Draft 0.4 P802.17/D0.4
November 8, 2001

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 - Gandalf

Draft 0.4 -November 8, 2001

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 MAU types 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

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.
365 March Road
Kanata, Ontario
Canada K2K 2C9
Phone: +1.613.271.3399
FAX: +1.613.271.3333
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:

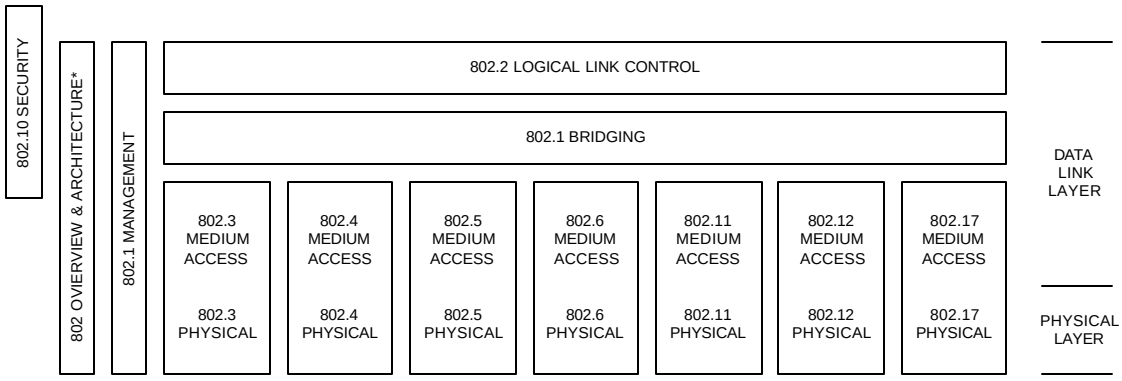
Gunes Aybay, Riverstone Networks
Mark Bordogna, Agere Systems
David Cheon, Sun Microsystems
Preminder Cohan, Infineon Technologies
Spencer Dawkins, Fujitsu Network Co
Martin Green, Cisco Systems
Yongbum Kim, Broadcom
Sateesh Kumar, Redwave Networks
Dave Meyer, Mindspeed
Gal Mor, Corrigent Systems
Chuck Lee, Appian Communications
Ashwin Moranganti, Appian Communications
Bob Sultan

Contributors:

Leon Bruckman, Corrigent Systems
Bob Castellano, Jedai Broadband Networks
Jim Kao, Cisco Systems
Carey Kloss, Cisco Systems
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

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

Proposals for Resilient Packet Ring (RPR)

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 scaleable LAN/MAN architecutre 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, scaleable 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 priorities
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 [1][2]
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 Application Areas

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

1.4 Conformance Requirements

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

1.5 Definitions

(? Definitions from Terms & Definitions committee to be added here?)

1.6 Abbreviations

(? Definitions from Terms & Definitions committee to be added here?)

1.7 References

[1] ANSI X3T9 FDDI Specification

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

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

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

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

(? Copy usual suspects from other standard + RPR specific ones here?)

1.8 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.

1.8.1 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 speci.ed here by describing the service primitives and parameters that characterize each service. This definition of service is independent of any particular implementation (see Figure1).

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.

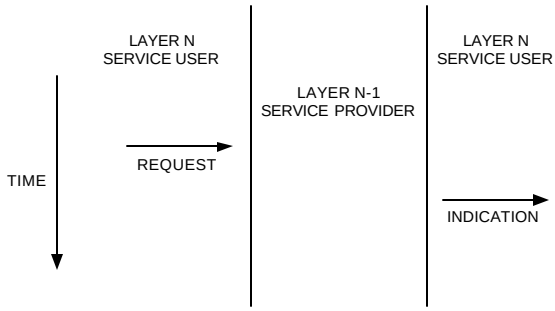


Figure 1—Service Definition

1.8.1.1 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 signi.cant 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 de.ned 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 de.ned or implicitly available.

1.8.2 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 Figure2)

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.

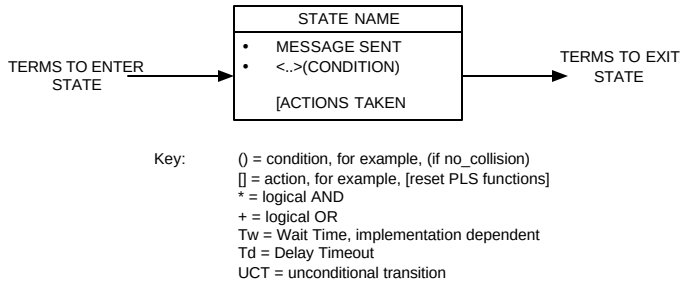


Figure 2—-State diagram notation example

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 (e.g., repeaters) 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.

1.9 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/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 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.

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.

2. Media Access Control (MAC) service specification

2.1 Scope

This clause specifies the services provided by the MAC sublayer and the MAC Control sublayer to the client of the MAC (see Figure3). 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 Figure4). 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.

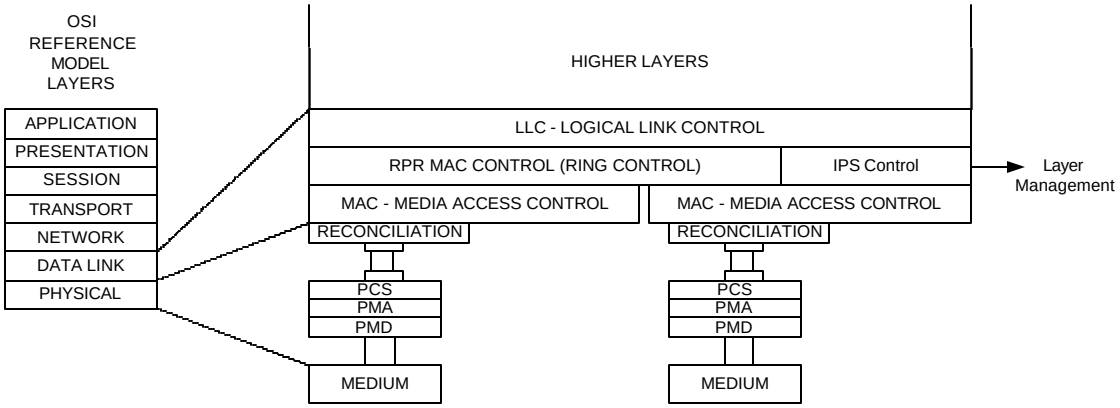


Figure 3—Service specification relation to the LAN model

2.2 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.

2.2.1 High Priority Service

The MAC provides a high priority delivery service bounded end-to-end delay and jitter specifications. This channel is intended to allow the client to implement a Synchronous Traffic class. 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.

2.2.2 Medium Priority Service

The Medium priority service is provided to implement a Guaranteed Traffic Class (GTC). 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, EIR and BIR limits. However the priority class on the ring transit path will be different depending on whether the particular frame is in or out of its agreed CIR/EIR/BIR profile. In profile frames will be delivered on the high priority, low-delay transit path while out of profile traffic will transit on the low priority, best effort path.

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.

2.2.3 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, but may or may not be tolerant to frame loss.

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.

2.3 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 two channels: high priority or low priority

2.3.1 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 channel.

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

2.3.2 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.

2.4 MAC services to the LLC

Two service primitives are defined for the LLC interfaces.

a) MA_UNITDATA.request
b) MA_UNITDATA.indication.
c) MA_CONTROL.request (used by MAC Control sublayer).
d) MA_CONTROL.indication (used by MAC Control sublayer).

2.5 MAC services to the IEEE 802.1D Bridge

Two service primitives are defined for the bridge interfaces.

a) M_UNITDATA.request
b) M_UNITDATA.indication.

The formats for the M_UNITDATA.indications and M_UNITDATA.requests are the same as formats for MA_UNITDATA.indication and MA_UNITDATA.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.

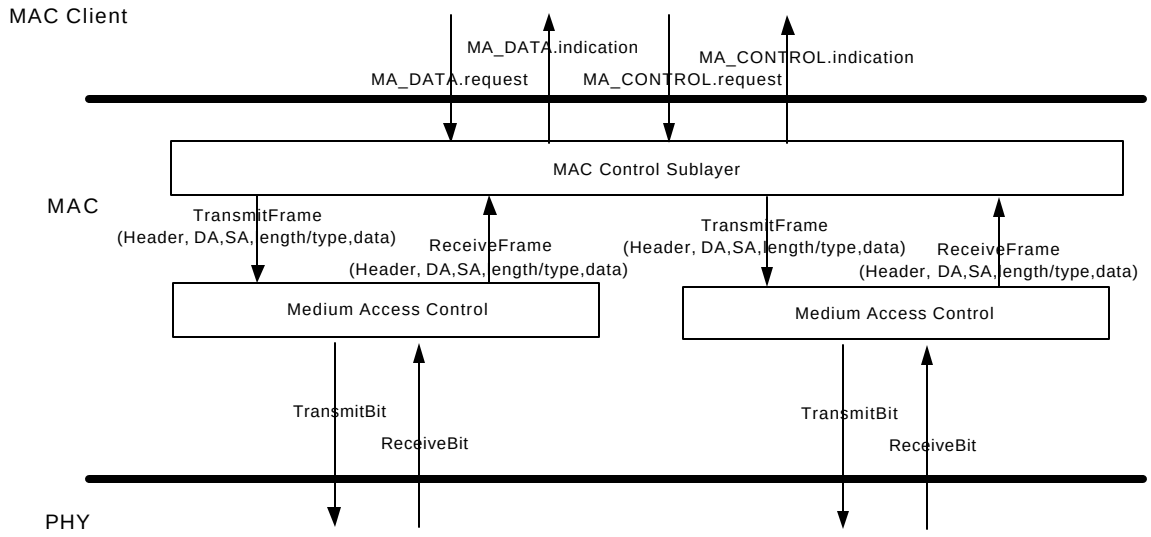


Figure 4—MAC service model

2.6 MAC client interface considerations

In its canonical form, an 802.17 network consists of dual, counter-rotating rings. 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.

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 Output Queueing to avoid head-of-line blocking of PDUs destined to nodes that are not physically situated beyond the point of congestion.

An RPR MAC will have processes to police 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 not accept the request and will return an indication to that effect. It is up to the MAC client to resubmit the request when the channel is available again and to ensure that data is not lost.

