IEEE 802.17 Resilient Packet Ring Background and Overview

Fredrik Davik, Mete Yilmaz, Stein Gjessing, Necdet Uzun

Abstract — **The IEEE Working group P802.17 is standardizing a new ring topology network architecture, called the Resilient Packet Ring (RPR), to be used mainly in metropolitan and wide area networks. This paper presents a technology background, gives an overview, and explains many of the design choices the RPR working group faced during the development of the standard. Some major architectural features are illustrated and compared by showing performance evaluation results using the RPR simulator developed at Simula Research Laboratory using the OPNET Modeler simulation environment.**

Index Terms— **Communications, Networking, MAN, WAN, Ring networks, Spatial reuse, Fairness.**

Fredrik Davik is with Simula Research Laboratory and Ericsson Research
Mete Yilmaz is with Cisco Systems
Stein Gjessing is with Simula Research Laboratory and is a visiting scholar at Department of Computer Engineering, San Jose State University
Necdet Uzun is with Cisco Systems

A shorter version of this report is to appear in IEEE Communications Magazine March 2004.

Simula Research Laboratory, Technical Report 11-2003, December 2003

I. INTRODUCTION

The Resilient Packet Ring (RPR, IEEE 802.17) is the latest development in a series of ring based network protocols standardized by IEEE [6]. IEEE 802.5 Token Ring [7] and IEEE 1596 Scalable Coherent Interface (SCI) [8] are examples of other ring based IEEE standards. Packet ring based data networks were pioneered by the Cambridge Ring [10], followed by other important network architectures, notably MetaRing [2], FDDI [12], ATMR [13] and CRMA-II [14].

Rings are in general built using several point-to-point connections. When the connections between the stations are bidirectional, rings allow for resilience (a frame can reach its destination even in the presence of a link failure). A ring is also simpler to operate and administrate than a complex mesh or an irregular network.

Networks deployed by service providers in the MANs or WANs are often based on SONET/SDH rings. Many SONET rings consist of a dual-ring configuration in which one of the rings is used as the back-up ring that remains unused during normal operation and utilized only in the case of failure of the primary ring [1]. The static bandwidth allocation and network monitoring requirements increase the total cost of a SONET network. While Gigabit Ethernet does not require static allocation and provides cost advantages; it cannot provide desired features such as fairness and fast (<50ms) auto-restoration.

In order to provide efficient, carrier class packet transport, some companies started to develop proprietary ring technologies. For example, Cisco Systems developed the Dynamic Packet Transport (DPT) [20] technology based on the Spatial Reuse Protocol (SRP) [19], and Nortel Networks developed the OPTera Packet Edge technology [16].

In order to standardize these new initiatives, IEEE was approached. One of the goals of the new standard is to utilize the simplicity of ring networks and use the bandwidth of the dual-ring as efficiently as possible for high-speed data transmission in MANs and in WANs. Important goals also includes distribution of bandwidth fairly to all active stations while providing fast auto restoration. For a rapid and widespread deployment, the reuse of existing physical layers is another important goal. To achieve all of this, the IEEE working group P802.17 was formally started in March 2001 under the name Resilient Packet Ring.

Since RPR is being standardized in the IEEE 802 LAN/MAN (Ethernet) families of network protocols, it can inherently bridge to other IEEE 802 networks and mimic a broadcast medium. RPR implements a Medium Access Control (MAC) protocol, for access to the shared ring communication medium, which has a client interface similar to that of Ethernet's.

The rest of this paper is organized as follows: In section II and III respectively ring network basics and RPR station design is discussed. The so-called fairness algorithm is the topic of section IV, while sections V, VI and VII treat topology discovery, resilience and bridging. Finally frame formats are outlined in section VIII, and a conclusion is given. In order to demonstrate different operational modes, some performance figures are included and discussed. The scenarios have been executed on the RPR simulator model developed at Simula Research Laboratory and implemented in OPNET Modeler [15], according to the latest RPR draft standard as of November 2003 (v3.0).

II. RING NETWORK BASICS

To facilitate discussion of design choices that were made in the development phase of RPR, this section introduces some basic ring networking principles, not all implemented in RPR.

