remain unchanged on the following rising edge of RFCLK.

D.8.15 RDAT[31:0] (receive packet data bus)

RDAT is an 8-bit or 32-bit bus which carries the packet octets that are read from the receive FIFO and the in-band port address of the selected receive FIFO. RDAT is considered valid only when RVAL is asserted. Data is received in big endian order.

D.8.16 RPRTY (receive parity)

RPRTY signal indicates the odd parity calculated over the RDAT bus.

D.8.17 RMOD[1:0] (receive word modulo)

RMOD is required only when RDAT is a 32-bit bus. RMOD indicates the number of valid bytes of data in RDAT. The RMOD bus should always be zero except during the last word transfer of a packet on TDAT. When REOP is asserted the number of valid packet data bytes on RDAT is decoded from RMOD as:

RMOD[1:0] = ì00îRDAT[31:0] is valid

RMOD[1:0] = ì01îRDAT[31:8] is valid

RMOD[1:0] = ì10îRDAT[31:16] is valid

RMOD[1:0] = ì11îRDAT[31:24] is valid

RMOD is considered valid only when RVAL is asserted

D.8.18 REOP (receive end of packet)

REOP is used to delineate the packet boundaries on the RDAT bus. When REOP is high, the end of the packet is present on the RDAT bus. REOP is required to be present at the end of every packets and is considered valid only when RVAL is asserted.

D.8.19 RERR (receive error indicator)

RERR is used to indicate that the current packet is in error. RERR shall only be asserted when REOP is asserted. Conditions that can cause RERR to be set may be, but are not limit to, FIFO overflow, abort sequence detection and FCS error.

D.8.20 RSX (receive start of transfer)

RSX indicates when the in-band port address is present on the RDAT bus. When RSX is high, the value of RADAT[7:0] is the address of the receive FIFO to be selected by the PHY. Subsequent data transfers on the RDAT bus will be from the FIFO specified by this in-band address. For single channel PHY devices the RSX signal is optional. For multi-port PHY devices, RSX must be asserted at the beginning of each transfer. When RSX is high RVAL must be low.

D.9 SPI-3 Management Functions

Management functions for this interface involves setting of FIFO thresholds used as triggers for flow control indicators.

D.10 SPI-4 Signaling function specifications

SPI-4 is an interface for packet and cell transfer between a physical layer (PHY) device and a link layer device, for aggregate bandwidths of OC-192 ATM and Packet over SONET/SDH (POS), as well as 10 Gb/s Ethernet applications.

On both the transmit and receive interfaces, FIFO status information is sent separately from the corresponding data path. By taking FIFO status information out-of-band, it is possible to decouple the transmit and receive interfaces so that each operates independently of the other. Such an arrangement makes POS-PHY L4 suitable not only for bidirectional but also for unidirectional link layer devices.

In both the transmit and receive interfaces, the packet's address, delineation information and error control coding is sent in-band with the data.

SPI-4 has the following general characteristics:

a) Point-to-point connection (i.e., between single PHY and single Link Layer device).
   1) Support for 256 ports (suitable for STS-1 granularity in SONET/SDH applications (192 ports) and Fast
   2) Ethernet granularity in Ethernet applications (100 ports)).
   3) Transmit / Receive Data Path:
      i) 16 bits wide.
      ii) In-band port address, start/end-of-packet indication, error-control code.
      iii) LVDS I/O (IEEE 1596.3 – 1996 [1], ANSI/TIA/EIA-644-1995 [2]).
      iv) 622 Mb/s minimum data rate per line.
      v) Source-synchronous double-edge clocking, 311 MHz minimum.

b) Transmit / Receive FIFO Status Interface:
   1) - LVTTL I/O or optional LVDS I/O (IEEE 1596.3 – 1996 [1], ANSI/TIA/EIA-644-1995 [2]).
   2) - Maximum 1/4 data path clock rate for LVTTL I/O, data path clock rate (double-edge
   3) clocking) for LVDS I/O.
   4) - 2-bit parallel FIFO status indication.
   5) - In-band Start-of-FIFO Status signal.
   6) - Source-synchronous clocking.

Data and control lines are driven off the rising and falling edges of the clock [TDCLK].

D.10.1 TDCLK (Transmit Data Clock)

TDCLK is a clock associated with TDAT and TCTL. TDCLK provides the datapath source-synchronous double-edge clocking with a minimum frequency of 311 MHz. Data and control lines are driven off the rising and falling edges of the clock. TDCLK is sourced by the MAC to the PHY.

D.10.2 TDAT[15:0] Transmit Data

TDAT is a 16-bit bus used to carry payload data and in-band control words from the Link Layer to the PHY device. A control word is present on TDAT when TCTL is high. The minimum data rate for TDAT is 622 Mb/s.

D.10.3 TCTL (Transmit Control)

TCTL is high when a control word is present on TDAT, otherwise it is low. TCTL is sourced by the MAC to the PHY.

D.10.4 TSCLK (Transmit Status Clock)

TSCLK is a clock associated with TSTAT providing source-synchronous clocking. For LVTTL I/O a maximum clockrate restraint is ¼ that of the data path clock rate. LVDS I/O allows a maximum of that equal to the data path clock (double-edge clocking).

D.10.5 TSTAT[1:0] (Transmit FIFO Status)

TSTAT is a 2-bit bus used to carry round-robin FIFO status information, along with associated error detection and framing. The maximum data rate for TSTAT is dependent on the I/O type, either LVDS or LVTTL, and is limited to its respective TSCLK restraints. TSTAT is sourced by the PHY to the MAC. The FIFO status formats are:

TSSTAT[1:0] = ì11î Reserved for framing or to indicate a disabled status link.