<There is one shaper each for HP, MP, and LP traffic. The shapers are simple token buckets, and if a bucket empties 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 interface between MAC and MAC client will be specified. in the MAC_CONTROL.indication. -- needs to be moved to MAC section>

2.6.1 Overview of interactions

MA_DATA.request

MA_DATA.indication

MA_CONTROL.request (used by MAC Control sublayer)

MA_CONTROL.indication (used by MAC Control sublayer)

2.6.2 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.

2.6.3 Detailed service specification

2.6.3.1 MA_DATA.request

2.6.3.1.1 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.

2.6.3.1.2 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 (see 2.3.1.5). The ringlet_id parameter 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.

2.6.3.1.3 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.

2.6.3.1.4 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.

2.6.3.1.5 Additional comments

The RPR MAC protocol provides three quality of services in service_class requested.

2.6.3.2 MA_DATA.indication

2.6.3.2.1 Function

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

2.6.3.2.2 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 who optionally use the information, which ringlet the m_sdu was received from.

2.6.3.2.3 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.

2.6.3.2.4 Effect of receipt

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

2.6.3.2.5 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).

2.6.3.3 MA_CONTROL.request

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

2.6.3.3.1 Function

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

2.6.3.3.2 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 pre appended to the frame by the local MAC sub-layer 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 page26.

2.6.3.3.3 When generated

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

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
0x05-0xFF	TBD	TBD

2.6.3.3.4 Effect of receipt

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

2.6.3.4 MA_CONTROL.indication

2.6.3.4.1 Function

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

2.6.3.4.2 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

2.6.3.4.3 When generated

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

2.6.3.4.4 Effect of receipt

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

3. Media access control frame structure

3.1 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.

Three 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 MAC frame format for Tagged MAC frames, i.e., frames that carry QTag

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. The minimum transfer length for data packets is 22octets.

These limits include everything listed in Figure5 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.

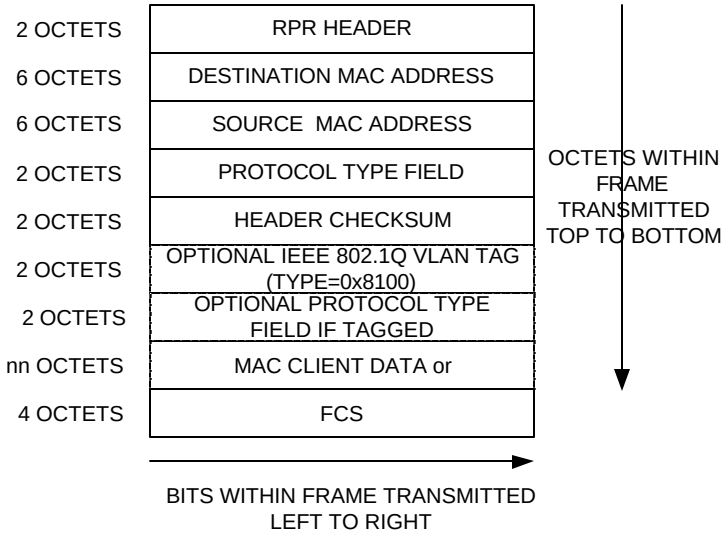


Figure 5—RPR Frame Format

3.2 RPR packet header format

Each packet has a fixed-sized header. The packet header format is shown in Figure6

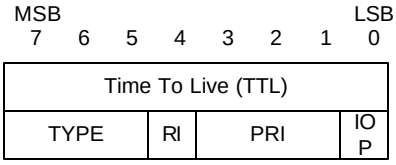


Figure 6—RPR Packet Header Format

The fields are described below:

3.2.1 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. This allows for a total node space of 256 nodes on a ring. However, due to certain failure conditions (e.g. when the ring is wrapped) the total number of nodes that are supported by RPR is 128. When a packet is first sent onto the ring the TTL should be set to at least twice the total number of nodes on the ring.

3.2.2 Type field

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

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

Table 3—Type Values

These modes will be further explained in later sections.

3.2.3 Ring Identifier

This bit indicates that the originated ring identifier for the packet.

3.2.4 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).

3.2.5 IOP

This bit could be used to mark if the packet is in or out of profiles.

3.3 Overall packet format

The overall packet format is show in Figure 5:

3.3.1 Destination address

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

3.3.2 Source address

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

3.3.3 Protocol type

The protocol type is a two-octet field like that used in Ethernet Type field representation. All value defined in Ether Type still is valid here besides there is some additional value will be allocated for RPR as in Table4

Value	Protocol Type
0x2007	RPR Control
0x0800	IP version 4
0x0806	ARP
0x8100	Vlan Tagged Frame
TBD	Payload with Customer Seperation ID

Table 4—Defined protocol type

3.3.4 HEC field

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

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

The HEC is computed over the RPR header, destination address, source address and protocol type.

3.3.5 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 computed over the whole payload. It does not include the RPR header.

3.3.6 Addressing

All nodes must have a globally unique IEEE 48 bit MAC address. A multicast bit is defined using canonical addressing conventions i.e. the multicast bit is the least significant bit of the most significant octet in the destination address. It is acceptable but not advisable to change a node's MAC address to one that is known to be unique within the administrative layer 2 domain (that is the RPR ring itself along with any networks connected to the RPR ring via a layer 2 transparent bridge).

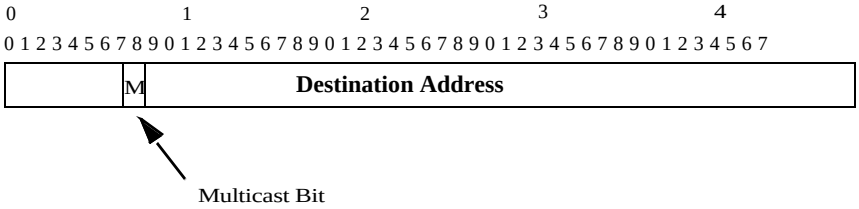


Figure 7—Multicast Bit Position

Note that for SONET media, the network order is MSB of each octet first, so that as viewed on the line, the multicast bit will be the 8th bit of the destination address sent. (For RPR on Ethernet media, the multicast bit would be sent first).

3.4 RPR control packet format

If the MODE bits are set to 101 then this indicates a control message. RPR control packet could be hop by hop or destined to a particular node. If the control packet is a hop by message, they are by definition unicast, and do not need any addressing information. The destination address field for control packets should be set to 0's. 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 Figure8.

Figure 8—Control Packet Format

The priority (PRI) value in RPR header shall be set to 0x7 (all one's) when sending control packets and should be queued to the highest priority transmit queue available. The Time to Live is not relevant since all packets will be received and stripped by the nearest downstream neighbor and can be set to any value (preferably this should be set to 001).

3.4.1 Control ver

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

3.4.2 Control type

This one octet field represents the control message type. Table5 contains the currently defined control types.

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

Table 5—Control Types

3.4.3 Control TTL

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

Note that the control layer hop count is separate from the RPR L2 TTL. The originator of the control message should set the initial value of the control TTL to the RPR L2 TTL normally used for data packets.

3.4.4 Payload

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

3.5 Order of bit transmission

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

3.6 Invalid RPR frame

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

a HEC is not match with the frame is received.
b FCS is 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.

3.7 Elements of tagged RPR frame

Tagged RPR frame format is shown as in Figure9. This format is an extension of the RPR frame format.

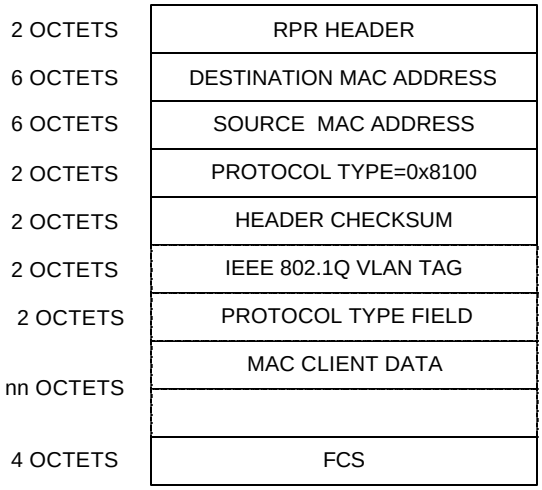


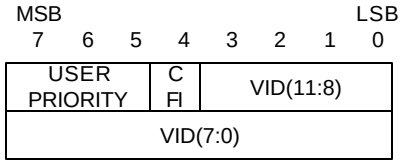
Figure 9—Tagged RPR Frame Format

3.7.1 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.

3.7.2 Tag Control Information field (informative)

The Tag Control Information field is subdivided as follows



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.

3.7.3 Payload Type

The Payload Type field contains the original Protocol Type from RPR frame prior to add the QTag Prefix.

3.7.4 RPR Fairness Frame Format

RPR Fairness Frame is sent to MAC neighbors to convey fairness algorithm specified in Claus 7.

2 OCTETS	RPR HEADER (TYPE=0x6)
6 OCTETS	SOURCE MAC ADDRESS
2 OCTETS	FAIRNESS CONTROL HEADER
2 OCTETS	CONTROL VALUE
4 OCTETS	FCS

4. Terms and Taxonomy

4.1 Ring terminology

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

The bidirectional ring allows for ring-wrapping in case of media or station failure, as in FDDI [1] or SONET/SDH [3]. The wrapping is controlled by the Protection Switching protocol.

To distinguish between the two rings, one is referred to as the “inner” ring, the other the “outer” ring. 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 ring.

