TABLE OF CONTENTS

1. Media Access Control

1.1 Congestion avoidance

For a shared ring topology in which the 802.17 MAC is used, each ring segment carries both local client traffic and the traffic from the clients of other MACs upstream. Unless the upstream MACs control their access rates, their traffic may consume the entire ring segment bandwidth, creating congestion and hence blocking the local client from gaining access to the media.

The 802.17 MAC employs congestion avoidance mechanism to prevent congestion before it occurs. The congestion avoidance is a proactive mechanism. It does not wait until congestion takes place and then reacts. The mechanism constantly monitors and controls the access rate from each upstream MAC as well as from its local client.

The congestion avoidance is a two phases mechanism: bandwidth allocation &the access rate control. Bandwidth allocation is carried out to ensure fair access to the ring segment and more importantly to allow maximum utilization of the ring segment. The MAC attached to the ring segment allocates the bandwidth between its local client and other competing clients of upstream MACs according to the allocation method defined in 1.6.2.1. The access rate must be controlled by each MAC on the ringlet to regulate the traffic that the MAC allows on to the ring. The access rate control prevent upstream MACs from gaining access more than their allocated rate, creating congestion at a down stream ring segment. The access rate control shall be implemented in compliance with 1.6.3.1.

1.2 Virtual Output Queuing support and Head-of-Line (HoL) blocking prevention

The client of an 802.17 MAC may send traffic to multiple destinations traversing many ring segments. If the MAC does not allow an independent access rate per destination, it is possible that the MAC sets the access rate low to satisfy the bandwidth allocated by one congested destination and severely limits the access rates to other uncongested destinations.

Another common problem that the 802.17 MAC prevents is head of line (HoL) blocking problem, which normally occurs in a single FIFO access. If the MAC client implements a single FIFO to buffer frames awaiting access on to the ringlet, it is possible that a frame destined to uncongested destination may be waiting behind an HoL frame that cannot gain access on to the ring because its destinations is congested. Until the HoL frame is removed from the FIFO all frames behind are blocked.

A well-known solution to this HoL blocking problem is a virtual output queue (VOQ) implementation. The client can implement VOQ by maintaining a dedicated FIFO for each destination. With a per-destination buffer, one frame cannot be blocked by another frame for a different destination, hence eliminating HoL blocking completely.

In order to allow the client to maximize a spatial reuse property of the ring, the 802.17 MAC shall implement independent access rate control for each ring segment. The 802.17 MAC shall also support client VOQ implementation.

To support VOQ implementation, the 802.17 MAC shall make the RCF values of the entire ring segments available for its client.

1.3 Bandwidth Allocation Policy

The 802.17 MAC shall support the following requirements:

a) Committed bandwidth provisioning.
b) Weighted fair allocation of available bandwidth of each ring segment.
c) Maximum utilization of each ring segment.

The bandwidth allocation policy shall be followed:

d) Determine the amount of committed bandwidth currently utilized by each MAC on the ring.
e) Calculate the amount of committed and uncommitted bandwidth available, namely available bandwidth.
f) Allocate available bandwidth fairly according to a pre-configured weight associated with each traffic-originating MAC.

As a result, each MAC receives total allocated bandwidth as follows:

Total allocated bandwidth = committed bandwidth + weighted fair allocation of available bandwidth

Detailed allocation method is specified in 1.6.2.1.

1.4 Functional Model of the MAC method

1.4.1 Overview

The MAC sublayer defines a medium-independent facility, built on the medium-dependent physical facility provided by the Physical Layer, and under the access-layer-independent LAN LLC sublayer (or other MAC client). It is applicable to a shared media resilient packet ring (RPR) network.

The LLC sublayer and the MAC sublayer together are intended to have the same function as that described in the OSI model for the Data Link Layer alone. In an RPR network, the notion of a data link between two network entities does not correspond directly to a distinct physical connection. Nevertheless, the partitioning of functions presented in this standard requires three main functions generally associated with a data link control procedure to be performed in the MAC sublayer. They are as follows:

g) Data encapsulation (transmit and receive)
    i) Framing (frame boundary delimitation, frame synchronization)
    ii) Addressing (handling of source and destination addresses)
    iii) Error detection (detection of physical medium transmission errors)

h) Frame handling
    i) Header processing
    ii) Frame add-drop
    iii) Transit path

i) Media Access Management
    i) Media access rate control (congestion avoidance)
    ii) Link bandwidth allocation (congestion avoidance)
    iii) Frame insertion (contention resolution)
    iv) Congestion management

j) Protection switching

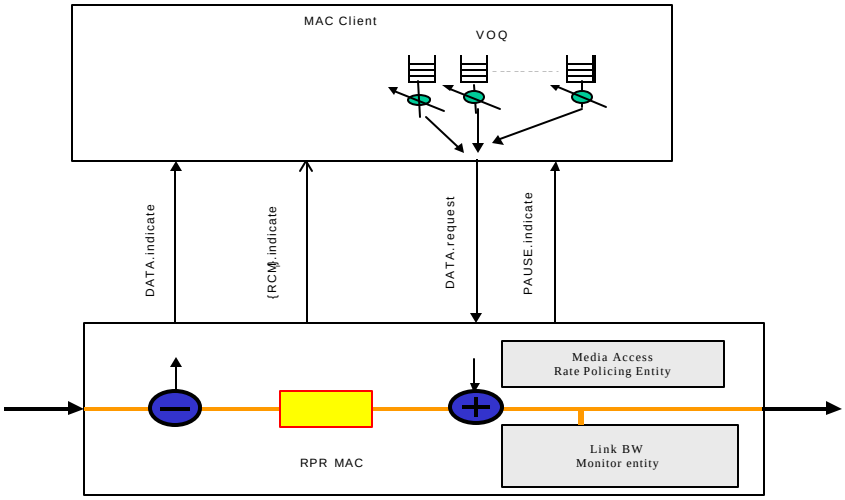


Figure 1.1. MAC model

Depicted above, the 802.17 MAC shall not implement frame buffering itself, rather its client maintains buffer for frames waiting to gain access to the ring. The MAC, however, may implement a staging buffer/s to store frames accepted temporarily while it waits for an opportunity to place them on the ring. Once accepted by the MAC, the frame is said to be already on the ring for a media accessing purpose.

1.4.2 MAC Architecture

The internal architecture of the MAC is shown in Figure 1.2. The MAC comprises of five entities. These entities can be categorized by their functionality into two categories. On a transit path, they are the header processor and the frame insertion arbiter entities. On the bandwidth management portion, they are the link bandwidth monitor, the fairness message management and the media access rate control entities.

The header process entity examines the header of an 802.17 frame to determine whether the frame is to be received for its local client and/or to be forward to the next MAC on the ring as a transit frame. The header process entity also determines the source station and the length of the frame that are required by the link bandwidth monitor entity. Transit frames are place in a small transit buffer waiting for the pending frame insertion to complete.

The frame insertion arbiter entity resolves contention among frames from the transit buffer, the insert buffer and the RCM frame from the fairness message management entity attempting to access the outgoing ring segment at the same time. The arbitration method shall be as specified in 1.5.2.4.

For the bandwidth management function, the link bandwidth monitor entity tracks the access rates going through the outgoing ring segment of all stations including of itself. Based on its measurement, the entity allocates available bandwidth in a weighted fair manner. The entity generates a rate control message periodically to update current bandwidth allocation.

The fairness message management entity is responsible for sending and receiving RCM frames to and from the other stations. This entity also responsible to pass the RCMs to the MAC client in case that the client may implement VOQ. In side the MAC, the RCMs is used by the media access rate control entity to limit the access rate to the media so that traffic from its client shall never exceed allocated bandwidth on any ring

segment. The PAUSE.indicate is used to signal the client that the MAC currently does not accept any frame from the client.