TSSTAT[1:0] = ì10î SATISFIED

TSSTAT[1:0] = ì01î HUNGRY

TSSTAT[1:0] = ì00î STARVING

D.10.6 RDCLK (Receive Data Clock)

RDCLK is a clock associated with RDAT and RCTL. RDCLK provides the datapath source-synchronous double-edge clocking with a minimum frequency of 311 MHz. Data and control lines are driven off the rising and falling edges of the clock. RDCLK is sourced by the PHY to the MAC.

D.10.7 RDAT[15:0] (Receive Data)

RDAT is a 16-bit bus which carries payload data and in-band control from the PHY to the Link Layer device. A control word is present on RDAT when RCTL is high. The minimum data rate for RDAT is 622 Mb/s.

D.10.8 RCTL (Receive Control)

RCTL is high when a control word is present on RDAT, otherwise it is low. RCTL is sourced by the PHY to the MAC.

D.10.9 RSCLK (Receive Status Clock)

RSCLK is a clock associated with RSTAT providing source-synchronous clocking. RSCLK is sourced by the Mac to the PHY. LVDS I/O allows a maximum of that equal to the data path clock (double-edge clocking).

D.10.10 RSTAT[1:0] (Receive FIFO Status)

RSTAT is a 2-bit bus used to carry round-robin FIFO status information, along with associated error detection and framing. The maximum data rate for RSTAT is dependent on the I/O type, either LVDS or LVTTL, and is limited to its respective RSCLK restraints. RSTAT is sourced by the Mac to the PHY. The FIFO status formats are:

TSSTAT[1:0] = ì11î Reserved for framing or to indicate a disabled status link.

TSSTAT[1:0] = ì10î SATISFIED

TSSTAT[1:0] = ì01î HUNGRY

TSSTAT[1:0] = ì00î STARVING

D.11 SPI-4 Management Functions

Parameter	Definition	P	CH	Units
CALENDAR[i]	Port address at calendar location i.	Yes	I	(N/A)
CALENDAR_LEN	Length of the calendar sequence.	Yes	I	(N/A)
CALENDAR_M	Number of times calendar sequence is repeated between insertions of framing pattern.	Yes	I	(N/A)
MAX_CALENDAR_LEN	Maximum supported value of MA_CALENDAR_LEN	No	I	(N/A)
MaxBurst1	Maximum number of 16 byte blocks that the FIFO can accept when FIFO Status channel indicates Starving.	Yes	C/I	16 byte blocks
MaxBurst2	Maximum number of 16 byte blocks that the FIFO can accept when FIFO Status channel indicates Hungry. MaxBurst2 <= MaxBurst1	Yes	C/I	16 byte blocks
a	Number of repetitions of the data training sequence that must be scheduled every DATA_MAX_T cycles.	Yes	I	(N/A)
DATA_MAX_T	Maximum interval between scheduling of training sequences on Data Path interface.	Yes	I	Cycles
FIFO_MAX_T	Maximum interval between scheduling of training sequences on FIFO Status Path interface.	Yes	I	Cycles

P = Provisionable, CH = Per channel (C) pr per interface (I)

Upon reset, the FIFOs in the datapath receiver are emptied, and any outstanding credits are cleared in the data path transmitter. After reset, but before active traffic is generated, the data transmitter shall send continuous training patterns. Transmission of training patterns shall continue until valid information is received on the FIFO Status Channel. The receiver shall ignore all incoming data until it has observed the training pattern and acquired synchronization with the data. Synchronization may be declared after a provisionable number of consecutive correct DIP-4 codewords are seen. Loss of synchronization may be reported after a provisionable number of consecutive incorrect DIP- 4 codewords is detected.

After reset but before active traffic is generated, the FIFO Status Channel transmitter shall send a continuous "1 1" framing pattern for LVTTL implementations, or continuous training patterns for optional LVDS implementations. Once the corresponding data channel has achieved synchronization, and a calendar has been provisioned, it may begin transmission of FIFO Status information. Once the data transmitter has received valid FIFO Status information (as indicated, for example, by a sufficient number of consecutively correct DIP- 2 codewords), it may begin transmission of data bursts to channels that have been provisioned and have space available.

In the event that the data path receiver is reset but the transmitter is still active, events at the receiver follow the same behavior as above. It shall ignore all incoming data until it has observed the training pattern and acquired synchronization with the data. It shall also send a continuous "1 1" framing pattern for LVTTL

implementations (or continuous training patterns for optional LVDS implementations) on its FIFO Status Channel, cancelling previously granted credits and setting them to zero. In this case the transmitter should send continuous training patterns to facilitate reacquistion by the receiver.

In the event that the data path transmitter is reset but the receiver is still active, events at the transmitter follow the same behavior as above. The transmitter shall send continuous training patterns until a calendar is configured and valid status information is received on the FIFO Status Channel. At the same time, the receiver may have lost synchronization with the data, and begun sending continuous framing patterns (or continuous training patterns for optional LVDS implementations) on the FIFO Status Channel. Once the data transmitter has received valid FIFO Status information (as indicated, for example, by a sufficient number of consecutively correct DIP- 2 codewords), it may begin transmission of data bursts to channels that have been provisioned and have space available.

Annex EPhysical MAC Client Interface

Annex FMIB

(normative)

Annex GBridging Conformance

(normative)

G.1 Bridging Overview

This section of the draft is not intended to define a new Bridging specification, but simply to extend the 802.17 MAC definition to demonstrate conformance to Transparent and VLAN Bridging , as defined in the IEEE 802.1D and 802.1Q standards respectively.

The MAC Bridging reference model for 802.17 is shown in Figure 1-1. The stations on the Ring act as the MAC Bridge, while the Ring acts as the shared media. Stations acting as Bridges are configured to do so through the appropriate Layer Management function.