4.2 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 Figure27. 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

4.3 Transit buffer

To be able to detect when to transmit and receive packets from the ring, RPR makes use of a transit buffer as shown in Figure10 below. There are two optional implementations, using either one or two transit buffers. 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

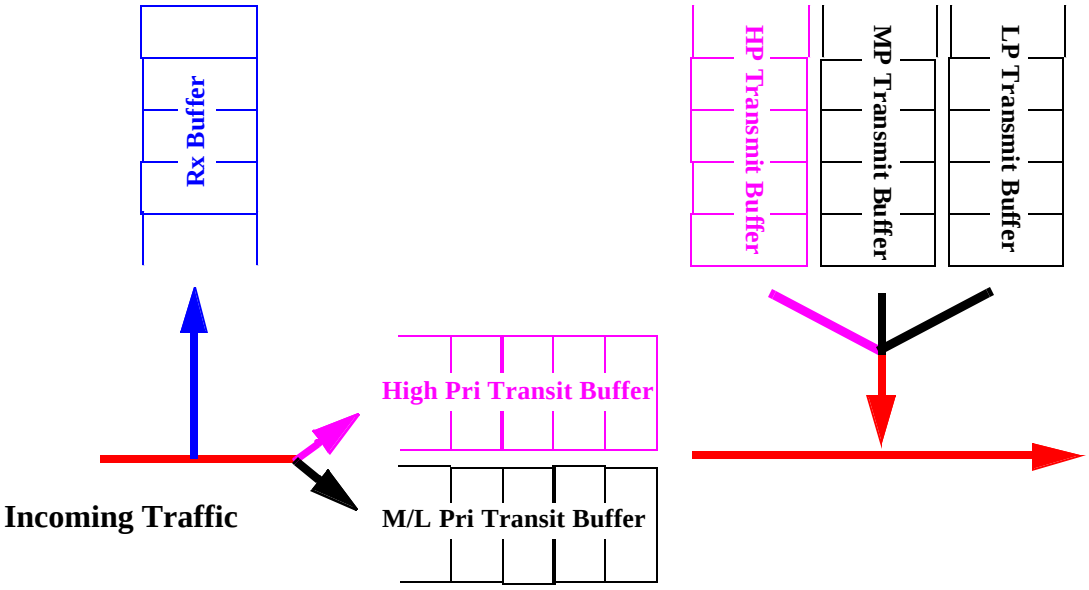


Figure 10—Transit / Transmit Buffer Design

transit buffer case, only one buffer will be used for all three priorities.

5. Media Access Control

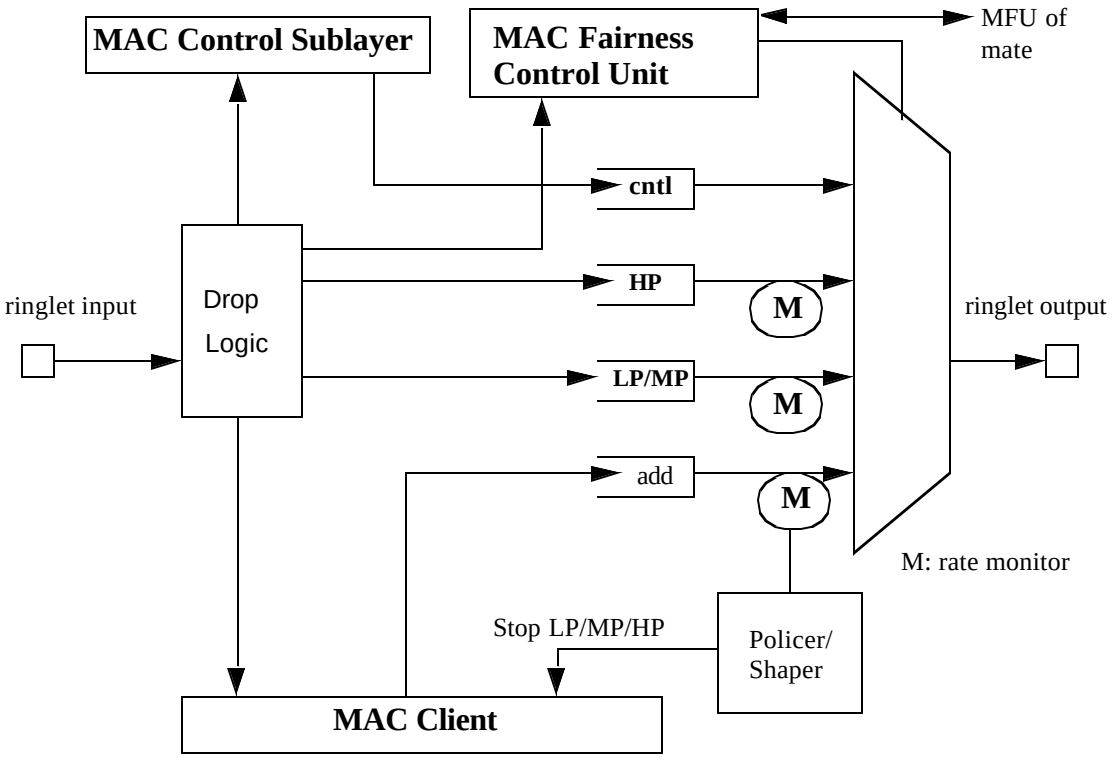


Figure 11—MAC Reference Model

5.1 Transmit and forwarding operation

A node 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 Access Rate (CAR.) Traffic within the CAR is treated as if it is high priority traffic while it is being accepted to the ring. Traffic above the CAR will be treated as low priority, and will be referred to as “excess” MP traffic, or eMP. The node will decide which traffic to send based on a priority scheme, which will differ between single and dual transit buffer implementations.

5.1.1 Single buffer implementation

In a single transit buffer implementation, forwarded 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 traf-

fic 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.

5.1.2 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 CAR is treated as high priority, and can be sent as long as the LPTB is not almost full.

Excess medium priority transmit traffic 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.

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 Figure12

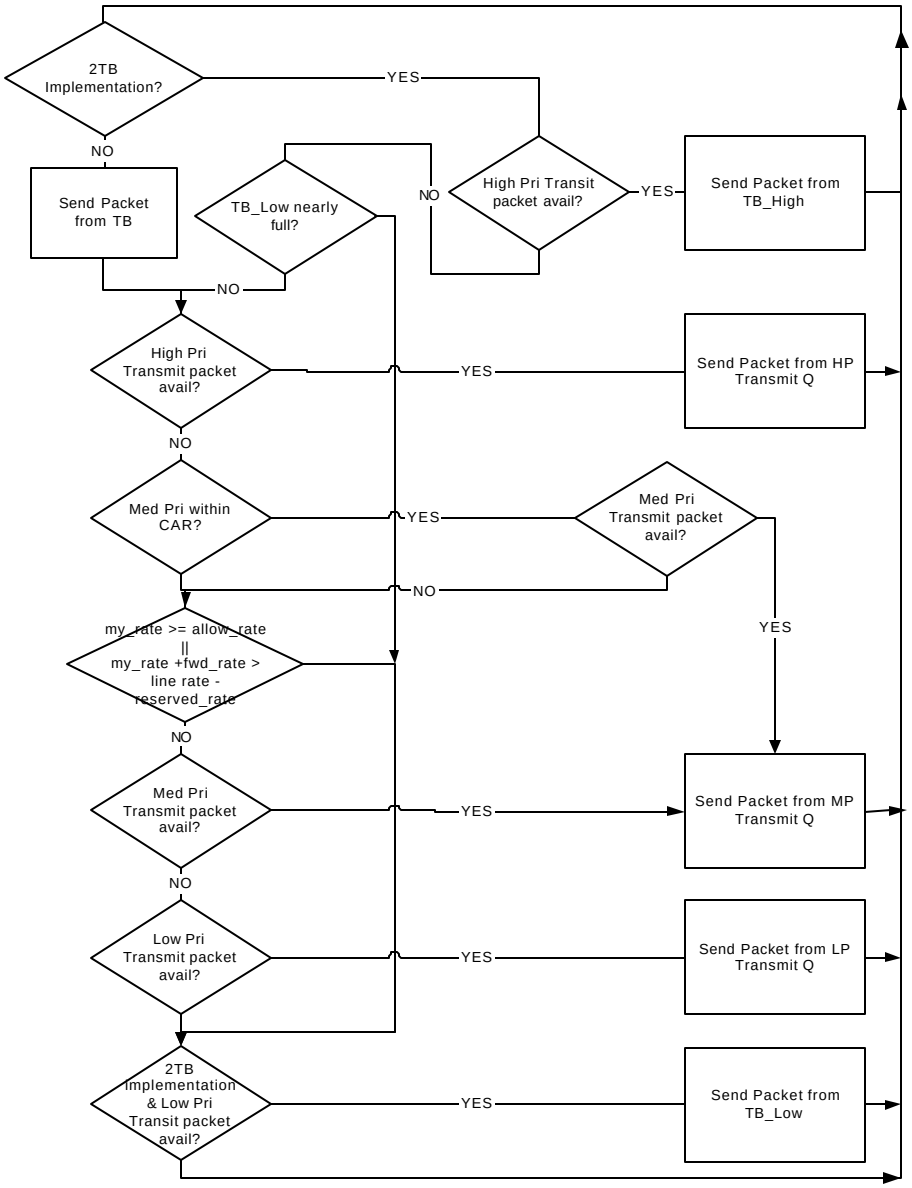


Figure 12—RPR transmit flowchart

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

5.3 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 control messages are broadcast to all nodes on the ring. All control packets (type=101) are hub by hub except the one with (control type = DA_strip). 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 Figure13.

1. Decrement TTL on receipt of a packet, discard if it gets to zero; do not forward.
2. Strip unicast packets at the destination station. Accept and strip “control” packets.
3. 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.
4. Do not process packets other than for TTL and forwarding if the type is not supported by the node (e.g. reserved types).
5. Transit packets will be discarded if there is a HEC error.
6. Packets accepted by the host due to destination address match may be 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.