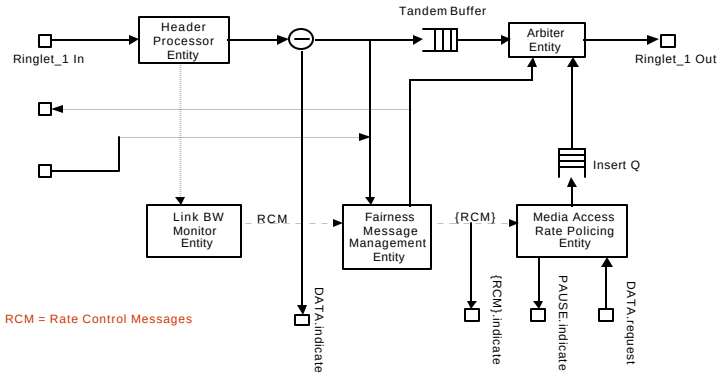


Figure 1.2. MAC Architecture

1.5 Frame Handling and Transit Path Functional Model

1.5.1 Objectives and requirements for the transit path

1. The transit path is part of the shared medium
2. The transit path is lossless.
3. The transit path implements destination and source stripping
4. The transit path implements broadcasting and multicasting: drop and pass mode
5. Minimal buffering in the transit path (Transit buffer only for collision avoidance) in order to
   - Minimize the cost of the standard RPR MAC chip saving memory cost
   - Minimize delay in the transit path
   - Maximize scalability as RPR MAC chip scales at higher-speed and multiple rings
6. Allow for different topologies in the RPR ring
   - Multiple rings

1.5.2 Basic design and mode of operation

Figure 1 shows on the transit path for each transit path consists of the header processing entity, the tandem buffer and the arbiter entity. The transit path is considered as an extension of a media, and hence is part of the ring. For multi-vendor interoperability, the transit path characteristic that all implementation must be conformed to the following:

The transit path includes the three following functionalities shown in Figure 1:
1. The RPR MAC frame header processing
2. The transit buffering
3. The add/pass arbitration

The RPR MAC frame header-processing block determines reception and transit conditions. The role of the transit buffer is exclusively to avoid collisions between Add and Pass frames. The Add/Pass Scheduler arbitrates access to the outgoing link between Add and Pass frames.

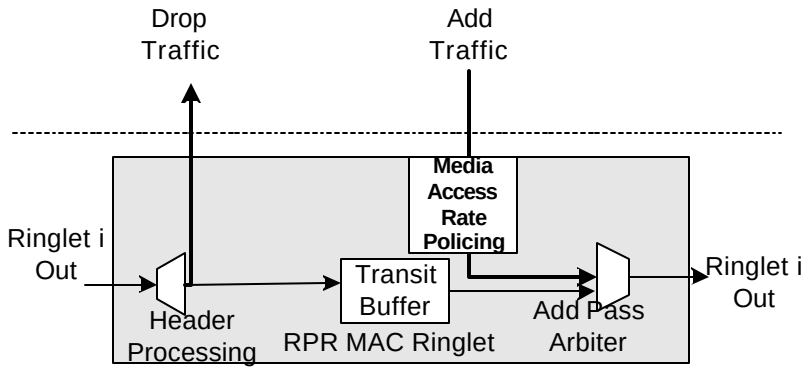


Figure 1: Transit Path Design

1.5.2.1 RPR MAC reception rules

When a frame arrives on an ingress port of an RPR MAC, the destination MAC address is matched with the RPR database in the header-processing block shown in Figure 1. The decision to strip or bypass the frame is accomplished as follows:

- If the frame DA matches the set of MAC addresses assigned to given instance of the MAC, the frame is stripped from the ring. The
- If the frame DA is a broadcast, multicast
    - If TTL>1, the frame is both stripped and copied.
    - If TTL=1, the frame is stripped.
- If the frame SA matches this RPR MAC address, then the frame is stripped, and discarded.
- If the frame has a bad CRC on the RPR MAC header, the frame is stripped and discarded. A bad CRC counter is incremented. These errors are accounted for signal degradation on the upstream link.
- If the DA MAC address of the incoming frame does not match the set of MAC addresses assigned to this instance of the MAC and if the TTL field in the header of the frame is TTL<=1 the frame is stripped and discarded.

Else, the frame is bypassed.

- The TTL field in the RPR MAC header is decremented by one.

The earliest time that the RPR MAC can decide whether to strip or bypass the frame is upon reception of the entire RPR header (which should be CRC protected). It does not wait for the entire frame to be received.

When the frame is stripped data.indicate signal is given to the MAC client to allow it to fetch the frame from the MAC.

Promiscuous Mode

RPR MAC allows all (or low priority class only) the transit traffic to be received to the MAC client in the promiscuous mode of operation. Promiscuous mode of operation is necessary for debugging operations. In addition, this mode allows MAC client to arbitrate and re-queue the transit traffic with the add traffic before transmitting back into the MAC. This mode of operation allows MAC client to provides fairness, class based services, and spatial reuse through vendor specific queuing and scheduling. Annex C shows how MAC client queuing allows for local congestion management by buffering traffic in case of congestion. Congestion management becomes then a local problem solved by the congested node and therefore it eliminates the need for generating the congestion or throttling messages to upstream nodes on the ring. On the other hand this mode allows congestion messages to throttle the media access insert rate.

In the promiscuous mode MAC Bandwidth management primitives are as follows: (perhaps Adisak you can accommodate it in the BW management section I'll move it to the BW management part after the group review it)

- MACs operating in promiscuous mode will generate RCM messages that contain the RCF value equals the outgoing segment link rate to upstream nodes on the ring. RCM messages will advertise the entire outgoing link segment rate to all other nodes.

- Media Access Rate control only polices traffic originating from its local client.

The Media access rate control entity polices the insert (add+transit) traffic from the MAC client on per destination segment basis on advertised RCF.

In a ring made of some nodes operating in promiscuous mode, while others in a normal mode. The MAC operating in normal mode will advertise the RCM message for fair sharing of attached outgoing link segment. The receiving MAC in the promiscuous mode will multiply the RCF factor with local weight to police the fair add rate for that link segment.

1.5.2.2 RPR MAC Transit Rules

Bypass frames are sent to the transit buffer in Figure 1. Frames are scheduled from the transit buffer by the arbiter that schedules between transit (or bypass) frames and transmit (or add) frames. The arbiter is a strict priority scheduler where the transit frames have highest priority.

The scheduling algorithm works as follows:

Step 1: choose a frame to be transmitted

If a transit buffer has at least one byte

Choose a frame from the transit buffer

Else if an insert buffer has at least one frame

choose a frame from the insert buffer

Step 2: transmit the chosen frame with no pre-emption

Step 3: complete the transmission, repeat step 1

In words, this works as following:

- If there is a transit frame waiting, it is sent out right away to the outgoing link on a given ringlet.
- If there is no transit frame waiting, an add frame is sent out. Thus the add frame waits until the transit path has no frame to send. Since there is a rate control at the insertion of each node, the amount of bandwidth in the transit path is limited and allows fair access for the add traffic.
- Add frames cannot interrupt the transit frame for the transmission (No pre-emption).
- Transit frame cannot interrupt the add frames under transmission.
- In the store and forward mode of operation transit frames are received entirely before they are sent out.

1.5.2.3 RPR MAC Discard Rules

The RPR MAC discard rules are as following:

- The FCS for the header, HEC, is incorrect the frame is discarded.
- If the source MAC address matches the RPR MAC database in Header Processing block, the frame is discarded.
- TTL value expired, the frame is discarded.

1.5.2.4 RPR MAC Add Rules

Add frames are queued in the MAC client before they access the RPR MAC please see the Annex A to consider different add path options. When MAC client has frame to send to the MAC client, MAC client indicates this by data.request primitive.

RPR MAC accepts the frame in the add path. If following conditions are satisfied.