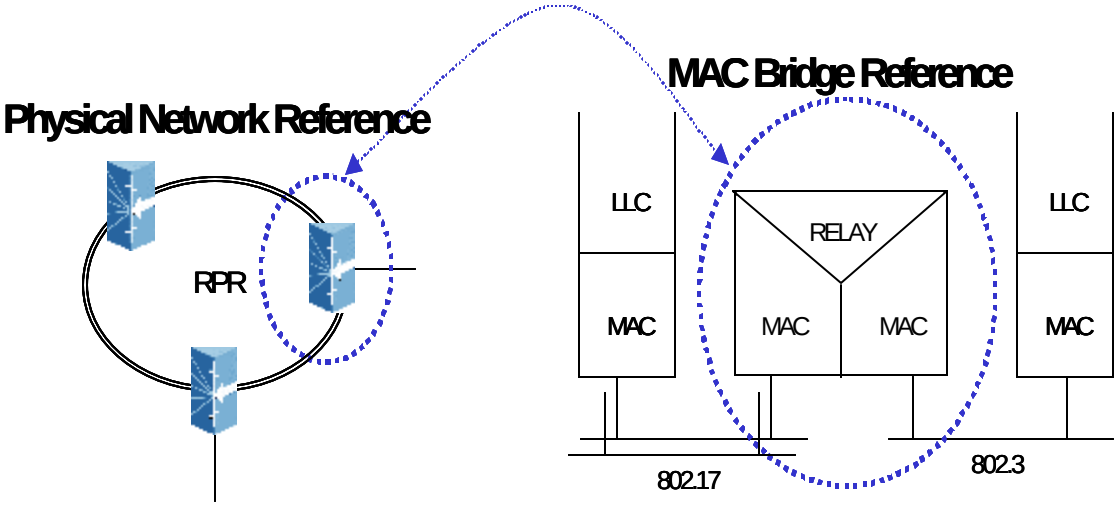


Figure C.1—Bridge Architecture Reference

The 802.17 MAC conformance to the aforementioned Bridging standards[1] can be achieved with the following 802.17 MAC capabilities:

1) The 802.17 MAC must provide an Internal Sub-Layer Service (ISS), which is used to interface with the Bridging Relay Entity. The ISS will conform to Section 6.4 of the IEEE Std 802.1D and Section 7.1 of the IEEE Std 802.1Q, when appropriate.
2) The 802.17 MAC must be able to communicate with the Bridge Protocol Entity via the LLC sub-layer, in conformance with the Bridging specifications.
3) The 802.17 MAC must be able to receives frames from the Ring (i.e., shared media) and determine whether they need to be bridged or not.
4) The 802.17 MAC must be able to transmit bridged frames appropriately over the 802.17 shared media. This includes handling of unknown unicast, broadcast, and multicast frames.

[1] IEEE 802.1D and 802.1Q Standards.

G.2 802.17 MAC Internal Sub-Layer Service

The ISS is provided by a MAC Entity to communicate with the MAC Relay Entity. The interface for this sub-layer is predefined in the 802.1D and 802.1Q specifications. The 802.17 MAC will adhere to these specifications.

G.2.1 802.17 MAC Support of Internal Sub-Layer Service

The 802.17 MAC access method is specified in the draft. Clause X specifies the MAC frame structure, and Clause Y specifies the MAC method.

On receipt of a MA_UNITDATA.request primitive, the local 802.17 MAC Entity performs Transmit Data Encapsulation, assembling a frame using the parameters supplied as specified below.

On receipt of an 802.17 MAC frame by Receive Media Access Management, the MAC frame is passed to the Reconciliation sub-layer which disassembles the frame into parameters, as specified below, that are supplied with the MA_UNITDATA.indication primitive.

The **frame_type** parameter takes only the value *user_data_frame* and is not explicitly encoded in MAC frames.

The **mac_action** parameter takes only the value *request_with_no_response* and is not explicitly encoded in MAC frames.

The **destination_address** parameter is either the address of an individual MAC entity or a group of MAC entities.

The **source_address** parameter is the individual address of the source MAC entity.

The **mac_service_data_unit** parameter is the service user data.

The **user_priority** parameter provided in a data request primitive is not encoded in the 802.17 MAC frame. The user_priority parameter provided in a data indication primitive takes the value of the Default User Priority parameter for the Port through which the MAC frame was received.

The **access_priority** parameter provided in a data request primitive is derived by a fixed 802.17 hod mapping. The fixed mapping is depicted in Table, “The access_priority parameter provided in a data request primitive is derived by a fixed 802.17 hod mapping. The fixed mapping is depicted in Table, “The access_priority parameter provided in a data request primitive is derived by a fixed 802.17 hod mapping. The fixed mapping is depicted in Table, “The access_priority parameter provided in a data request primitive is derived by a fixed 802.17 hod mapping. The fixed mapping is depicted in Table, “The access_priority parameter provided in a data request primitive is derived by a fixed 802.17 hod mapping. The fixed mapping is depicted in Table, “The access_priority parameter provided in a data request primitive is derived by a fixed 802.17 hod mapping. The fixed mapping is depicted in Table , “The access_priority parameter provided in a data request primitive is derived by a fixed 802.17 hod mapping. The fixed mapping is depicted in . The values shown are not modifiable my management or other means.,” on page181. The values shown are not modifiable my management or other means.,” on page180. The values shown are not modifiable my management or other means.,” on page 180. The values shown are not modifiable my management or other means.,” on page 180. The values shown are not modifiable my management or other means.,” on page180. The values shown are not modifiable my management or other means.,” on page180. The values shown are not modifiable my management or other means.