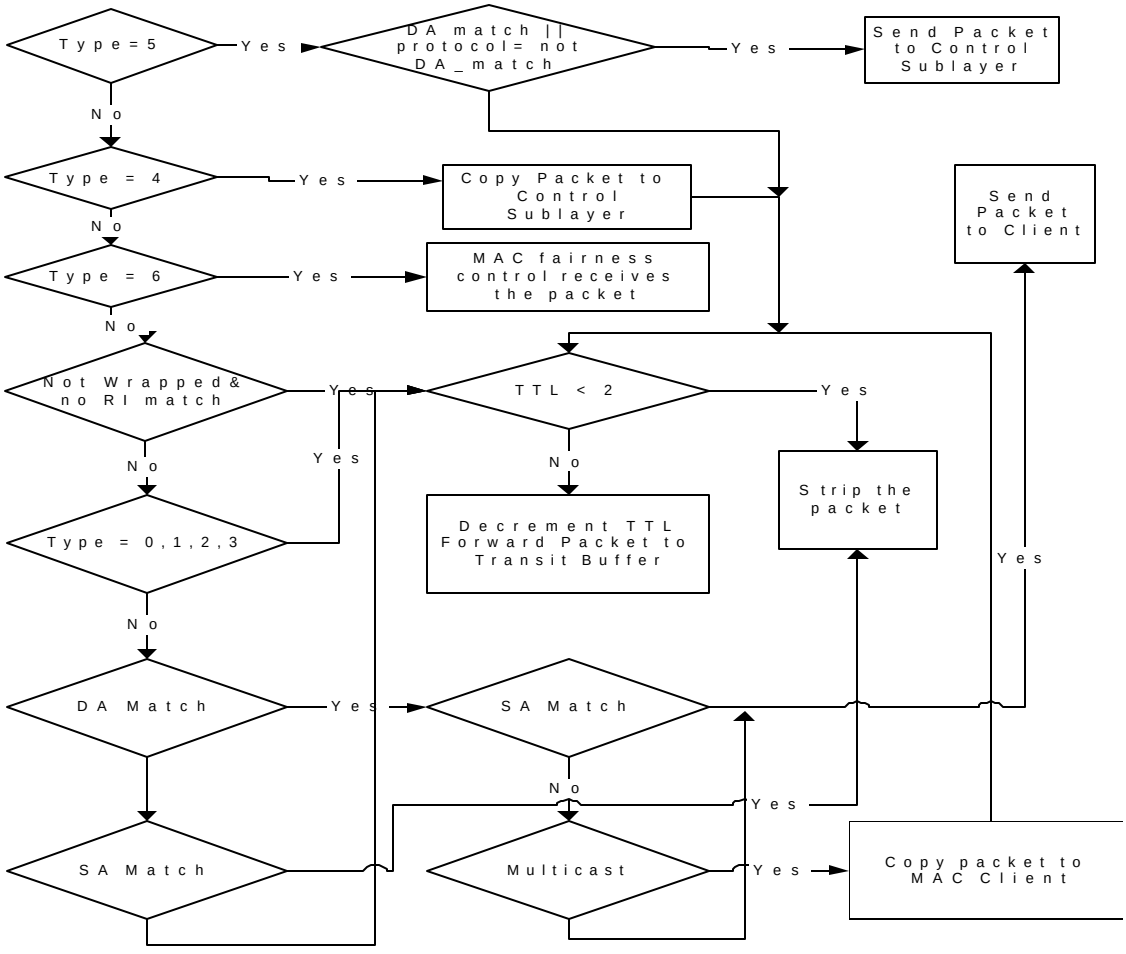


Figure 13 – RPR Receive Flowchart

8. Type 5 control messages are accepted and stripped if there is a DA match or protocol filed indicates that it is a point to point message
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.

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.

ii) Conditionals (if statements) in Figure13 branch to the right if true and branch down if false.

5.4 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 at least twice 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.

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.

5.5 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.

5.6 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 packets, and does not modify any of 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.

6. RPR fairness algorithms

The RPR-fa is a mechanism that enforces fairness among the nodes on the ring. It applies only to LP and excess Medium priority (eMP) traffic coming from the MAC client. Each node is assigned a weight, which allows the user to allocate more ring bandwidth to certain nodes.

The RPR-fa is implemented completely in the MAC Fairness control Unit (MFU). The MFU does not understand the ring topology, however it utilizes the implicit topology information (TTL value) passed by the MAC client to perform fairness and policing functions. 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.

In RPR-fa, if a node experiences congestion, it will advertise the value of its transmit fair rate counter (my_rate) to upstream nodes via the opposite ring. The fair rate counter is run through a low pass filter function and divided by the node's weight. The low-pass filter stabilizes the feedback, and the division by weight normalizes the transmitted value to a weight of 1.0. When they receive an advertised fair rate, upstream nodes will adjust their transmit rates so as not to exceed the advertised value (adjusted by their weights). Nodes also propagate the advertised value received to their immediate upstream neighbor. Nodes receiving advertised values who are also congested propagate the minimum of their normalized low pass filtered transmit fair rate and the received fair rate.

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.

Fairness messages are generated periodically and also act as keepalives informing the upstream station that a valid data link exists.

6.1 Congestion detection

Congestion is detected by using 3 criterion depending on the type of transit path design.

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

If the transit path has 2 buffers, then the node is congested when the depth of the low priority transit buffer reaches a congestion threshold. Or (my_rate + forward_rate) is more than (line_rate - reserved_rate).

6.2 Traffic policing function

RPR-fa utilizes the allow_rate_congestion and TTL_to_congestion registers and my_rate_congestion accumulator to police the add traffic to the ring. Allow_rate_congestion and TTL_to_congestion registers keep the values of fair rate and 255-TTL of the most recently received fairness packet, respectively. my_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 >255-TTL contribute to my_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, policing function can not be fooled.

6.3 Dynamic traffic shaping

RPR-fa controls ring access rate within allow_rate. Due to RPR-fa dynamic control nature, the allow_rate can vary in a wide range. When access traffic contains transmission bursts, ring access rate can vary between burst low rate and high rate as limited by allow_rate. As a result, traffic on the ring can be burst. To achieve better jitter performance, ring access rate can be dynamically shaped to a low-pass filtered value of allow_rate, which allows ring access rate conform to a fair rate as governed by RPR-fa while reduces extra burst 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 burst transmission rate. The token octets are generated at a rate equal to the low-pass filtered value of allow_rate. 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 medium 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. When the number of token reaches to the maximum token octets, token bucket shall saturate.

6.4 Pre-Provision bandwidth for high priority traffic

For some application, it will require to reserve certain bandwidth for high priority traffic. In RPR scheduling block, it will make sure that (my_rate+forward_rate) is less than (line_rate - reserved_rate) before low priority transmit/transit packet will be served. Also in the congestion detection, if (my_rate + forward_rate) is more than (line_rate - reserved_rate) then it is congested.

6.5 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 value received fair rate 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 6.3. 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.

6.6 Basic RPR-fa rules Of 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 CAR" 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 four counters/registers which control the traffic forwarded and sourced on the RPR ring. The counters are my_rate (tracks the amount of LP and eMP traffic sourced on the ring), my_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 forwarded on to the ring from the LP transit buffer), the registers are 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 policier shall not allow my_rate_congestion and my_rate to pass allow_rate_congestion and max_rate, respectively. It is accomplished by generating stop_low signal to the client when my_rate_congestion and my_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 my_rate value to upstream nodes. The value (nlp_my_rate) is obtained by passing my_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 is observed when the LP transit buffer depth crosses a threshold.

A node that receives a non-null fairness message (rcvd_rate) will set its allow_rate_congestion and TTL_to_congestion to the rcvd_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_rate is the same node that received it then the rcvd_rate shall be treated as a null value. When comparing the rcvd_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_rate generally propagate rcvd_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, then a null value is propagated to the upstream neighbor instead of the rcvd_rate.

Nodes that are congested propagate the smaller of normalized nlp_my_rate and rcvd_rate.

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

6.7 Multi-Choke implementation of RPR-fa

Multi-Choke implementation of RPR-fa uses the same algorithms for deciding fairness. The difference is that its client keeps track of up to the number of node on the ring congestion locations (choke points), and uses this information to increase ring utilization and spacial 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 my_rate values for all of the queues after the choke point are used as a total_my_rate value. This value is normalized and compared to the rcvd_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 section 6.3

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 rate for destination nodes 3, 4, 5,... against the rcvd rate for the choke point at 2. If that test passes, it will check the total rate for destination nodes 5, 6,... against the rcvd_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 rcvd_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

6.8 RPR-fa pseudo-code

A more precise definition of the fairness algorithm is shown below.

Variables:

```
typedef unsigned short Uint16;
typedef unsigned long Uint32;

Uint32 lo_tb_depth; //low priority transit buffer depth

Uint32 my_rate; //count of LP and eMP octets transmitted by client
Uint32 my_rate_congestion; //count of LP and eMP octets transmitted by
client and destined beyond the congestion point

Uint32 lp_my_rate; //my_rate run through a low pass filter
Uint32 nlp_my_rate; //lp_my_rate / WEIGHT
Uint32 lp_my_rate_congestion; //my_rate_congestion run through a low pass
filter
Uint32 nlp_my_rate_congestion; //lp_my_rate_congestion / WEIGHT


boolean my_rate_ok; // flag indicating that host is allowed to transmit

Uint32 allow_rate_congestion; // the fair amount each node is allowed to
transmit beyond the congestion point

Uint32 forward_rate; //count of LP+eMP octets forwarded from the LP tran-
sit buffer
Uint32 lp_forward_rate; // forward_rate run through low pass filter

Uint32 lp_allow; // allow_rate run through low pass filter
```

```
boolean congested; //node cannot transmit host traffic without the TB
buffer filling beyond its congestion threshold point.

Uint16 rcvd_rate; //the fair rate received from the downstream neighbor

Uint16 rev_rate; //the fair rate passed along to the upstream neighbor
Uint16 token_octets; // the number of octets in RPR-fa dynamic shaper

typedef struct _fairness_pkt_t {
   char[6] SA;
   Uint16 rate;
} fairness_pkt_t;

fairness_pkt_t fairness_pkt; //received fairness packet
```

Constants:

```
Uint16 WEIGHT; //configured weight for this node

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

Uint16 AGECOEFF = 4; // Aging coeff for my_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_ALLOW = 64; // LP filter for allow rate auto increment
Uint16 LP_ALLOW_COEF = 128; // low-pass filter for lp_allow

Uint16 NULL_RCVD_INFO = 65535; //All 1's in rcvd_rate field

Uint16 TB_LO_THRESHOLD; // TB depth at which no more LP host 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 reserved_rate; //high priority reserved rate
```

THESE ARE UPDATED EVERY CLOCK CYCLE:

```
// my_rateis increment by 1 for every LP/eMP octet that is
// transmitted by the host (does not include data
// transmitted from the Transit Buffer).
// my_rate_congestion is increment by 1 for every LP/eMP octet that is
// transmitted by the host (does not include data
// transmitted from the Transit Buffer) and destined beyond congestion
point (i.e., TTL > TTL_to_congestion).

// forward_rate is increment by 1 for every LP/eMP octet that enters the
// LP Transit Buffer

// token_octets is incriminated at a rate which equals to lp_allow/(AGE-
COEFF * DECAY_INTERVAL)
// token_octets is decremented by 1 for every LP/eMP octet that is trans-
mitted by the client.
```

```
if ((my_rate_congestion < allow_rate_congestion) &&
    (token_octets > 0) &&
   !((lo_tb_depth > 0) && (WEIGHT * forward_rate < my_rate)) &&
   (my_rate < MAX_ALLOWANCE)){
   my_rate_ok = TRUE; // true means OK to send client packets
}
```

UPDATED WHEN FAIRNESS_PKT IS RECEIVED:

```
if (fairness_pkt.SA == my_SA) &&
   ((node_state == wrapped)) {
   rcvd_rate = NULL_RCVD_INFO;
} else {
   rcvd_rate = fairness_pkt.rate;
}
```