- If there is no transit packet under transmission and transit buffer is empty.
- Media access rate control has not asserted PAUSE.

Media Access Rate Control

MAC client is expected to shape the frames according to the advertised rate by the RPR MAC for each destination. Bandwidth management described in section x. described how the media access bandwidth is computed on destination basis. The PAUSE shall be asserted to prevent the MAC client from exceeding the allocated bandwidth on any segment whenever the MAC client insert rate exceeds the allocated bandwidth on some segment downstream. Media access rate policing method is described in section x.

Virtual Output Queuing support and Head-of-Line (HoL) blocking prevention

One of the objectives of RPR is to maximize the spatial reuse in the RPR ring, e.g. to maximize the link utilization for frame flows with arbitrary (source, destination) pairs. The client of an 802.17 MAC may send traffic to multiple destinations (more and 1) traversing many ring segments.

If the MAC does not allow an independent access rate per destination, it is possible that the MAC sets the access rate low to satisfy the bandwidth allocated by one congested destination and severely limits the access rates to other uncongested destinations.

Another common problem that the 802.17 MAC prevents is head of line (HoL) blocking problem, which normally occurs in a single FIFO access. If the MAC client implements a single FIFO to buffer frames awaiting access on to the ringlet, it is possible that a frame destined to uncongested destination may be waiting behind an HoL frame that cannot gain access on to the ring because its destinations is congested. Until the HoL frame is removed from the FIFO all frames behind are blocked.

The proposed 802.17 architecture solves this problem through Virtual Output Queuing (VoQ), where there is an output queue for each destination node. A well-known solution to this HoL blocking problem is a virtual output queue (VOQ) implementation. The client can implement VOQ by maintaining a dedicated queue for each destination and a separate queue for unknown destination traffic. With a per-destination + unknown destination buffer, a frame cannot be blocked by another frame for a different destination, hence eliminating HoL blocking completely.

In order to allow the client to maximize a spatial reuse property of the ring, the 802.17 MAC shall implement independent access rate control for each ring segment. The 802.17 MAC shall also support client VOQ implementation. To enable VoQ, congestion avoidance messages are sent rate control for each link to each station in the ringlet. The 802.17 MAC shall have the rate control values of the entire ring segments available for its client.

Signaling between the RPR MAC and the RPR MAC client is described in detail in the MAC Service Definition section.

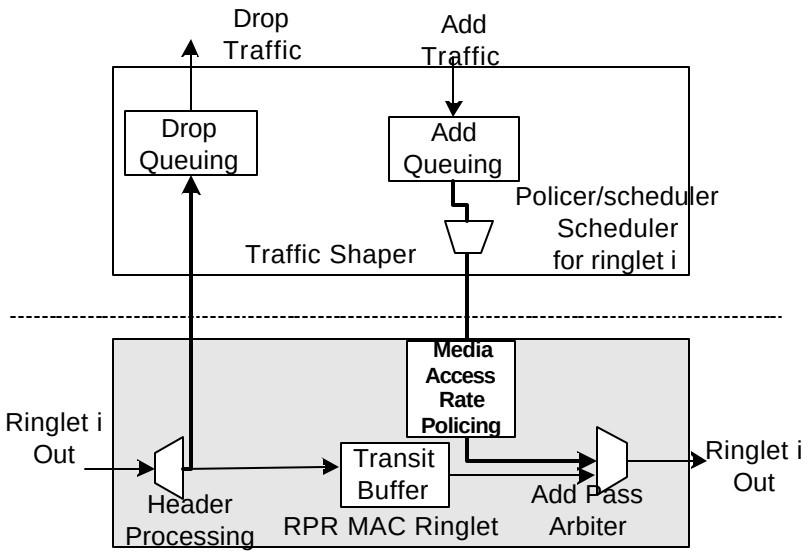


Figure 2: Transit path design and traffic shaper architecture

1.5.3 Transit Buffer Modes of Operation

1.5.3.1 Cut-through Mode:

In this mode of operation, frame transmission can begin before it is entirely received. At the minimum RPR header should be received before beginning transmission to the outgoing ringlet, since the header has to be

processed following the rules described in section 1.5.2.1. If the RPR MAC is sending out an add frame, while receiving a transit frame, the latter will be stored until the add frame has been completely sent out. Only Single MTU worth transit buffer required in the transit path.

The advantage of this option is that it reduces the delay that frames experience in the transit path.

1.5.3.2 Store and Forward Mode:

The store and forward mode of operation transit frames are received entirely before they are considered for transmission. This mode of operation allows FCS errored frames to be stripped and transit error counter incremented. This mode of operation requires, two MTU worth transit buffer in the transit path. The advantage of this option is that it eliminates degraded frames already in transit, so that they do not take up the bandwidth up to destination in the rare instance of FCS error.

1.6 Bandwidth Management

This clause provides detail requirements for the bandwidth Management entity for fair media access.

The goal of the fairness algorithm is to achieve local fairness on an RPR during congestion and to maximize the spatial reuse of the shared medium for an RPR. The algorithms achieve high throughput with minimal and bounded delay and jitter for all classes and BW requirements. It enables the VOQ construct that eliminates Head of Line (HoL) blocking.

In the un-congested state each station advertises ring bandwidth less the reserved bandwidth to all upstream stations. When congestion is detected, due to link utilization crossing its threshold or excess access delay. Rate control Factors (RCF) or fair rates per class are then advertised to the upstream stations the allowed rate to send on the output ring segment attached. Each station responds by limiting types of packets injected into the ring. This allows the congested link bandwidth to be divided fairly to all stations sharing it. Link utilization is continuously monitored for congestion alleviation. As BW resources are freed up, monitored by a decrease in link utilization the advertised RCF is increased until the maximum access rate is again advertised to the upstream stations to maximize all link utilization on the ring.

The key the fairness algorithm is a effective method of BW allocation estimation. If the BW estimation is too slow, it does not optimally allocating resources as congestion commences and terminates. The response lacks like an over damped system.

If the BW estimation is too aggressive then the fairness algorithm will falsely declare congestion thus reduces overall throughput. The response is too reactive like an under-damped system.

The recommended operating point for the fairness algorithm is critically damped error on the side of being over-damped for stability.

The Fairness procedure requires all stations to perform the following tasks as specified to prevent any unfair advantage:

1. A method of detection start and termination of congestion
2. A method of calculating the fair rate
3. A method of distributing and control the available resource fairly or unfairly

The BW management entity consist of the following blocks as shown in figure 1.2:

1. Link BW Allocation

    a. Monitors the link utilization per active source per class for the purpose of detecting the onset of congestion.

    b. Calculated the weighted fair rate to be advertised to other stations.

Monitors spare capacity in the media and reallocate the resources to stations that need it.

2. Fairness Message Management

Performs the tasks of calculating the “fair rate” and process the control messages, Rate Control Messages (RCM) for the purpose of executing fairness

This entities performs the following tasks:

    a. Send the message to other stations on the ring

    b. Receive Fairness Messages from other stations and calculate the rate control parameter for the Media Access Rate Control entity

    c. A method of receiving weight information from other stations.

3. Media Access Rate Policing

Media Access Rate Control entity performs the following function

    a. Police the ring access packets

    b. Performs the execution of fair access on to the ring as dictated by the fairness algorithm.

    c. Signal the MAC client sub-layer with RCF information for VOQ support

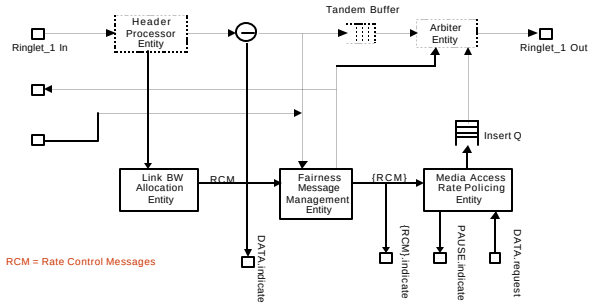


Figure 1.3 MAC architecture.