User Priority	Outbound Access Priority per MAC method									
	802.3	802.17	8802-4	8802-5 (default)	8802-5 (alternate)	8802-6	802.9a	8802.11	8802-12	FDDI
0	0	0	0	0	4	0	0	0	0	0
1	0	1	1	1	4	1	0	0	0	1
2	0	2	2	2	4	2	0	0	0	2
3	0	3	3	3	4	3	0	0	0	3
4	0	4	4	4	4	4	0	0	4	4
5	0	5	5	5	5	5	0	0	4	5
6	0	6	6	6	6	6	0	0	4	6
7	0	7	7	6	6	7	0	0	4	6

The **frame_check_sequence** parameter is encoded in the FCS field of the MAC frame. The FCS is computed as a function of the destination address, source address, length, RPR Header, and data fields. If a MA_UNITDATA.request primitive is not accompanied by this parameter, it is calculated in accordance with Clause Z of this draft.

FigureC.2 below shows the mapping of the MA-UNITDATA.request primitive parameters to the 802.17 MAC frame fields, and the mapping of the 802.17 MAC frame fields to the MA-UNITDATA.indication primitive parameters.

Table C.1—Outbound Access Priorities

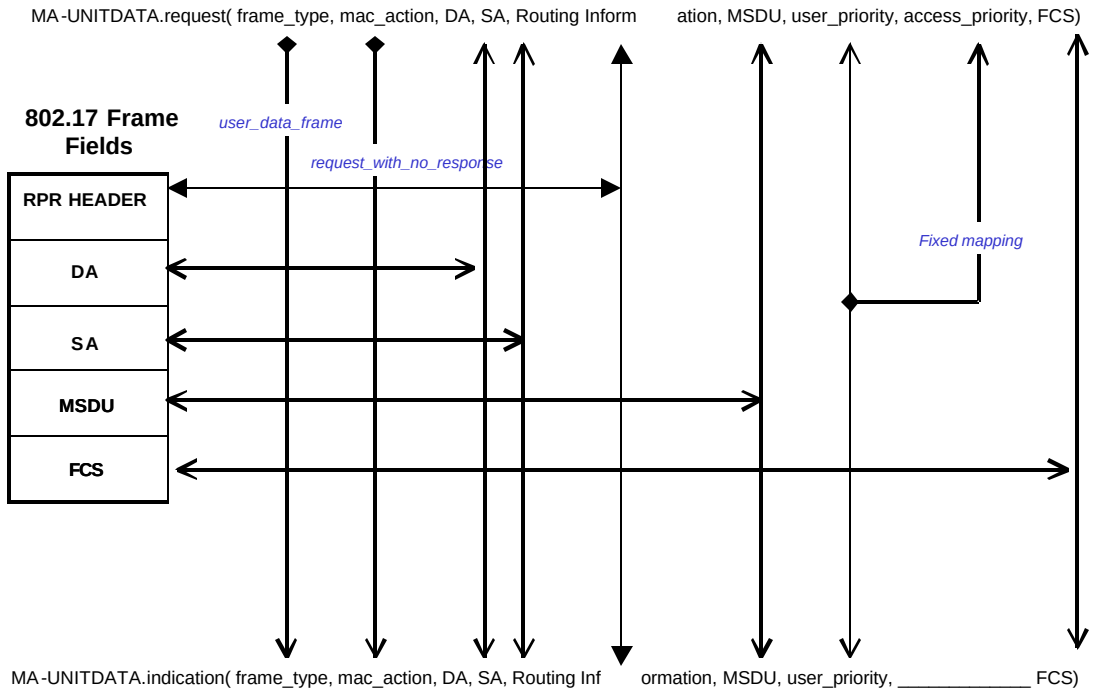


Figure C.2—Mapping of MAC Service Primitives

G.2.2 802.17 MAC Support of Enhanced Internal Sub-Layer Service

An Enhanced Internal Sub-Layer Service (E-ISS) is derived from the Internal Sub-Layer Service (ISS, defined in ISO/IEC 15802-3, 6.4) by augmenting that specification with elements necessary to the operation of the tagging and un-tagging functions of the MAC Bridge. The E-ISS provided by the 802.17 MAC will conform to Section 7.1 of IEEE 802.1Q Std.

FigureC.3 below shows the mapping of the EM-UNITDATA.request primitive parameters to the 802.17 MAC frame fields, and the mapping of the 802.17 MAC frame fields to the EM-UNITDATA.indication primitive parameters.