THE FOLLOWING IS CALCULATED EVERY DECAY_INTERVAL:

```
if ((lo_tb_depth > TB_LO_THRESHOLD/2)||
   ((my_rate + fwd_rate) > (MAX_LRATE - reserved_rate))){
   congested = TRUE;
} else {
   congested = FALSE;
}
lp_my_rate = ((LP_MU-1) * lp_my_rate + my_rate) / LP_MU;
nlp_my_rate = lp_my_rate / WEIGHT;

//my_rate is decremented by min(allow_rate/AGECOEFF, my_rate/AGECOEFF)

lp_fwd_rate = ((LP_FWD-1) * lp_forward_rate + forward_rate) / LP_FWD;

fwd_rate is decremented by forward_rate/AGECOEFF

//(Note: lp values calculated prior to decrement of non-lp values).

if (rcvd_rate!= NULL_RCVD_INFO) {
   allow_rate_congestion = (rcvd_rate*WEIGHT);
} else {
   allow_rate_congestion += (MAX_LRATE -
allow_rate_congestion)/(LP_ALLOW);
}
lp_allow = ((LP_ALLOW_COEF-1)*lp_allow + allow_rate)/LP_ALLOW_COEF;

if (congested) {
   if (nlp_my_rate < rcvd_rate) {
      rev_rate = nlp_my_rate;
   } else {
      rev_rate = rcvd_rate;
   }
} else if ((rcvd_rate != NULL_RCVD_INFO) &&
      (lp_forward_rate > (allow_rate_congestion/WEIGHT)) {
   rev_rate = rcvd_rate;
} else {
   rev_rate = NULL_RCVD_INFO;
}

if (rev_rate > MAX_LRATE) {
   rev_rate = NULL_RCVD_INFO;
}
```

6.9 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. According to simulation results, for 100km rings 256KB is adequate. For 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.

6.10 RPR fairness packet format

RPR fairness packets are sent out periodically to propagate allowed rate information to upstream nodes in a unicast packet format. RPR fairness packets also perform a keepalive function. RPR fairness packets should be sent periodically. 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.

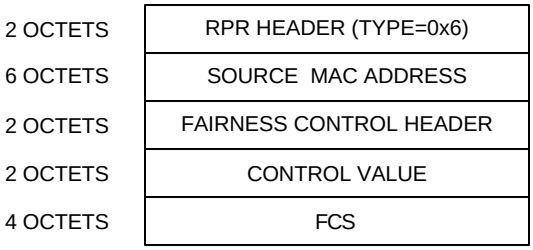
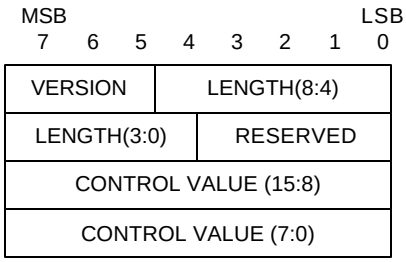


Figure 14—Fairness Packet Format

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



6.10.1 Version field (3 bits)

This field is to specify the version number of fairness packet. Table 6 shows the fairness message version values. Type 1 fairness message is used to implement basic RPR-fa. Type 2 fairness message is only needed

Table 6—Version values

Value	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

to support multi-choke implementation. Type 1 messages are propagated hub by hub 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 MFU and information contained is passed to the MAC clients whereas type 2 messages are not processed by MFU and passed to the MAC clients as well.

6.10.2 Length field (Optional 8 bits)

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

6.10.3 Reserved field (4 bits)

It is set to 0 for type 1 and 2 fairness packets.

6.10.4 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 FFFF indicates the availability of up to line rate bandwidth.

2 OCTETS	RPR HEADER(TYPE=0x5)
6 OCTETS	DESTINATION MAC ADDRESS
6 OCTETS	SOURCE MAC ADDRESS
2 OCTETS	PROTOCOL=0x2007
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
2 OCTETS	MAC ADDRESS
nn OCTETS	OTHER MAC BINDINGS
4 OCTETS	FCS

Figure 15—Topology Packet Format

7. 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).

7.1 Topology discovery packet format

7.1.1 Topology length

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

7.1.2 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.

7.1.3 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.

7.1.4 MAC type format

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

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	Steer Protection capable (1)
5-7	Fairness message version
8	Jumbo frame support(1)
9-15	Reserved

Table 7—MAC Type Format

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.

7.2 Topology discovery state transition

7.2.1 Constants

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

7.2.2 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

7.2.3 Timers

— Topology_Query_Timer

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

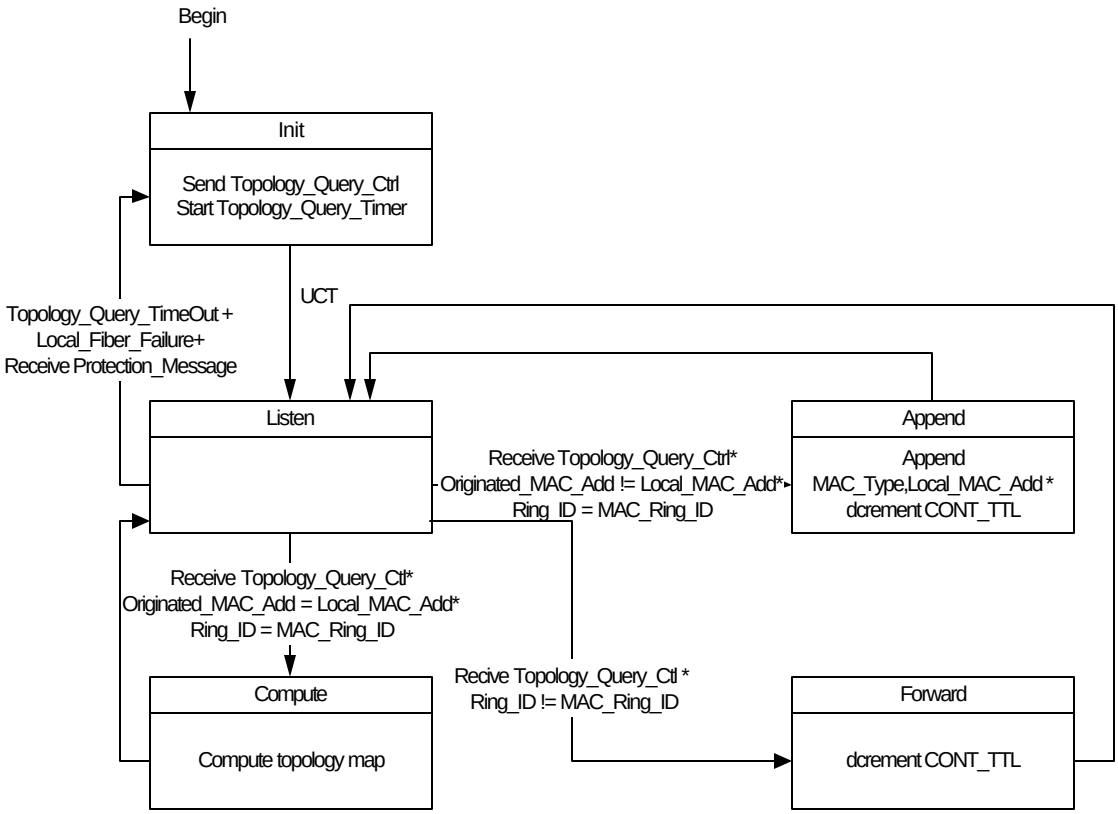


Figure 16—Topology discovery state transition diagram

8. Protection switching protocol description

Resiliency is one of RPR objectives. The goal is to provide protection within 50ms in case of ring or node failure. There are two known protection mechanisms: wrapping and steering. Protection switch protocol will support both mechanisms.

During the topology discovery, every node will indicate if it supports wrap protection or not. If all nodes are able to wrap protection, Protection switch protocol will use the wrap protection scheme. Otherwise, Protection switch protocol will use the steering protection scheme.

Nodes communicate between themselves using protection switch protocol signaling on both inner and outer ring.

8.1 Wrap protection

An 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.

An example of the data paths taken before and after a wrap are shown in Figure 17 and Figure18. Before the fiber cut, N4 sends to N1 via the path N4->N5->N6->N1.

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.

Subsequently a new ring topology is discovered and a new optimal path N4->N3->N2-N1 is used, as shown in Figure19. Note that the topology discovery and the subsequent optimal path selection are not part of the protection switch protocol.

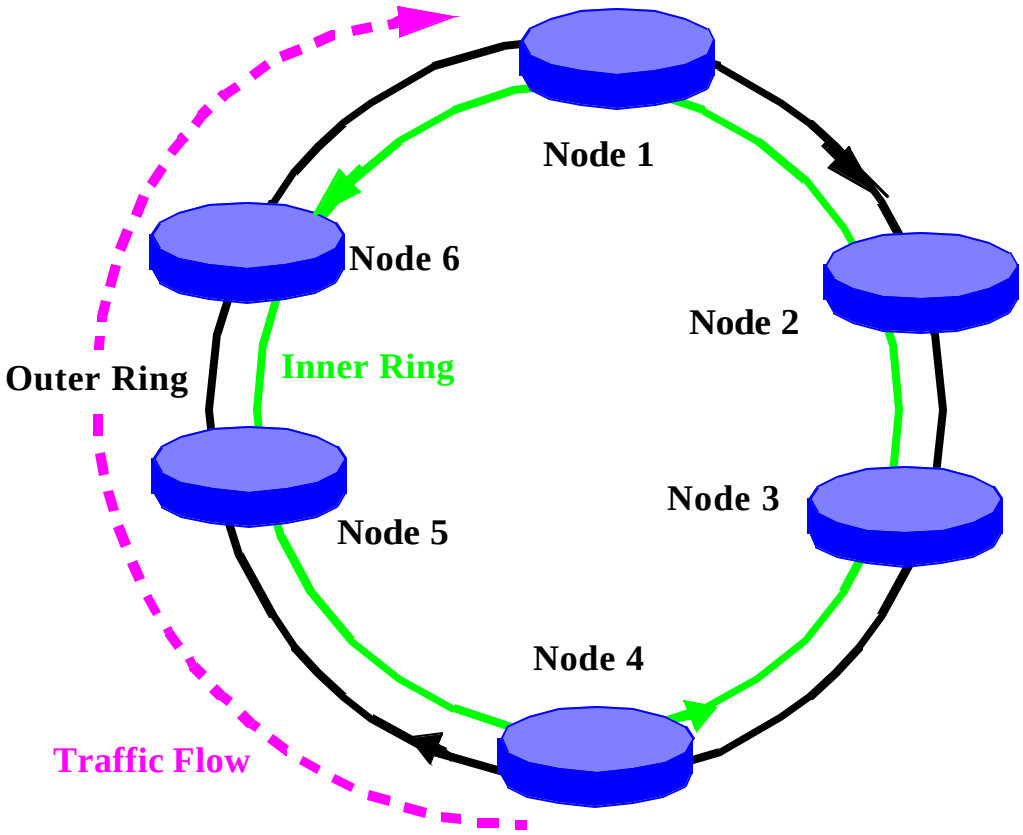


Figure 17—Data flow before fiber cut.