1.6.2 Link Bandwidth Allocation Entity

The link bandwidth allocation entity monitors traffic utilizing the attached outgoing segment for both passthru and ring ingress. Traffic from each source MAC is tracked independently to determine the activity and rate at which the source MAC puts traffic through the segment. The bandwidth allocation method relies on this information to allocate bandwidth fairly based on a given weighted allocation method in 1.6.2.1.

The bandwidth allocation procedure allows other MACs on the same ringlet to utilize the segment bandwidth fully under all traffic patterns. This must be achieved without violating committed bandwidth allocation and weighted fair allocation bandwidth of available bandwidth of the other MACs.

1.6.2.1 Weighted Fair Bandwidth Allocation Method

The clients of other MACs may or may not utilize all the bandwidth allocated. They may even not have enough traffic to utilize their committed bandwidth. In order to maintain maximum utilization of each ring segment at all time, this unused bandwidth must be re-allocated to other stations that need more bandwidth. Hence, it is the task of the link bandwidth allocation entity to detect the demand of each upstream MAC and to allocate available bandwidth in a prescribed manner.

To achieve this ring access fairness the following procedures are required to determine the fair rate:

1. Measure BW per observed clients within a ring

When determining the fair rate any reserved BW must be subtracted from available ring BW.

The available ring BW may be provision at the provision ringBW management object, less the reserved BW parameter for each station on the ring:

If this sum of committed bandwidth is represented by

$$\sum_{active} r_i$$

and C is the capacity of the ring segment, then

$$C - \sum_{active} r_i$$

reflects the amount of instantaneous available bandwidth to be allocated. If the committed BW is not to be shared, then difference is the available BW to be partitioned.

2. Determine the Advertised rate using by the prescribed equation. The method is as follow:

The weighted allocation of this available bandwidth can be given only each active station as follows:

$$weighted\_fair\_allocation\_of\_station\_i = \frac{w_i(C - \sum_{active} r_i)}{\sum_{active} w_i},$$

where w_i is a weight given to station *I*, and $\sum_{active} w_i$ is the total weights of active stations combined. This is the total of number of active stations requesting using the resource.

In order to reduce the amount of the rate update that each MAC sending out to other MACs, the above term can be further compacted. Only the RCF is needed to be sent to all other stations.

$$weighted\_fair\_allocation\_of\_station\_i = w_i * RCF ,$$

where RCF is defines as

$$RCF = \frac{(C - \sum_{active} r_i)}{\sum_{active} w_i}$$

The utility of RCF is specified in 1.7

3. management objective for sample rate is calcINTERVAL, weightOLD, weightNEW are used to determine the Rate Control Factor (RCF).

The RCF value is a filter value. This value is obtained by either filter the individual sample rate as specified by in clause 1.6.2.2 or it is determined by individual sampled rate ri and filter the *weighted_fair allocation_of station*_i.

The RCF filtering equation is described in clause 1.6.2.2.

4. Derive the rate for the Media Access Rate Control entity.

.

Each source station determines the access rate at which it can send traffic through this ring segment as follows:

$$f_i = r_i + w_i * RCF ,$$

where f_i denotes the access rate and r_i is a committed rate for station i.

It is worth noting that stations do not wish to participate in available bandwidth allocation can set their weights to zero. In this case, the stations can always access the media at but not above their committed rates.

The following state diagram describes the method of calculating RCF for one ring segment.

To be converted into a state diagram.

@packet arrival

bucket(source node) += packet length

@calc_interval (1us)

for each source node (I=1 to 256)

drain bucket(i)

bucket(I) -= calc_interval(Ri+Wi*RCF)*

if (bucket(I) <0), bucket(I) = 0

if (bucket(I) >0)

SUM Ri += Ri

SUM Wi += Wi

end FOR

RCF = (link capacity – SUM Ri) / SUM Wi

SUM Ri = 0

SUM Wi = 0

Figure 1.4 A state diagram of the weighted fair bandwidth allocation method

The 802.17 MAC shall implement a link bandwidth allocation entity and calculate an instantaneous RCF as specified in this clause.

1.6.2.1.2 Determining the local client activity

The attached client is said to be active if the PAUSE.indicate is asserted or the corresponding credit usage for the segment directly attach to the MAC (local segment) is less than an ActiveTHRESHOLD. As specified in 1.6.3.1, Having a credit usage of less than the ActiveTHRESHOLD meaning that the client is inactive for the moment.

1.6.2.2 RCF Filtering Method

The measured RCF is an instantaneous value that may fluctuate greatly depending on traffic dynamic and the fact that the buckets may transition from an active state to inactive state frequently capturing the percentage utilization of allocated bandwidth for each source station. A filtering (average) method is require to translate the instantaneous RCF into a more stable and smooth RCF that can be sent to other station on the ring. The following defined a first order averaging method:

$$RCF_{new} = weightOLD * RCF_{old} + weightCURRENT * RCF_{measured},$$

For a sampling period of 1 usec, it is recommended that the value of *WeightOLD* should be set at 0.95 and *weightCURRENT* at 0.05.

1.6.3 Media Access Rate Control Entity

The 802.17 MAC shall implement a media access rate control entity to limit client traffic according to a method described in 1.6.3.1. The PAUSE.indicate shall be asserted to prevent the MAC client from exceed-

ing the allocated bandwidth on any segment whenever the MAC client insert rate exceeds the allocated bandwidth on some segment downstream.

1.6.3.1 Policer Method

The policer method in figure 16 is a credit based rate pacer. It monitors the insert rate from the MAC client to all ring segments. For each the segment, the media access rate control entity maintains a credit counter.

At every T_{update} (usec) interval, credits are added to each counter based on the bandwidth allocation received for the segment, RCF. The number of credits loaded to the counter divided by the T_{update} is equal to the allowed fair rate for that segment.

To support VOQ, for each frame received from the client, credits for all segments that the frame needs to traverse, including the local segment attached to the MAC, are deducted based on the size of the packet.

To interwork with simple local fairness control method where only the congested station sends RCF, all counter segments from the head to the tail is loaded with the same RCF value. The simple local fairness implementation reduces the over throughput with reduced complexity.

The transmission is frame based, hence the pacer operates like a deficit token bucket. In the current T_{update} a station may transmit more than it fair rate, but by keeping a deficit the average rate is the fair rate.

When a RCM is expected but not received due to timer expiration, this is the classical Byzantine failure scenario, where a station is failed but does not cease operation but performs improperly.

To be converted to an SDL state diagram

```
At each T_update intervcal
  for each link segment
    calculate the node (for this MAC) allowed BW, fj.
      fj = rj + wj*RCF
    give credit for each segment
      if ( segment_credit) < 15,000,000
        segment_credit += fj
      if ( segment_credit) < 0   // client BW exceeds limit
        assert PAUSE.indicate
  end FOR

attach DATA.request
   if no PAUSE.indicate asserted, accept DATA.request
       for each segment between this and the dest nodes
        deduct segment credit
          segment_credit -= frame_length*10,000
```

end FOR

Note all BW are in Kbyte/sec unit

Figure 1.5 A state diagram of the policer method

1.6.4 Fairness Message Management Entity

The Fairness Message Management entity is responsible for

1. receiving fairness messages and validated them
2. Determine the RCF as shown clause 1.6.2.1
3. formatting transmit message with Rate Control Factors as received from the

1.7 Rate Control Message (RCM)

The rate control messages are send used to send target rates to each station to indicate the allowed rates

Message sending period depends on the tracking of the traffic pattern

1.7.1 Message Protocol

At each transmission time interval, a message is sent to the station upstream. For dual counter rotation ring one message is sent in one direction and another message is sent in the opposite direction.

For the purpose of ring efficiency and independence with the topology entity this message can be sent as a broadcast message to all station on the ring and source stripped or it is send unicast and forwarded by each station to its upstream station.