In unicast addressing (broadcast will be covered later), frames are added on to the ring by a sender station, that also decides on which of the two counter rotating rings (called ringlet 0 and ringlet 1 in RPR) the frame should travel to the receiving station. The destination address in the frame header might not be the exact address of the receiving station itself. However, the station should, based on the destination address, recognize that it is the receiving station for this frame. In this way the transmission on the ring (from sender to receiver) might be only one hop in a multi hop transmission from source to destination.

If a station does not recognize the destination address in the frame header, it transits the frame, i.e. the frame is forwarded to the next station on the ring. In RPR, the transit methods supported are cut through (the station starts to forward the frame before it is completely received) and store and forward.

To prevent frames, with a destination address recognized by no stations on the ring, to circulate forever, a time to live field (TTL) is decremented by all stations (as in RPR) or by one station (as in SCI) on the ring. Frames received with a TTL value of 0 is not passed on to downstream stations (is stripped from the ring).

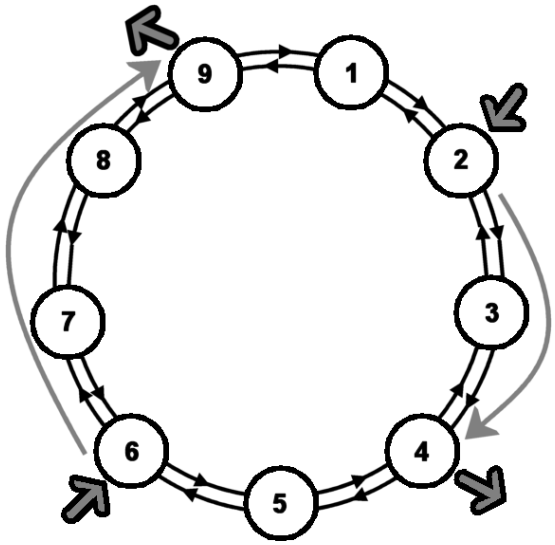

Figure 1. Destination Stripping and Spatial Reuse illustrated on the outer ringlet.

When a station recognizes that it is the receiver of a frame, it may copy the contents of the frame and let the frame traverse the ring back to the sender (like in the Token Ring), it may send back only an acknowledgement (if the station is able to receive the frame) or a

negative acknowledgement (if the station is unable to receive the frame) back to the sender (like in SCI), or it may remove the frame completely from the ring (like in RPR). When the receiving station removes the frame from the ring, the bandwidth otherwise consumed by this frame on the path back to the source, is available for use by other sending stations. This is generally known as spatial reuse.

Figure 1 shows an example scenario where spatial reuse is obtained on the outer ringlet; station 2 is transmitting to station 4 at the same time as station 6 is transmitting to station 9. Destination stripping with spatial reuse was previously exploited in systems like MetaRing [2], ATMR [13], CRMA-II [14] and SCI [8].

The ring access method is an important design choice. A token may circulate the ring, so that the station holding the token is the only station allowed to send (like in Token Ring). An alternative access method, called a "buffer insertion" ring, was developed as early as 1974 [5][17]. Every station on the ring has a buffer called an "insertion buffer" (called a "transit queue" in RPR, see Figure 2) in which frames transiting the station may be temporarily queued. The station must act according to three simple rules. The first principle is that, the station will not add packets to the ring as long as there are packets in the insertion buffer or packets in transit. Secondly, when there is no frame in transit, the station itself is allowed to add a frame. Thirdly, if a passing frame arrives at a station when it has started to add a frame, the frame in transit is temporarily (for as long as it takes to complete the sending of the added frame) queued in the insertion buffer.

Obviously these three simple principles need some improvement to make up a full, working protocol that distributes bandwidth fairly. This has been studied before [3][11][4][9], and how this is achieved in RPR will be revealed in section IV when the RPR fairness algorithm is discussed.