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.

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

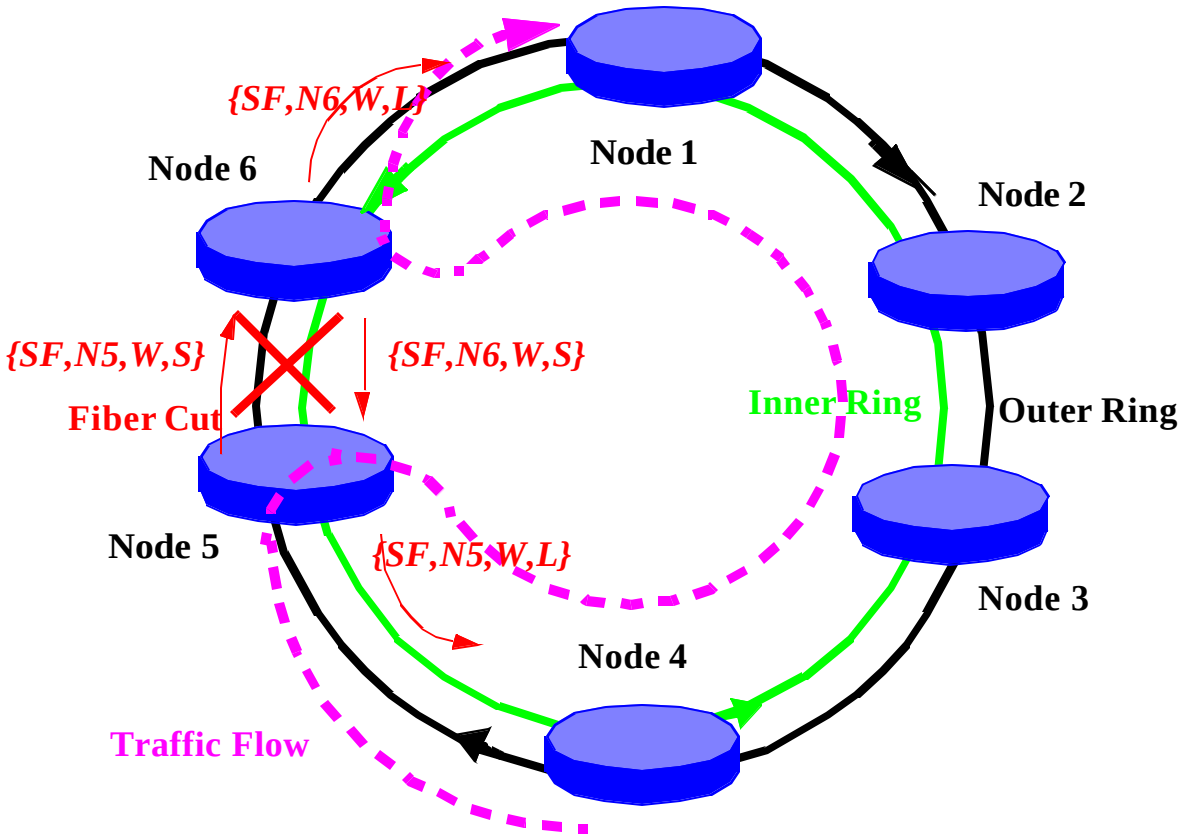


Figure 18—Data Path after Wrap

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

If a protection (either automatic or operator originated) is requested for a given span, the node on which the protection has been requested issues a protection request to the node on the other end of the span using both the short path (over the failed span, as the failure may be unidirectional) and the long path (around the ring).

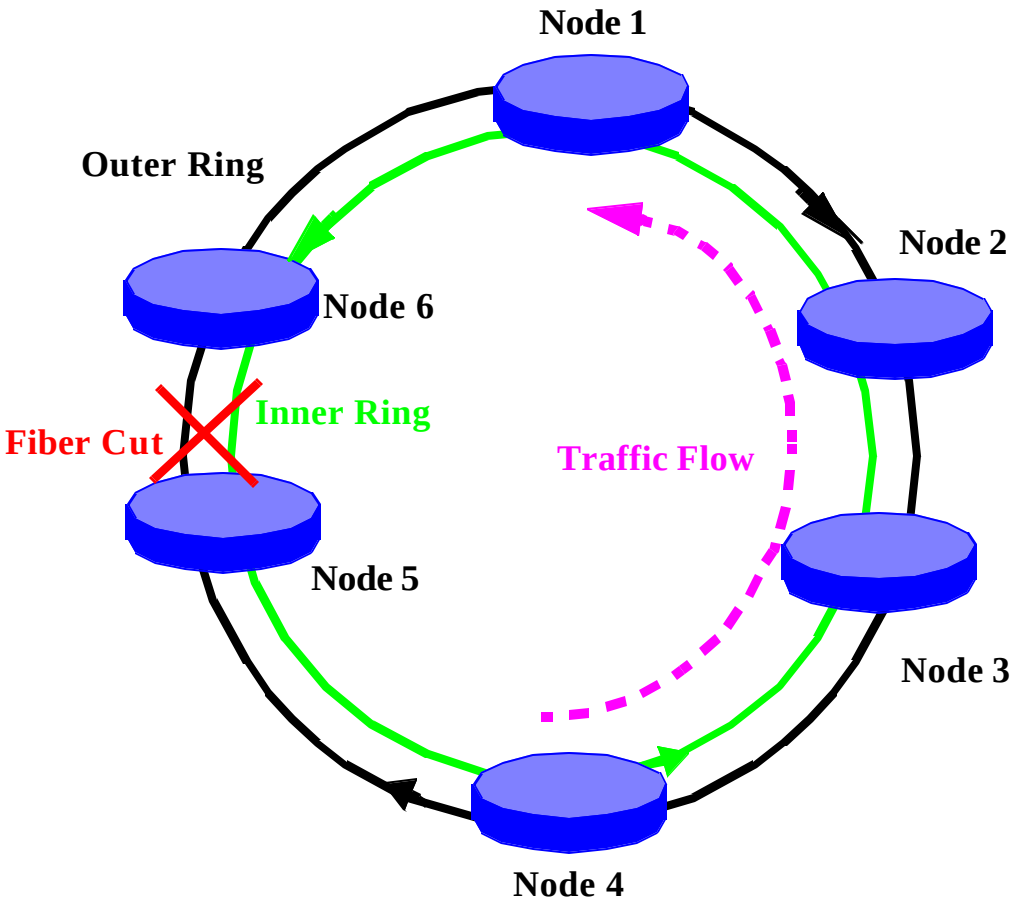


Figure 19—Data Path after new Topology Discovery

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 (wraps at both ends of a span for both transmit and receive directions, even if a failure is only unidirectional).

8.2 Steering protection

Steering protection will not wrap the failed span. An 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.

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.

8.3 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 Figure20 below.

2 OCTETS	RPR HEADER(TYPE=0x4)
6 OCTETS	DESTINATION MAC ADDRESS
6 OCTETS	SOURCE MAC ADDRESS
2 OCTETS	PROTOCOL=0x2007
2 OCTETS	HEADER CHECKSUM
1 OCTET	CONTROL VERSION(0x0)
1 OCTET	CONTROL TYPE(0x2)
2 OCTETS	CONTROL TTL
1 OCTET	Protection Message Octet
1 OCTET	Reserved
4 OCTETS	FCS

Figure 20—Protection Switch Packet Format

The protection switch specific fields are detailed below.

8.3.1 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.

8.3.2 Source MAC address

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

8.3.3 Protection message octet

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

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

Table 8—Protection message Octet Format

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

8.3.4 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 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.

8.3.5 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}

8.4 RPR protection protocol states

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

8.4.1 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}

8.4.2 Wrapped

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

8.5 Protection protocol rules

8.5.1 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.

8.5.2 RPR protection signaling and wrapping mechanism

R S.1:

Protection switch signaling is performed using protection control packets as defined in Figure20 “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.

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).

8.5.3 Example

In Figure21, 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.

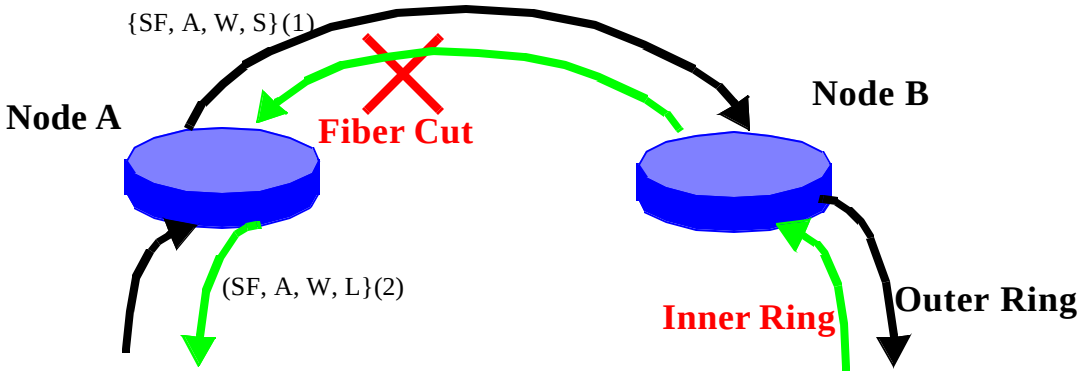


Figure 21—RPR Protection Switch Signaling

8.6 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.

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 wrapped node with a request >= than that in the request; otherwise it passes it through and unwraps.