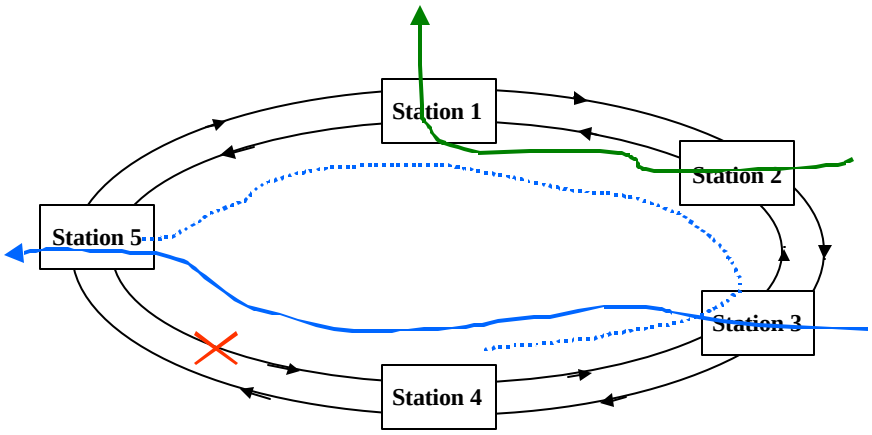
The decision to forward the message is dependent on the passthru BW utilization factor. If the passthru BW usage is less than a programmable threshold then its contribution to the congestion is negligible then the message is not required to be sent to the upstream station for efficient spatial re-use.

1.7.1.1 Single Link Failure: A special case

For the two cases of message propagation, in the case of single link failure, the Ring Control Message are handled as follow:

For source broadcast in both direction, the short path message will not reach its destination. The long path message is already launched hence no action is required from any station.

In the case of unicast message, the station detecting the failure will signal the source station and it will then send the Rate Control Messages in both the long path and short path.



Congestion Span 1, short path

Congestion Span 1, long path

Congestion Span 2, short path

Link Failure

1.7.1.2 Multiple failure Scenario

All other failure scenario may lead to the decision of isolated stations. This leads to segmented ring.

1.7.2 RCM Format

A type field in the RPR Header is reserved for the RCM messages

RPR Header	
Length	2 bytes
Station ID	6 bytes
Sequence Number	4 bytes
control	4 bytes
RCF Ring 1	4 bytes
RCF Ring 2	4 bytes
RCF ….	
RCF Ring N	4 bytes
CRC-32	4 bytes

Figure 1.6. RCM frame format

1.7.2.2 Field Definition

1) RPR Header is as defined in the Header Frame Format

2) Length: the bit fields are TBD : 16 bits

   i. The length field is used to identify the length of the RCM message

3) Node ID: This is the source station ID: 48 bits

4) Sequence number: The sequence number is used to identify the latest message. It allow the synchronization of all station automatically to the latest RCM: 32 bits

   32 bits is a big enough field that guards against Byzantine failure. The RCM is periodically transmitted thus when a node receives a message is starts an auto aging process. When the aging timer expires, this tells the station that any RCM received contains new information than what is currently in the database.

5) Control bits are used for various ring failure signaling: 32 bits

   i. 16 bits: Provision weight of station: wi

      Each station advertises its weight to all other stations

   ii. 8 bits: TTL used in unicast packet mode to determine hop count to RCM source. This value is incremented if the fairness message is propagated.

   iii. 2 bits: The version field is used to identify the message type

   iv. 1bits: BW allocation/ BW reservation methods

v. 1 bit: Forward/not forward.

In simplex mode, this bit indicates whether the RCF message is to be forward upstream based on the passthru threshold setting.

vi. 1 bit: Downstream rx failure

This bit indicates whether the immediate neighbor has a far end receive error

vii. 1 bits: Fairness Span failure

For single segment failure, the fairness message is looped back to the in the long path. The source can monitor this packet for source redirect. When the failure condition is cleared, this message is no long send.

This message is sent for the destination station. It can also to striped at the source and a source redirect message is generated.

viii. 2bits reserved.

6) Ring Control Factors (RCF). For each field this is the allowed rate for each ringlet. There are as many RCF fields as there is ringlet in a ring.

Each RCF factor is 32 bits

All "0" indicates that there is no RCF factor.

All "1" indicates that there is link is not active.

All other values indicates valid RCF in Kbytes

7) CRC-32: RCM packet integrity verification.

1.7.3 When generated

RCM messages are sent under the following conditions:

1. RCM messages are periodically transmitted as a Transmit Timer Expires. This is a soft state operation that provides highly robust operation. For ring operating on BW allocation this message is transmitted at very short intervals.

2. RCM is also transmitted on up congestion threshold crossing.

    1) Onset of Congestion is declared when the BW utilization on the ring crossed its threshold. This is equivalent to the activation of the stations exceeding the ring capacity.

    2) Congestion is also declared if the Ring Access Delay Timers expires.

Annex1: MAC client Add Drop Options

1.7.4 Add/Drop path design

The normal mode of operation avoids congestion and provides flows fair access through inter-node signaling. Inter-node signaling blocks or rate limits traffic in one or more add buffers at the source node. For a single queue add buffer shown in Figure 3, there is a leaky bucket shaper that limits the rate and the maximum burst size of add traffic inserted onto the ring to a single MTU sized packet.

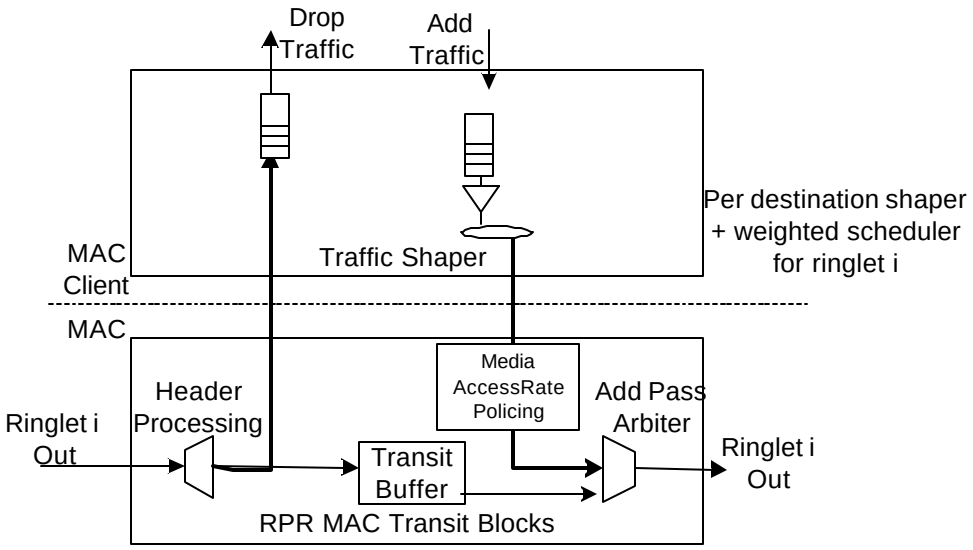


Figure 3: Single Output Queuing in the Add Path in the traffic shaper

1.7.4.1 Virtual Output Queuing

One of the objectives of RPR is to maximize the spatial reuse in the RPR ring, e.g. to maximize the link utilization for packet flows with arbitrary (source, destination) pairs. When a single node is sending traffic to two or more nodes, head of line blocking occurs if packets bound for multiple destinations are queued in a single add queue that has been blocked for at least one of the destinations due to ring signaling messages.

Head of line blocking can be explained very simply. In the ring shown in **Error! Reference source not found.**, Node 0 has two aggregate flows to send to Node 2 and Node 5. This is a common scenario for many applications such as Transparent LAN Services and peer-to-peer routing.