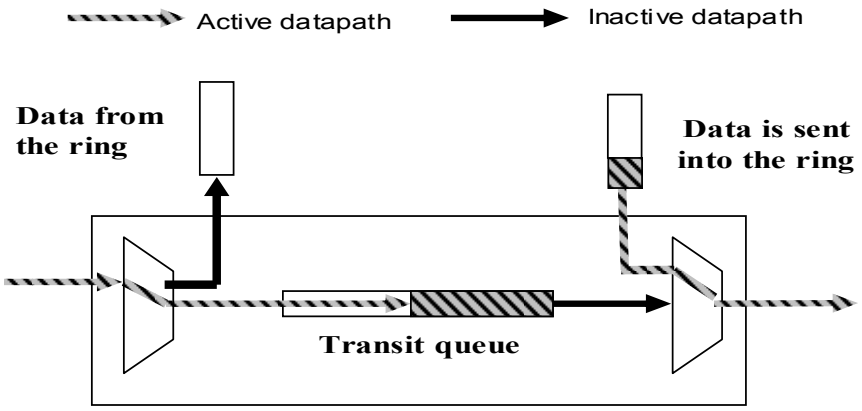

Figure 2. The "insertion buffer" or "transit queue" stores frames in transit, while the station itself adds a frame (here a stations attachment to only one ring is shown).

III. STATION DESIGN AND PACKET PRIORITY

The stations on the RPR ring implements a medium access control (MAC) protocol that regulates the stations access to the ring communication medium. All links around the ring is mandated to have the same capacity (called the full rate). The P802.17 working group has defined several physical layer interfaces (reconciliation sublayers) for Ethernet (called PacketPHYs) and SONET/SDH [6]. The MAC also implements access points that clients can call in order to send and receive frames and status information.

The RPR working group decided to implement a three level, class based, traffic priority scheme. The objectives of the class based scheme is to let class A be a low latency, low jitter class, class B be a class with predictable latency and jitter, and finally class C be a best effort transport class. It is worth to note that the RPR ring does not discard frames to resolve congestion. Hence when a frame has been added onto the ring, even if it is a class C frame, it will eventually arrive at its destination. The design decision behind this choice will be explained later

Class A traffic is divided into classes A0 and A1, and class B traffic is divided into class B-CIR (Committed Information Rate) and B-EIR (Excess Information Rate). The two traffic classes C and B-EIR are called Fairness Eligible (FE), for reasons that will become clear in section IV.

The bandwidth around the ring is pre-allocated in two ways. The first is called "reserved" and can only be used by class A0 traffic, and is equally reserved all around the ringlet. If stations are not using their pre-allocated A0 bandwidth, this bandwidth is wasted. In this way TDM-like traffic can be sent by RPR stations as A0 frames.

The other pre-allocated bandwidth is called “reclaimable”. A station that has class A1 or B-CIR traffic to send, pre-allocates “reclaimable” bandwidth for these types of traffic. If not in use, such bandwidth can be used by FE traffic. In addition, any bandwidth not pre-allocated is also used to send FE traffic. The distribution and use of unallocated and unused reclaimable bandwidth (FE bandwidth) is dynamically controlled by the fairness algorithm.

A station’s reservation of class A0 bandwidth is broadcasted on the ring using topology messages (topology messages will be discussed later). Having received such topology messages from all other stations on the ring, every station calculates how much bandwidth to reserve for class A0 traffic. The remaining bandwidth, called “unreserved rate” can be used for all other traffic classes. This in general means that the unreserved rate is the full rate minus the sum of all station’s A0 reservations. The standard claims that spatial reuse of A0 bandwidth is possible but does not specify how this can be obtained.

An RPR station implements several traffic shapers that limit and smooth the add- and transit traffic. There is one shaper for each of the traffic classes A0, A1, B-CIR as well as one for FE traffic and one for non-A0 traffic. The shapers for class A0, A1, B-CIR and non-A0 traffic are preconfigured, while the FE shaper is dynamically adjusted by the fairness algorithm. The non-A0 shaper (also called the Downstream shaper) limits the amount of non class A0 traffic transmitted, ensuring that reserved bandwidth is available all around the ring. The rate of this shaper is the full rate minus the bandwidth reserved for class A0 traffic (i.e. unreserved rate).

When a station tries to send more class B traffic than the B-CIR shaper allows, the rest of the class B traffic is sent as class B-EIR. Class B-EIR traffic has higher add-priority than class C traffic.