R P.9:

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

R P.10:

When a wrapped 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:

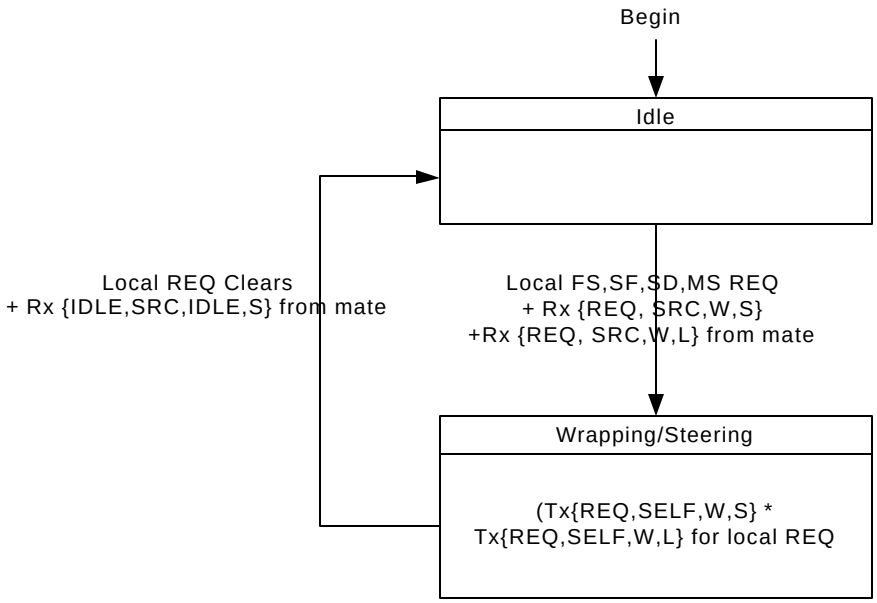
If a short path FS request is present on a given side and a SF/SD condition takes place on the same side, accept and process the SF/SD condition ignoring the FS. Without this rule a single ended wrap condition could take place. (Wrap on one end of a span only).

R.P.17:

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

8.7 Protection state transition

Figure22 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 22—Simplified Protection State Transitions Diagram

8.8 Failure examples

8.8.1 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.

8.8.1.1 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

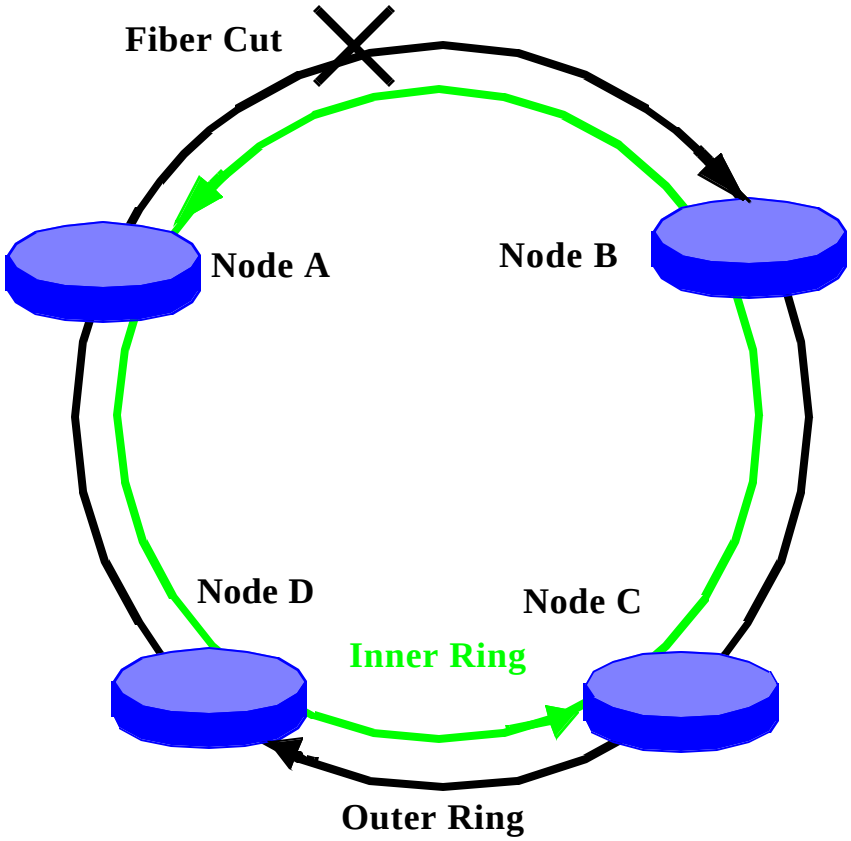


Figure 23—An RPR Ring with a fiber cut in outer ring

8.8.1.2 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

8.8.2 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.

8.8.2.1 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}

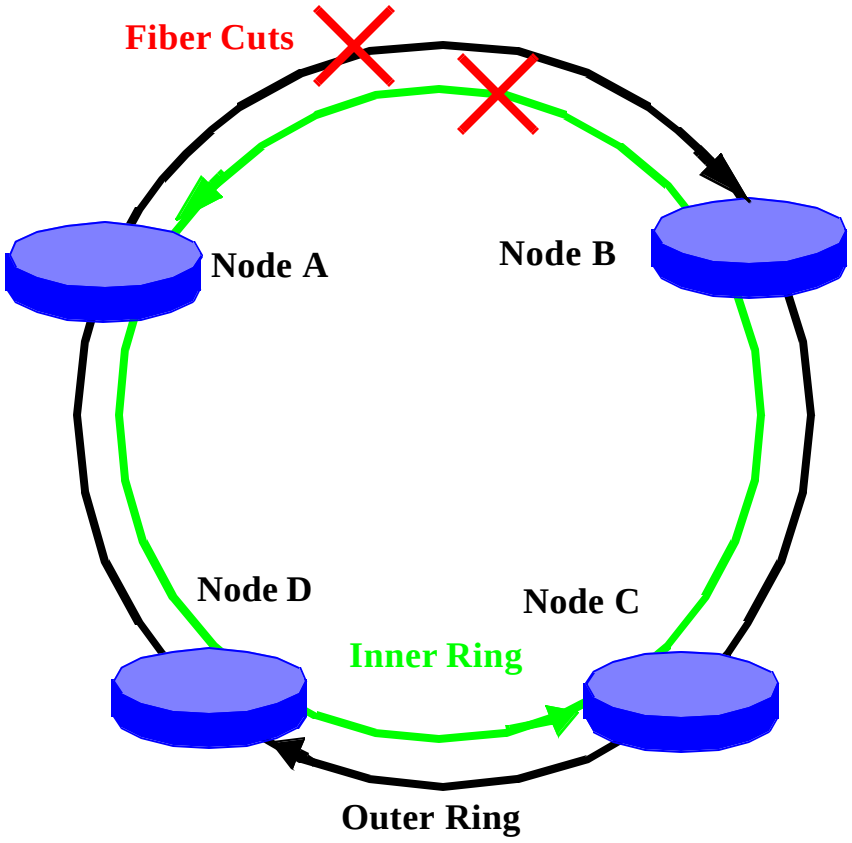


Figure 24—An RPR Ring with bidirectional fiber cut

3. Steady state is reached

8.8.2.2 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

8.8.3 Failed node scenario

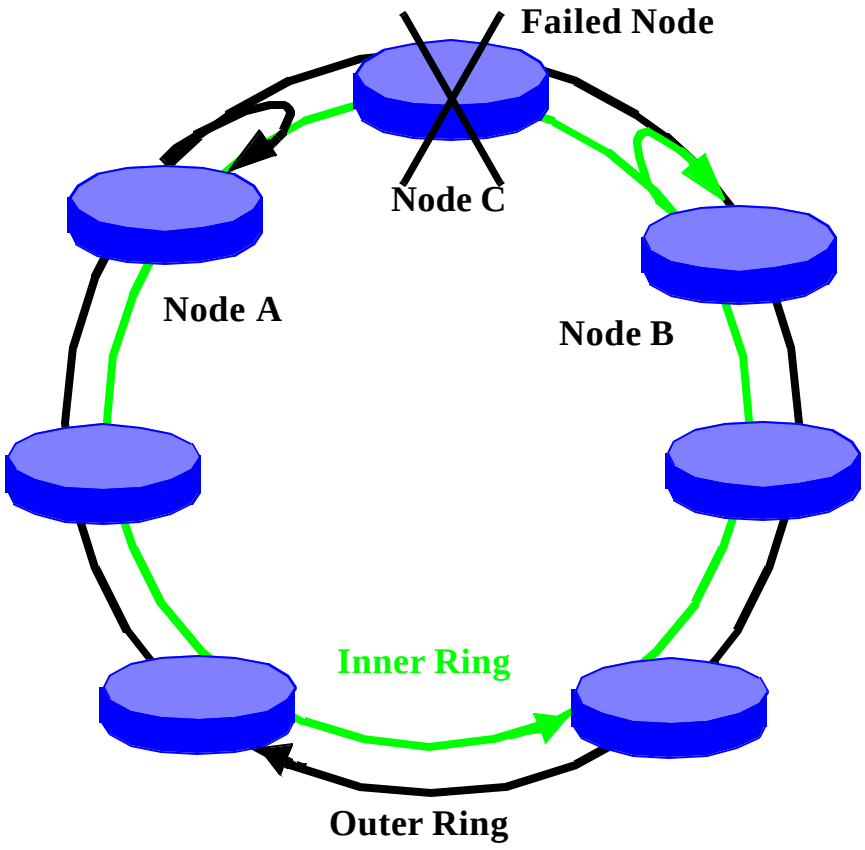


Figure 25—An RPR Ring with a failed node

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

8.8.3.1 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. 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
4. Steady state is reached

8.8.3.2 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 and enters a Pass Through state passing requests between C and B
5. Steady state is reached

8.8.3.3 Second span returns to service

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

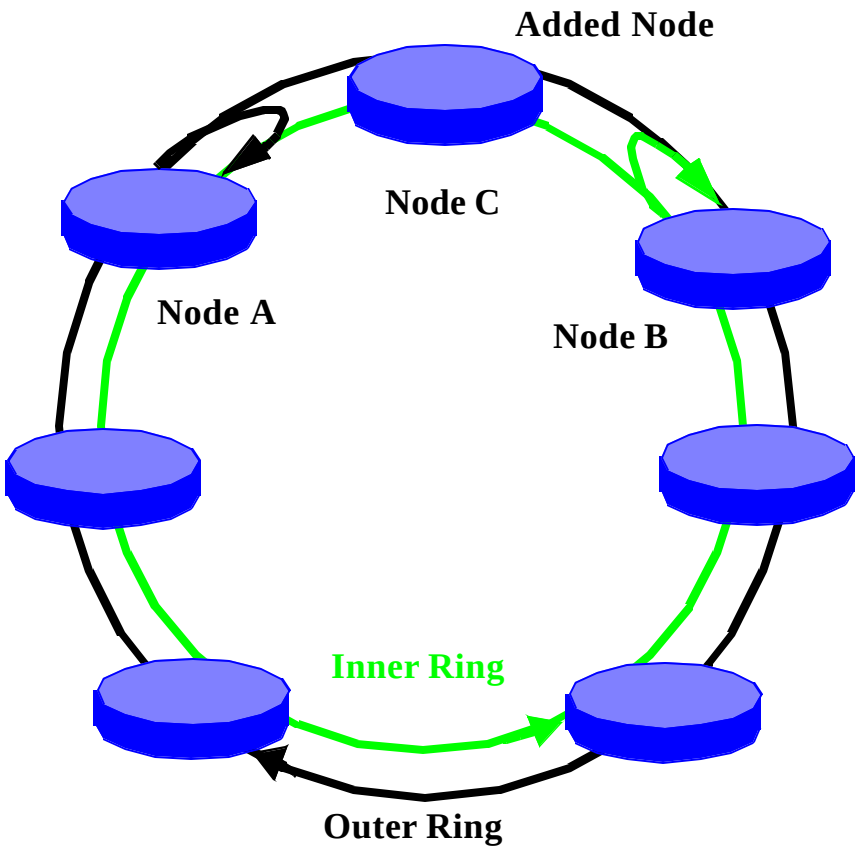


Figure 26—An RPR Ring with a failed node

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.