If the link between the Node 4 and Node 5 is congested, Node 4 will send the appropriate congestion avoidance/rate shaper parameter information to Node 0. This will slow or block the packets destined for Node 5. When packets destined for Node 5 reach the head of the queue in Figure 3, the packets destined for Node 2 will be slowed or blocked by the packet(s) destined for Node 5. This is head of line blocking. This problem has been long addressed in high-speed crossbar switches. The proposed 802.17 architecture solves this problem through Virtual Output Queuing (VoQ), where there is an output queue for each destination node shown in Figure 4. To enable VoQ, signaling propagates accurate congestion information for each link. Rate shaping is performed on a per destination basis at each source, based on the bandwidth availability on all of the links that must be traversed to reach each destination.

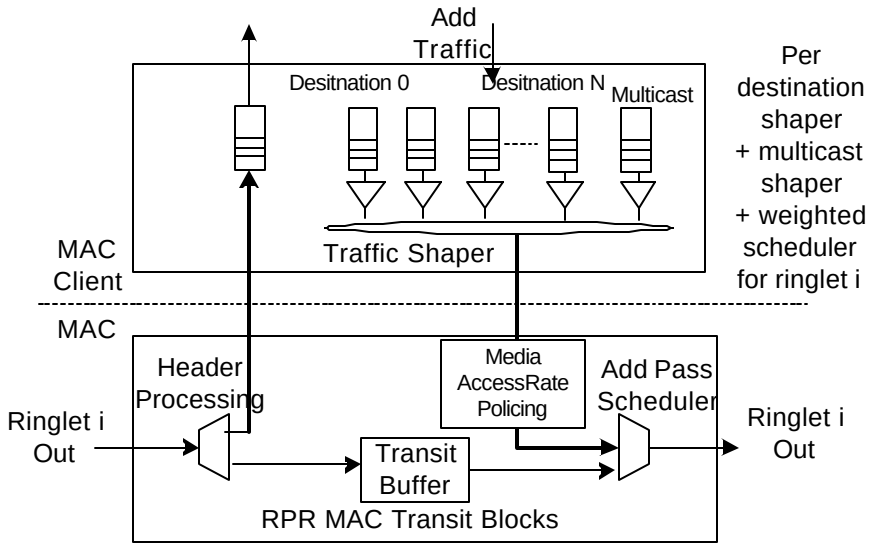


Figure 4: Virtual Output Queuing in the Add Path in the MAC Client

1.7.4.2 Class of service

Services on the Internet can be categorized as follows:

- Provisioned delay sensitive service
- Provisioned delay insensitive service
- Over committed provisioned delay insensitive service
- Best effort service

To manage and support all of these services simultaneously, the 802.17 framework in this proposal supports all of these services using differentiated classes of service. Once an incoming packet is classified, marked, and policed according to the relevant SLA, the only information appended to these packets are the COS bits on the ring header associated with each class. In the MAC Client shown in Figure 5, packets are sent to 8 (or fewer than 8) class queues on a per destination basis (VoQ). The 8 class queues corresponding to a given destination are scheduled on a strict priority basis.

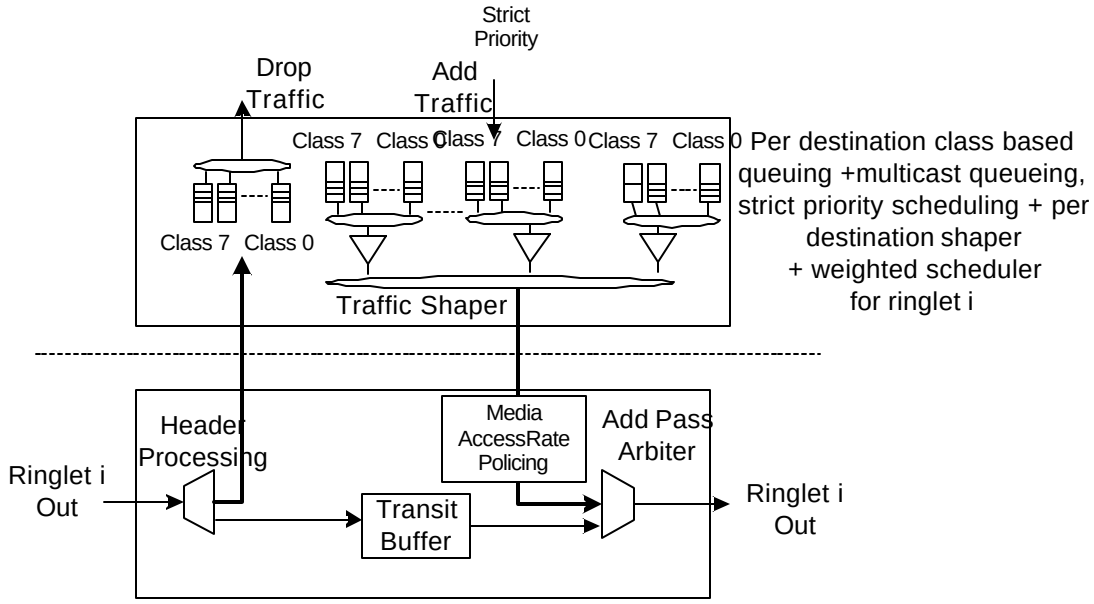


Figure 5: Class of service based queuing per destination and scheduling in the add path in the MAC Client

1.7.4.3 Multicast support

Multicast traffic is expected to be source stripped. Thus the furthest node is the node it self. Thus there has to be a separate destination queue structure for multicast add traffic in the add path shown in Figure 5. This destination queue structure will be shaped according to an available bandwidth all around the ringlet to the source node.

Annex B: Multiple Rings

Multiple ringlets are a generalized case of bi-directional rings. Multiple provide more efficient bandwidth management as it allows N+1 protection option, where protection of N transport paths is enabled with only 1 backup path. Figure 6 shows multiple transit paths, one for each of the N ringlets. Bandwidth management is performed on each ringlet independently as discussed in section x. Add traffic in the MAC client is destined for one of the N +1 multiple rings. Add traffic options are discussed in the Annex section x. Link aggregation is a special case of multiple rings. If every node in the is tapping into all N ringlets than it will the case for link aggregation