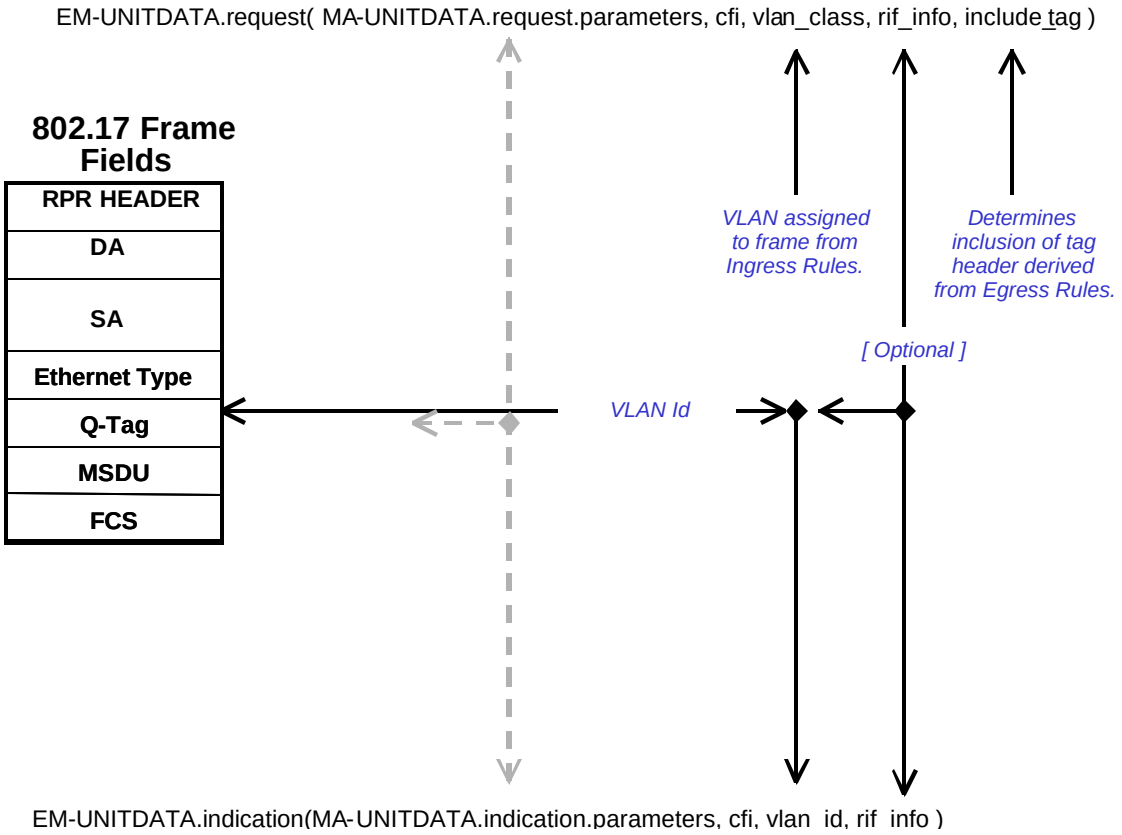


Figure C.3—Mapping of Enhanced MAC Service Primitives

G.3 Bridge Protocol Entity Interactions

The 802.17 MAC will provide a MAC sub-layer that will conform to Section 2.2.2 of IEEE 802.2 Std.

G.4 802.17 MAC Handling of Frames to be Bridged

The 802.17 MAC transit data path needs to incorporate logic to determine whether the received frame should be:

Dropped. The frame is stripped from the Ring and passed to a MAC client.
Discarded. The frame is stripped from the Ring and not passed to any MAC client. The frame is discarded.
Passed Through. The frame is passed to the tandem buffer and dispatched to the outgoing Ringlet.
Replicated. The frame is replicated prior to the transit path drop point. One copy of the frame is Passed Through, and the other copy is Dropped.

FigureC.4 depicts a simplified depcition of the 802.17 MAC transit data path. Refer to Clause W of the 802.17 Draft for a complete description of the 802.17 MAC transit path.

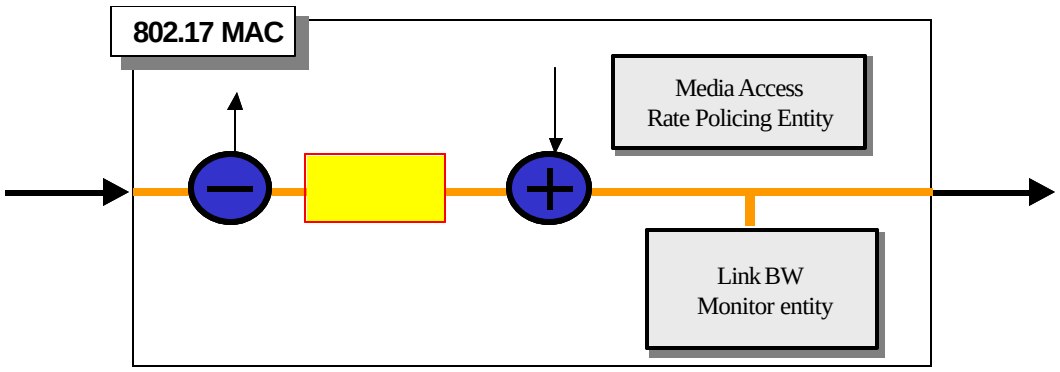


Figure C.4—Simplified 802.17 MAC Transit Path

When Bridging is provision on the RPR station, the Layer Management Element sets a state used by the 802.17 MAC transit path to indicate that Bridging is configured. The Drop/Discard point in the 802.17 MAC transit path needs to incorporate the following logic:

a) If the Destination Address (DA) associated with the received frame is that of the RPR Station, and Bridging is configured on the Station, the frame is Dropped and passed to the MAC Relay Entity.
b) If the DA of the received frame is not that of the RPR Station, and Bridging is configured on the Station, the frame is Replicated.

The aforementioned logic is an extension of the RPR MAC reception rules outlined in Clause W of the 802.17 Draft.

G.5 802.17 MAC Transmission of Bridged Frames

Bridge relayed frames are submitted for transmission by the Bridging Forwarding Process. The Service request primitive associated with such a frame conveys the values of the source and destination address fields received in the Service indication primitive. Refer to Figure 1-2 and Figure 1-3 for the mappings.

Bridged frames with a Multicast and Broadcast destination address are broadcast around the RPR.

Bridged frames with a destination address of a station on the RPR (i.e., a known destination address) is forwarded to the station using internal 802.17 MAC topology and steering tables.

Bridged frames with an unknown destination address (e.g., a destination address not matching a RPR station address) are flooded over the RPR.

G.5.1 Flooding Packet over 802.17

An 802.17 MAC floods a packet of the 802.17 shared media, by replicating and dispatching the packet(s) over both directions of the Ring. The TTL field, found in the RPR header, is set such that each station on the Ring only sees the packet once. Figure 1-5 below illustrates the operation of flooding a packet over the RPR.