8.8.3.4 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

8.8.3.5 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.

8.8.3.6 Second span put into service

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

9. System Considerations

9.1 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.

Figure27 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 who 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.

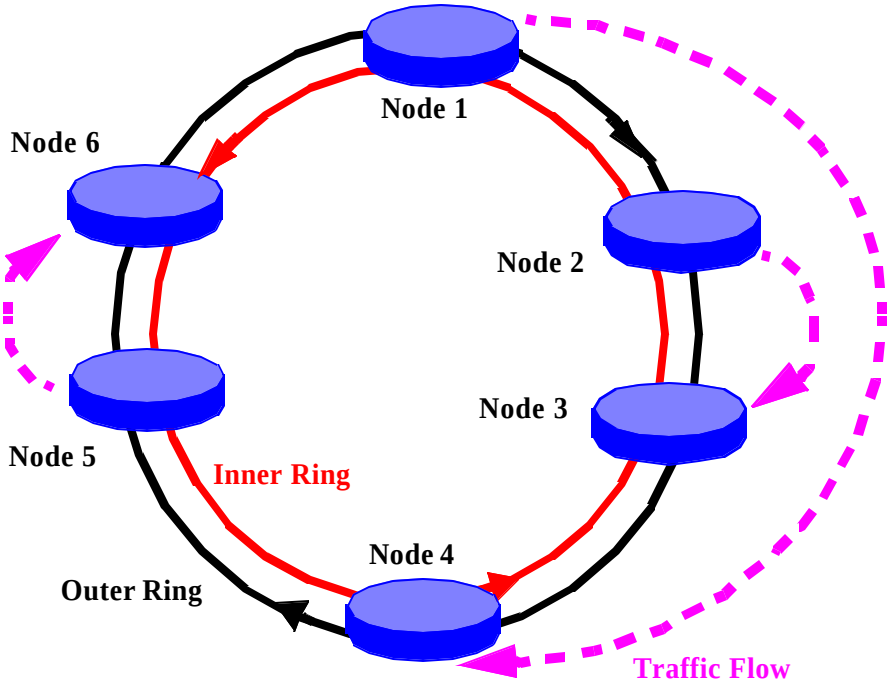


Figure 27—Global and Local Reuse

10. Physical media

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

10.1 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 100 msec in order to allow the SONET/SDH network to protect. Only if a failure persists for over 100 msec (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.

10.1.1 POS framing

Flag delimiting on SONET/SDH uses the octet stuffing method defined for POS. 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 POS packet delimiting allows RPR to be used directly over fiber or through an optical network (including WDM equipment).

10.1.2 GFP framing

RPR frame will be able to encapsulate using the GFP frame format

10.2 Ethernet

RPR frames can be sent over Ethernet physical media.

10.3 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.

11. OAM

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 physical media lacks OAM functions completely.

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. In some cases the flows may be further segmented using the Class Of Service, and the optional subscriber identification tag.

OAM frames are grouped in four groups:

— Fault Management
— Performance Management
— Activation/Deactivation

11.1 Fault Management

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

11.2 Activation/Deactivation

Note: The need for this flow is to be discussed

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.

11.3 OAM functions of the RPR layer

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 and PM

11.3.1 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).

Among others, Loss Of Continuity (LOC) is one defect 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 mis-configuration/failure. Such failure/mis-configuration 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.

Fault Management frames include: RDI, LB and CC

11.3.1.1 RDI defect indication

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.

11.3.1.2 Continuity Check

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 using Activation/Deactivation frames or by configuration. 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 shall have separate CC capability, LOC declaration shall be per ringlet. It shall be possible to activate CC without specifying the side, in this case the Station should activate it in the shortest path, and steer the CC flow in case of failure.

CC shall be always bidirectional.

11.3.1.3 Loopback capability

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 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 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.

11.3.2 Activation/Deactivation procedures

Since Continuity Check can be activated at any time, an initialization procedure is needed between the two endpoints of the flow to properly initialize the OAM process. 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.

11.4 OAM frame handling during failures

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

11.4.1 Steer protection

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

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

If the CC was activated using the unspecified side option, then the CC flow will be steered and no LOC will be declared.

11.4.2 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.

11.5 OAM frame

The OAM frame includes a common part and a function specific part. Figure28 shows the general frame format.

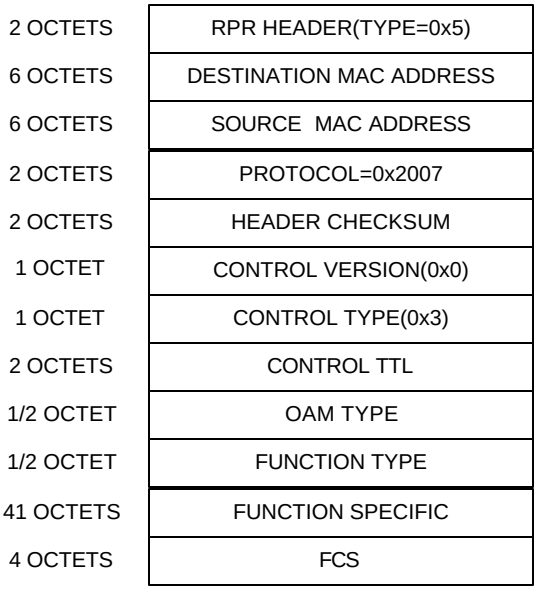


Figure 28—OAM frame format

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

11.5.1 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.

11.5.2 OAM Type

The OAM type identifies the OAM group of the OAM frame. Table9shows the possible values of the OAM type field.

11.5.3 Function Type

This field indicates the actual function performed by this frame within the group indicated by the OAM Type. Table9 shows the possible values of the Function Type field.

—

Table 9—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

11.6 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 to determine the type of OAM frame received
— Silently discard OAM frames with unsupported type or Function

11.6.1 OAM frames support

All RPR compliant Stations should support Loopback OAM frames.

CC and RDI support is optional. When implemented, support must be of both functions.

Activation/Deactivation support is optional.

11.7 Specific fields for OAM frames

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

11.7.1 Fault Management frame

The Function Type field for the Fault Management frame will be used to identify the following possible functions: RDI, CC and LB.

11.7.1.1 RDI Fault Management frame

The function specific fields for RDI fault management frames are illustrated in Figure 29

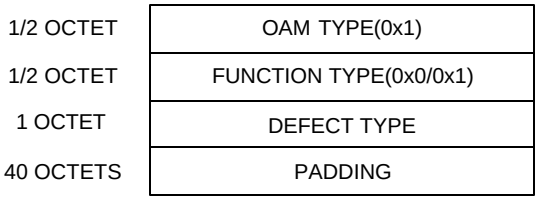


Figure 29—RDI fault management frame

11.7.1.1.1 Defect type

Optional field used to provide further information about the nature of the failure. Examples of this information are specified in Table10

Table 10—Defect Types

Defect Type	Coding
Defect not specified	11111111
Defect in the RPR layer. For example loss of continuity	00000000

11.7.1.1.2 Defect Location

This field identifies the ringlet in which the LOC was detected. Examples of this information are specified in Table11

Table 11—Defect Location

Defect Location	Coding
Ringlet 0	00000000
Ringlet 1	00000001
Unspecified	11111111

11.7.1.2 Continuity Check Fault Management frame

No fields are specified for the Continuity Check Fault Management frame

11.7.1.3 Loopback Frame

The function specific fields for Loopback frames are illustrated in Figure30

Figure 30—Loopback frame

1/2 OCTET	OAM TYPE(0x1)
1/2 OCTET	FUNCTION TYPE(0x0/0x1)
1 OCTET	REQUEST TYPE
1 OCTET	IDENTIFIER
1 OCTET	SEQUENCE NUMBER
38 OCTETS	PADDING

11.7.1.3.1 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. Table12 shows the possible values of the Loopback type and the required action.

Table 12—Loopback Request Type values

Request Type	Action
0x00	Reply through shortest path
0x01	Reply through Inner ring
0x02	Reply through Outer ring
0x03	Reply on same ring
0x04	Reply on opposite ring

11.7.1.3.2 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.

11.8 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 31

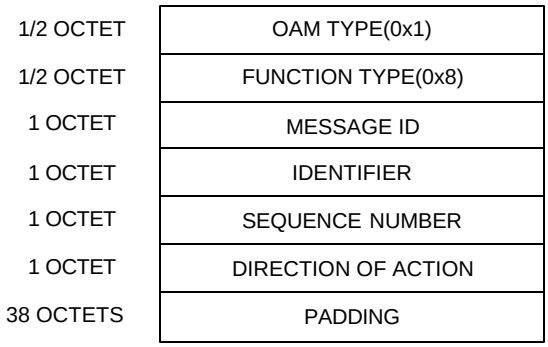


Figure 31—Activation/Deactivation frame

11.8.1 Message ID

This field indicates the Message ID for activating or deactivating specific OAM functions. Code values for this field are shown in Figure 13.

Table 13— 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

11.8.1.1 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.

11.8.2 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 Table14

Table 14—Direction values

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