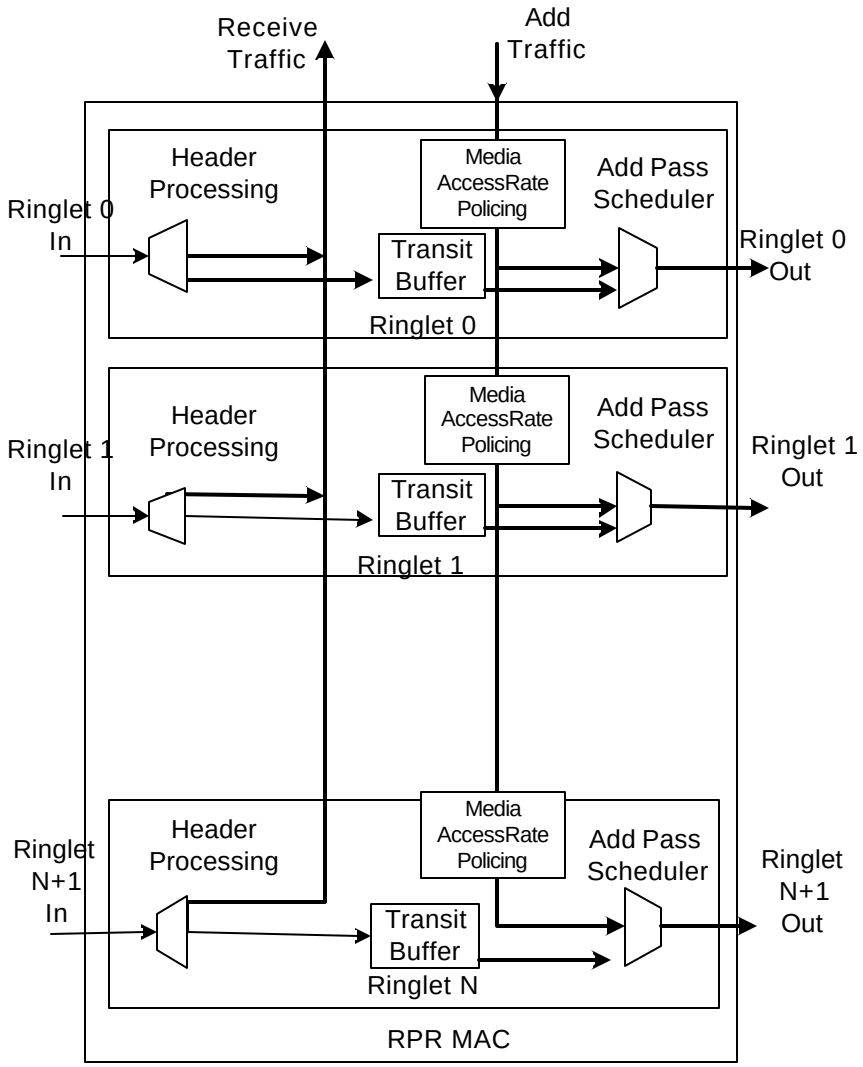


Figure 6: Multiple ringlets make one single RPR MAC. Each ringlet is shown with its own transit path.

Choosing a transit path among multiple ringlets

MAC that consists of multiple ringlets shown in Figure 7 allows MAC client to choose a ringlets independently on per destination basis. MAC client is aware of bandwidth constraint of each ringlet as RPR MAC advertises the rate information (RCM messages) to the Mac client queues on per destination basis. RCM messages are sent with ringlet IDs. Protected traffic should choose a ringlet that has maximum bandwidth available in the least cost direction while making sure least this much bandwidth is available in least one other ringlet in the other direction to serve the protected bandwidth in case of protection event while providing maximum special redundancy (against link and node failures). Unprotected traffic on the other hand can choose a ringlet that provides the maximum bandwidth (Maximum RCF factor) even though equal amount is not available in any other ringlet. Unprotected traffic doesn't have to choose a least cost direction.

Protected traffic chooses the "least-cost" ringlet at the time of provisioning. Protected traffic will not automatically switch from ringlet to ringlet or from direction to direction if another path becomes lower cost after

the connection is already provisioned. A manual command needs to be provided for this purpose. Unprotected traffic of a variety of service types also should not automatically switch. Some services can be load-balanced (if they are not loss sensitive or sensitive to re-ordering).

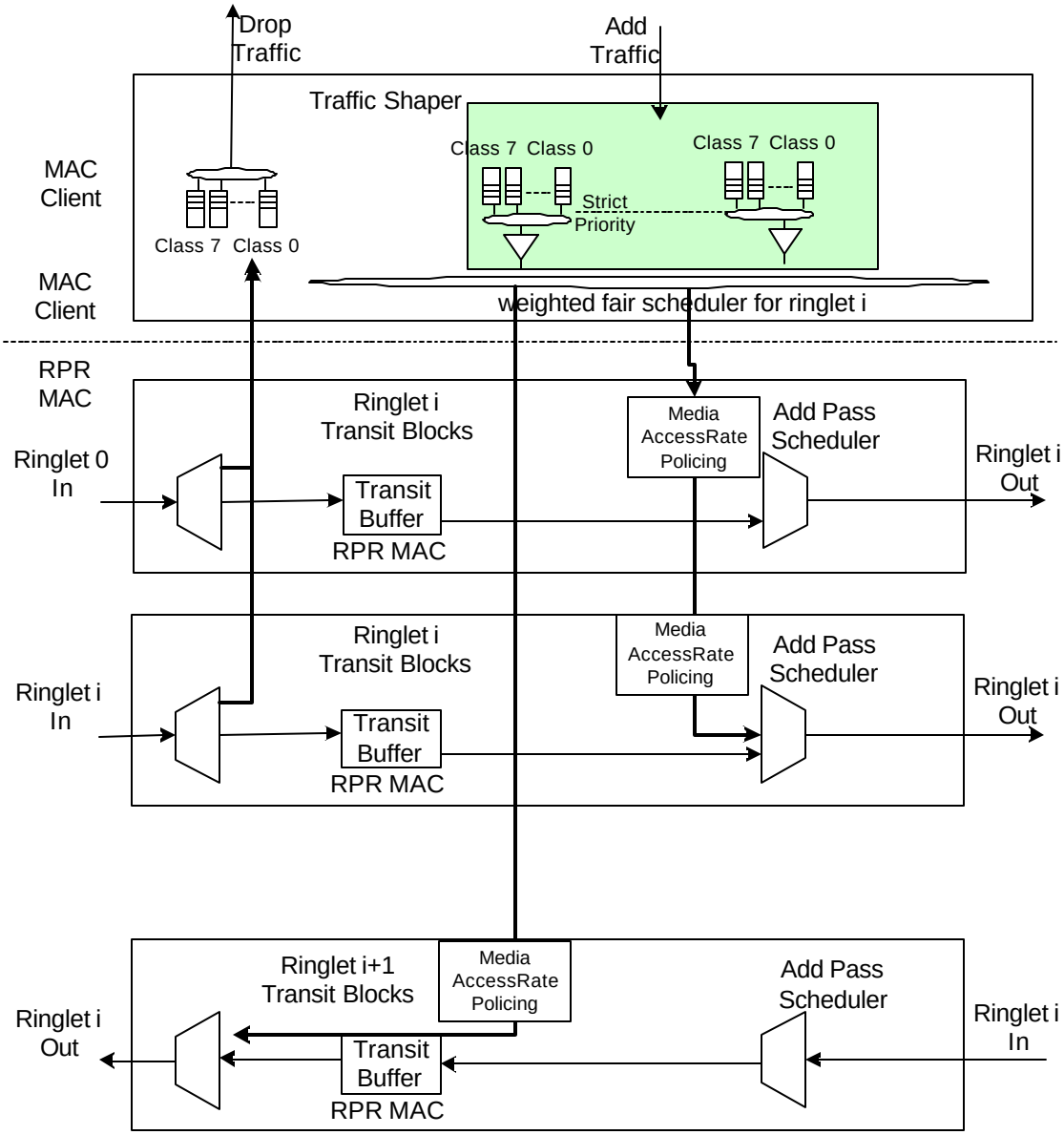


Figure 7: Choosing ringlet among N ringlets

1.7.4.4 Queuing options for protected and unprotected traffic

Protection is another form of quality of service. Since protected and unprotected traffic can be choosing different ringlets, MAC client can also queue traffic on protected and unprotected basis as shown in Figure 8. This increases number of queues as now we have queues = number of destination*number of classes* 2 (for protection traffic and unprotected traffic).

We can reduce number of queues with following options:

- Associated protection with class of service. Certain classes are protected and others are not.
- Associate certain destination with protection. Certain destination are protected others are not.

- All traffic is protected.