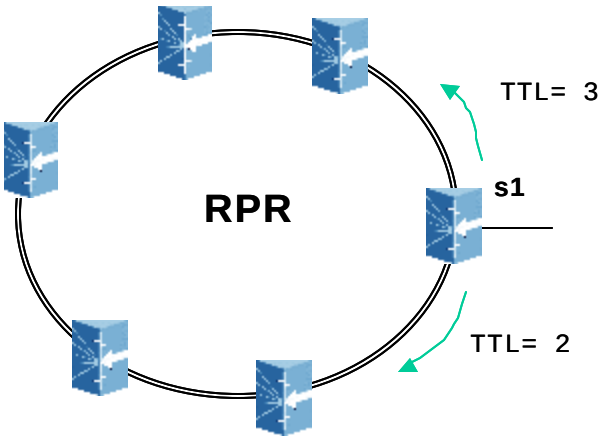


Figure C.5—Flooding Packets over RPR

Annex HCRC Calcuation

(normative)

Annex IStratum Clock Distribution

(normative)

Annex JCode Examples

(informative)

J.1 RPR-fa C code example

Variables:

```
typedef unsigned short Uint16;
typedef unsigned long Uint32;
typedef short Int16;

Uint32 lo_tb_depth; //low priority transit buffer depth
Uint32 add_rate;    //count of low priority (LP) and excess medium
                    //priority eMP octets transmitted by client
Uint32 add_rate_congestion; //count of LP and eMP octets transmitted by
                   //client and destined beyond the congestion point
Uint32 lp_add_rate;    //add_rate run through a low pass filter
Uint32 nlp_add_rate; //lp_add_rate / WEIGHT

Uint32 lp_add_rate_congestion; //add_rate_congestion run through a low
                               //pass filter
Uint32 nlp_add_rate_congestion; //lp_add_rate_congestion / WEIGHT
boolean add_rate_ok; //flag indicating that client is allowed to trans-
mit
Uint32 allow_rate_congestion; //the fair amount each node is allowed to
                              //transmit beyond the congestion point
Uint32 fwd_rate;  //count of LP+eMP octets forwarded from the LP
                  //transit buffer
Uint32 lp_fwd_rate; //fwd_rate run through low pass filter
Uint32 rate_iMP;  //count of iMP octets forwarded and added
Uint32 lp_rate_iMP; // rate_iMP run through low pass filter
Uint32 lp_allow_congestion; // allow_rate_congestion run
                            // through low pass filter
boolean congested; //station experiences congestion based on transit
            //buffer occupancy, link utilization or HoL timer expiration
Uint32 access_delay;

Uint16 rcvd_advertised_rate; //the fair rate(# of bytes in a decay
                       // interval)received from the downstream neighbor
Uint16 advertised_rate;  //the fair rate (# of bytes in a decay
                      // interval) passed along to the upstream neighbor
Int16 token_octets;  //the number of octets in RPR-fa dynamic
             // shaper/policer for eMP and LP traffic beyond congestion
point
Int16 token_octets_low; //the number of octets in RPR-fa
                        //shaper/policer for low priority
Int16 token_octets_med; // the number of octets in RPR-fa
                        //shaper/policer for medium (iMP+eMP) priority
Int16 token_octets_high; // the number of octets in RPR-fa
                         //shaper/policer for high priority
Uint16 Ns;  // Number of active sources
```

```
Boolean add_packet_enqueued_high, add_packet_enqueued_low;

typedef struct _fairness_pkt_t {
char[6] SA;
bit RI;
Uint16 rate;
} fairness_pkt_t;
fairness_pkt_t fairness_pkt; //received fairness packet

enum states {"FAIRNESS_OFF",
             "FAIRNESS_ON",
             "RAMP_UP",
             "RAMP_DOWN",
             "DOWNSTREAM_CONGESTED" };
state current_state, next_state;
boolean initial;


Constants (Tunable, provisioned LME objects):
Uint16 WEIGHT; //configured weight for this station
               //an integer between 1 and 63
Uint16 w[n];   // weight vector for all stations on the ring
Uint16 BUCKET_SIZE;    // provisioned RPR-fa dynamic shaper
                       //leaky bucket size in octets
Uint32 MAX_ALLOWANCE; //configured value for max allowed rate for this
node
Uint32 DECAY_INTERVAL; //   8,000 octet times @ OC-12,
                       //  32,000 octet times @ OC-48,
                       // 128,000 octet times @ OC-192
Uint32 ADVERTISEMENT_INTERVAL;//between 1 MTU time and DECAY_INTERVAL

Uint16 AGECOEFF = 4; // Aging coeff for add_rate and fwd_rate
Uint16 LP_FWD = 64; // Low pass filter for fwd_rate
Uint16 LP_MU = 512; // Low pass filter for my fair rate
Uint16 LP_MU_history = 511 //Low pass filter for history rate
Uint16 LP_ALLOW = 64; // LP filter for allow rate auto increment
Uint16 LP_ALLOW_COEF = 128; // low-pass filter for lp_allow
Uint16 NULL_RATE = 0xFFFF; //All 1's in rcvd_advertised_rate field.

Uint16 TB_LO_THRESHOLD; // TB depth at which no more LP
                        //client traffic can be sent
Uint32 MAX_LRATE; //AGECOEFF * DECAY_INTERVAL =
                  // 512,000 for OC-192
                  // 128,000 for OC-48
                  //  32,000 for OC-12

Uint32 AD_THRESHOLD; // default value 1ms


Uint32 reserved_rate; //high priority reserved rate
                      //(# of bytes in a decay interval
//Each station should have knowledge of reserved rate

Bit mode; // Advertising mode: 1 conservative, 0 (default) aggressive
```