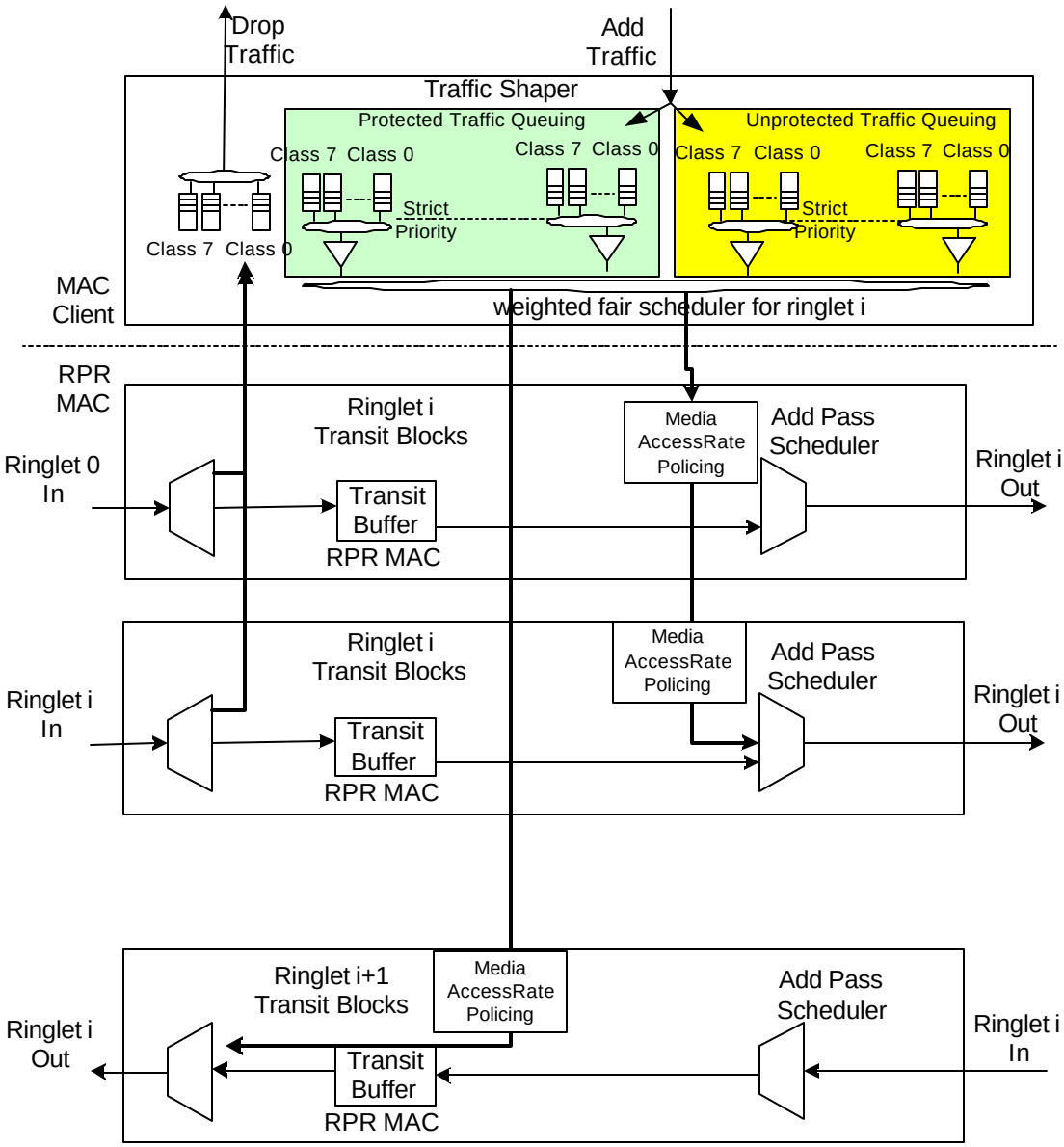


Figure 8: Protection based queuing in the MAC Client add path for multiple rings.

Annex C: Different Span Bandwidth Option

(Sanjay: as we discussed we will not send this ANNEX C to wider distribution outside our group for their support for this September meeting. But we will keep it in our working document for future discussion)
There are two cases for different span bandwidth:

Outgoing link speed > Incoming link speed

Outgoing link speed < Incoming link speed

If the outgoing link speed is greater then the incoming link speed, the cut through is not supported. Only mode supported is store and forward then every transit packet has to be received completely before it is retransmitted.

If the outgoing span from the node is lower speed than the incoming span. Bandwidth provisioning makes sure that upstream nodes do no send

Transit bandwidth = Outgoing ringlet link capacity – add bandwidth. [1]

Instantaneous traffic burst from incoming high-speed link can result into packet drop. Thus there is a need for the rate adaptation buffer that accommodates at most max burst size in the transit. Maximum burst size seen by a single node was given in equation [2]

Largest burst contributed from all nodes
*<= (N/2)*Maximum burst size transmitted by each upstream station in normal operation*
*= N * Maximum burst size transmitted by each upstream station in normal operation* [2]

Thus,
Max Transit buffer = N* *Maximum burst size transmitted by each upstream station* [3]

Where:
N is the total number of nodes in the ring.

Please note that required amount of transit buffer for different span bandwidth support stays constant as the link speed scales, ringlet in and out difference changes. With Max burst size = 10k (RPR MTU), 64 nodes in the ring. Max Transit buffer = 64*MTU = 640k = 5.12 Mbit. This much buffer can be easily accommodated on the onchip SRAM memory block.

Thus if we allow the option of flexibility in the transit path leave it as an implementation choice we can accommodate the different span bandwidths on the incoming and outgoing ringlet spans.

Nodes with different transit buffer that have transit buffer=N*MTU in the transit path shown in Figure 9 can accommodate different span bandwidths on incoming and outgoing spans. The nodes that only have single MTU sized buffer cannot accommodate above features. Thus transit buffer option only affect local add traffic.

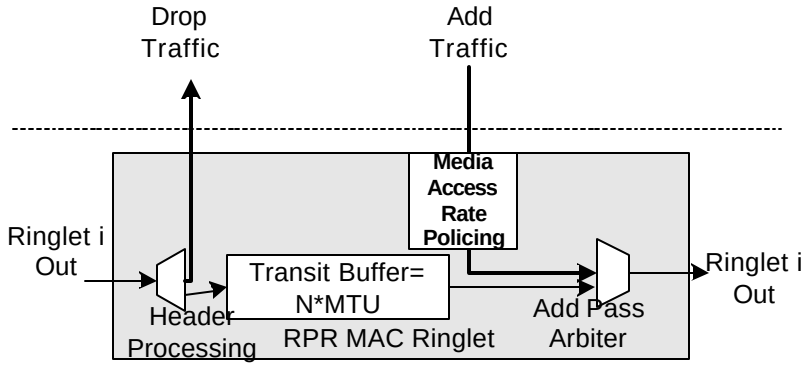


Figure 9: Transit Path design with the option of N*MTU worth of buffer in the transit path

Annex C: Transit Path design in Promiscuous Mode of Operation

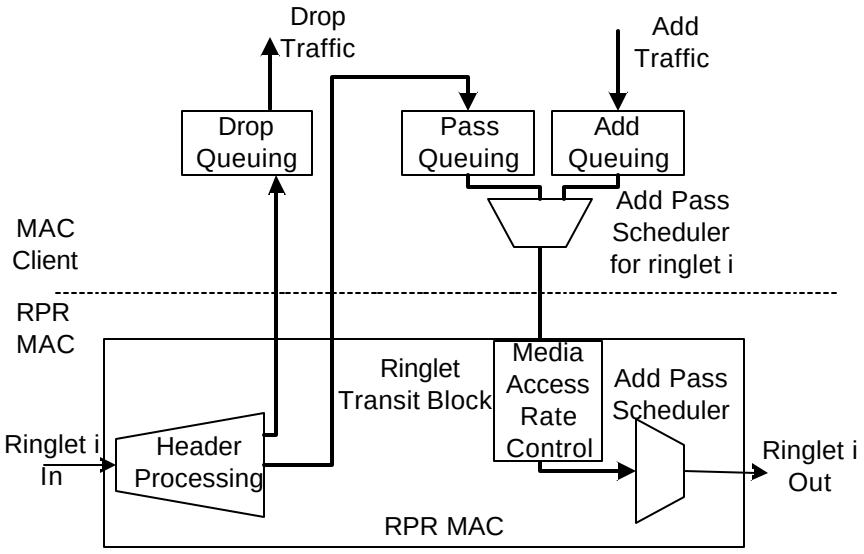


Figure 10: Optional transit path design in promiscuous mode