THESE ARE UPDATED EVERY CLOCK CYCLE:

```
// add_rate is increment by 1 for every LP/eMP octet that is
// transmitted by the client (does not include data
// transmitted from the Transit Buffer).

// add_rate_congestion is increment by 1 for every LP/eMP octet that is
// transmitted by the client (does not include data
// transmitted from the Transit Buffer) and destined beyond congestion
// point (i.e., TTL > TTL_to_congestion).

// fwd_rate is increment by 1 for every LP/eMP octet that exits the
// LP Transit Buffer

// token_octets is decremented by 1 for every LP/eMP octet that is
// transmitted by the client.
//A packet is sent to completion, why is the below check here?
if ((add_rate_congestion < allow_rate_congestion) &&
    (fwd_rate+add_rate+rate_iMP < MAX_LRATE - reserved_rate) &&
    (token_octets > 0) &&
    !((lo_tb_depth > 0) && (WEIGHT * fwd_rate < add_rate)) &&
    (add_rate < MAX_ALLOWANCE)){
    add_rate_ok = TRUE; // true means OK to send client packets
}

// access_delay is started when an SOP arrives
// access_delay is reset when packet is transmited
// access_delay will not increase when a packet is enqueued

if (add_packet_enqueued)
    access_delay++;   // Any packet being added except high priority


if (access_delay >= AD_THRESHOLD)  //access delay expires
    congested = TRUE;
```

UPDATED WHEN FAIRNESS_PKT IS RECEIVED:

```
if ((fairness_pkt.SA == my_SA) &&
    ((node_state == wrapped || fairness_pkt.RI == my_RI))){
    rcvd_advertised_rate = NULL_RATE;
    }else{
    rcvd_advertised_rate = fairness_pkt.rate;
    }

steering case:
   do not forward on,
   for wrapping case pass set
treat single link as segment failure.
```

THE FOLLOWING IS CALCULATED EVERY DECAY_INTERVAL:

```
if((lo_tb_depth > TB_LO_THRESHOLD/2)|| // LPTB crosses threshold
   ((add_rate + fwd_rate) >
```

```
    (MAX_LRATE - reserved_rate))|| //link utilization cross high thresh-
old
    (access_delay >= AD_THRESHOLD)){ // access delay
    congested = TRUE;
} else {
    congested = FALSE;
}

lp_add_rate = (lp_add_rate/(LP_MU_-1/LP_MU)+
              (add_rate / LP_MU)); //more generic weights than LP_MU-1

nlp_add_rate = lp_add_rate/WEIGHT; // modified weight function
//add_rate is decremented by min(allow_rate/AGECOEFF, add_rate/AGECOEFF)

lp_fwd_rate = ((LP_FWD-1)*lp_fwd_rate+fwd_rate)/LP_FWD;
// fwd_rate is decremented by fwd_rate/AGECOEFF
//(Note: lp values calculated prior to decrement of non-lp values).

// token_octets is incremented at a rate which equals to
// lp_allow /(AGECOEFF * DECAY_INTERVAL)

lp_allow = ((LP_ALLOW_COEF-1)*lp_allow+allow_rate_congestion)/
            LP_ALLOW_COEF;

if(rcvd_advertised_rate!=NULL_RATE){
    allow_rate_congestion = (rcvd_advertised_rate*WEIGHT);
}else{
    allow_rate_congestion += (MAX_LRATE-reserved_rate-
                              allow_rate_congestion)/(LP_ALLOW);
}

if (fairness_on && mode == 1)
    allow_rate_congestion = advertised_rate * WEIGTH;

// fwd_rate_iMP is increment by 1 for every iMP octet that
// exits the LP Transit Buffer

THE FOLLOWING IS CALCULATED EVERY ADVERTISEMENT_INTERVAL:
advertise();


=====================================================================

if((add_rate + fwd_rate) <
   (MAX_LRATE - reserved_rate - hysteresis)){
    Below_Low_Threshold = TRUE;
    Above_Low_Threshold = FALSE;
} else {
    Below_Low_Threshold = FALSE;
    Above_Low_Threshold = TRUE;
}

if((add_rate + fwd_rate) >=
   (MAX_LRATE - reserved_rate)){
```

```
    Above_High_Threshold = TRUE;
    Below_High_Threshold = FALSE;
}else{
    Below_High_Threshold = TRUE;
    Above_High_Threshold = FALSE;
}

advertise (){          // rate advertisement machine
next_state = current-state;    // not change
switch(current_ state){
    case FAIRNESS_OFF:
        if(rcvd_advertised_rate != NULL_RATE)
            //downstream_congested
            next_state = DOWNSTREAM_CONGESTED;
            //next state is downstream congested

        if (congested)
            next_state = FAIRNESS_ON;
            // next state is congested

        advertised_rate = NULL_RATE;
        // advertises is still maximum bw

        //allow_rate_congestion +=
        //(MAX_LRATE -allow_rate_congestion)/(LP_ALLOW);
        initial = TRUE;

    case FAIRNESS_ON:
        if(initial){
       advertised_rate = mode ? nlp_add_rate :
                              (MAX_LRATE - reserved_rate)/(F1);
        // calculated initial rate:fair rate per unit weight
        initial = FALSE;
        }

        // advertise min(low pass add rate, received
        //              advertise rate)

        if(nlp_add_rate < rcvd_advertised_rate){
            advertised_rate = (nlp_add_rate);
        }else{
            advertised_rate = rcvd_advertised_rate;
        }

        if(rcvd_advertised_rate != NULL_RATE)
            // downstream_congested
            next_state = DOWNSTREAM_CONGESTED;
            // next state is downstream congested

        if(Below_Low_Threshold)
            next_state = RAMP_UP;

        //allow_rate_congestion
```

```
    case RAMP_UP:
        advertised_rate = mode ? NULL_RATE:
                          lp_add_rate +lp_add_rate*I/J;

        if(advertised_rate > (MAX_LRATE- reserved_rate))
            advertised_rate = NULL_RATE;

        if(Above_High_Threshold)
            next_state = RAMP_DOWN;

        if(Above_Low_Threshold)
            next_state = FAIRNESS_ON;

        if(advertised_rate == NULL_RATE)
            next_state = FAIRNESS_OFF;

        //allow_rate_congestion

    case RAMP_DOWN:
        advertised_rate = mode ? lp_add_rate:
                          lp_add_rate - lp_add_rate *I/J;

        if(Below_High_Threshold)
            next_state = FAIRNESS_ON;

        if(Below_Low_Threshold  )
            next_state = RAMP_UP;

        //allow_rate_congestion

    case DOWNSTREAM_CONGESTED:
         if(lp_fwd_rate > (allow_rate_congestion/WEIGHT))
             advertised_rate = rcvd_advertised_rate;
         else
        advertised_rate = NULL_RATE;

         allow_rate_congestion = (rcvd_advertised_rate*WEIGHT);
         // WEIGHT is local weight * received control message

         if(rcvd_advertised_rate = NULL_RATE)
             next_state = FAIRNESS_OFF;

         current_state = next_state;
     }
}

Uint16
F1(weight_station) {
    // input weight_station;
    // mode 1: weight == station_weight;

    Uint16 weight_total_active = 0; // local variable weight_total_active
    Uint16 weight_total = 0;    // local variable weight_total
```

```
    for (i = 1; Ns; ++i ){          // Ns is number of active sources
   weight_total_active += w[i];}  //add all active weight global vari-
able
    weight_total_active+= weight_station; // add local station weight
    return weight_total_active;
}
```

Annex KImplementation Guildelines

(informative)

K.1 MAC client behavior

The behavior of MAC client is presented for descriptive purposes and this section does not impose any behavior for MAC client. RPR standard defines a set of primitives at the MSAP interface. The number of queues and the queue managers at the MAC client are a matter of choice.

The simplest MAC client can have one queue for each traffic class. The STOP_LO, STOP_MED, STOP_HI signals will indicate the traffic class that cannot be sent. If MAC client sends a packet of a stopped traffic class, the MAC policing functionality will not allow any more packets to be sent until that traffic class is allowed. However, MAC client can decide to send a medium priority packet as an excess bandwidth packet, in which case MAC will treat that packet as a low priority packet and the status of STOP_LO signal will be important. This means that a MAC client is allowed to send a medium priority packet even when there is STOP_MED signal, provided that STOP_LO is not asserted. When a medium priority packet must be treated as low priority, MAC will mark the IOP bit in the header as out-of-profile, and the packet will consume rate shaping and fairness credits of the low priority traffic class in the MAC. It is possible to starve low priority traffic by sending excess medium traffic in place of low priority traffic. MAC Client should choose, how much it could schedule excess medium priority traffic to starve or not to starve low priority traffic.

An example client can behave as follows:

Table D.1—MAC client interface

Stop_L	Stop_M	Stop_H	Which queue to select
0	0	0	If there is a packet in high class traffic queue, schedule high class. If there is a packet in medium class traffic queue, schedule medium class. If there is a packet in low class traffic queue, schedule low class
0	0	1	If there is a packet in medium class traffic queue, schedule medium class. If there is a packet in low class traffic queue, schedule low class.
0	1	0	If there is a packet in high class traffic queue, schedule high class If there is a packet in medium class traffic queue, schedule medium class(will be treated as low priority) If there is a packet in low class traffic queue, schedule low class
0	1	1	If there is a packet in medium class traffic queue, schedule medium class(will be treated as low priority) If there is a packet in low class traffic queue, schedule low class
1	0	0	If there is a packet in high class traffic queue, schedule high class If there is a packet in medium class traffic queue, schedule medium class
1	0	1	If there is a packet in medium class traffic queue, schedule medium class
1	1	0	If there is a packet in high class traffic queue, schedule high class
1	1	1	Stop scheduling any more packets

Optionally, MAC client may implement a more sophisticated queuing scheme to avoid head-of-line blocking and to utilize more bandwidth. This can be accomplished through the use of network congestion information transmitted by the nodes on the network and the collected topology information. For this purpose MAC client can implement virtual output queues for each destination on the ring for low and/or medium priority.

MAC client is allowed to send a packet from a virtual output queue for low priority or excess medium priority queues if it can satisfy the necessary condition, which is for each congestion point before the selected destination, the total usage beyond the congestion point should be less than the congested nodes fairness value.

At any time there can be more than one virtual output queue that will satisfy the condition, in this case a round robin approach can be chosen to simplify the solution. However, a better approach will be using deficit round robin, which will avoid possible unfairness among virtual output queues.

The calculations of queue add rates and allowed rates beyond a congested point are also important factors to increase utilization and obtain a stable behavior. An acceptable approach is to choose a similar algorithm to update and increment these values as in RPR MAC client for "allowed_rate" and "add_rate" for each virtual output queue. In addition one should low pass filter the value of per queue add rates to smooth out instantaneous variations. Once MAC client receives a fairness message from MAC about a congested node, it will update the allowed_rate of that destination, which represents the total allowed_rate beyond that node. That value will then be increment periodically as long as MAC client does not receive another fairness message for that node. In essence, MAC Client implements a copy of the MAC fairness algorithm for each destination.

Depending on the client's behavior, the assertion of STOP signals will vary in the MAC. For virtual destination queuing, ideally STOP signals will never be asserted other than rate shaping purposes. If a client goes insane, MAC policing functions will prevent the client from abusing the ring.