It is important that the preconfigured shapers are correctly set, in particular that the sum of the bandwidth allocated for A0, A1 and B-CIR traffic does not surpass the maximum bandwidth (the full rate).

It is not obvious that class A0 pre-allocated and reserved bandwidth may be spatially reused. For example stations 2 and 6 in figure 1 may both send 10Mbit/sec class A0 traffic to respectively stations 4 and 9. If this is the only class A0 traffic reserved, then ideally, the spatial reuse feature should allow us to reserve only 10Mbit/s all around the ring. However, when the constant "reserved bandwidth" is calculated, it is calculated as the sum of all A0 allocations. If the downstream shaper is set to unreserved rate (full rate minus reserved rate), no spatial reuse will be achieved for A0 traffic. If, on the other hand, the downstream shapers are set lower than the unreserved rate (because of known spatial reuse of A0 traffic), spatial reuse of A0 traffic is indeed achieved. RPR does not contain mechanisms to ensure the intended spatial reuse, and it may even not be mandated by the final standard to set the downstream shaper to any other value than the unreserved rate. Anyhow, if stations 2 and 6 can not be trusted to send to non-overlapping segments of the ring, a total of 20 Mbit/sec must be reserved around the ring for class A0 traffic. Misconfiguration of the downstream shaper may cause serious problems at run time.

Also class A1 and C-BIR traffic may be spatially reused, so that the total pre-allocated bandwidth on any link may be calculated taking spatial reuse into consideration, in the same way as explained above for class A0 traffic. But also for these classes of service, it is outside the scope of the RPR standard to enforce spatial reuse. Hence, also here it might be wise to assume that all stations send all their class A1 and B-CIR traffic all around the ring, and that the total pre allocated class A1 and B-CIR bandwidth is the sum of all station's allocations. The difference between allocation of A1 and B-CIR bandwidth with or without spatial reuse, only affects the calculation that is needed to ensure that the total ring capacity is not surpassed. Since unused class A1 and B-CIR bandwidth is reclaimable, this unused bandwidth may anyhow be used by the fairness algorithm to send FE-traffic (class B-EIR and class C traffic).

The minimum transit queue size is the maximum transfer unit that a station itself may add (because this is the maximum buffer size needed by the frames in transit while the station adds a new frame). Some flexibility for scheduling of frames from the add- and transit-path can be obtained by increasing the size of the transit queue. For example, a station may add a frame even if the transit queue is not completely empty. Also a larger queue may store lower priority transit frames while the station is adding high priority frames. The transit queue could have been specified as a priority queue, where frames with the highest priority are dequeued first. This was considered too complex and instead the working group decided that a station optionally may have two transit queues. Then high priority transit frames (class A) are queued in the Primary Transit Queue (PTQ), while class B and C frames are queued in the Secondary Transit Queue (STQ). Forwarding from the PTQ has priority over the STQ and most types of add traffic. Figure 3 shows one ring interface, with the three add- and two transit queues. The numbers in the circles indicate a crude priority on the output link. Regarding priority between add traffic and the STQ, as the STQ fills up, it will have increasingly higher priority (this is not a linear function, but based on thresholds). Since class A frames have priority over all other

traffic, a class A frame traveling thee ring will usually experience not much more than the propagation delay and some occasional transit delays waiting for outgoing packets to completely leave the station (RPR does not support pre-emption of packets).

When in transit, both class B and C frames are stored in the STQ, hence, once added to the ring, they experience delay values within the same range. The difference between class B and class C frames is the scheduling at the ingress. Class B-CIR (as well as class A1) add frames have priority over the class B and class C frames in the STQ (as long as the STQ is not completely full). The worst case delay for class B-CIR frames is the propagation delay, plus the maximum delay that both B and C frames can experience in passing through the (FIFO) STQs on their way around the ring to the receiver. Class B-EIR and class C frames may have to wait very long to get onto the ring, depending on how much bandwidth is consumed by A and B-CIR frames, and how many other stations are adding class B-EIR and class C traffic. Hence it is very hard to give any bound on the latency of class B-EIR and class C frames.

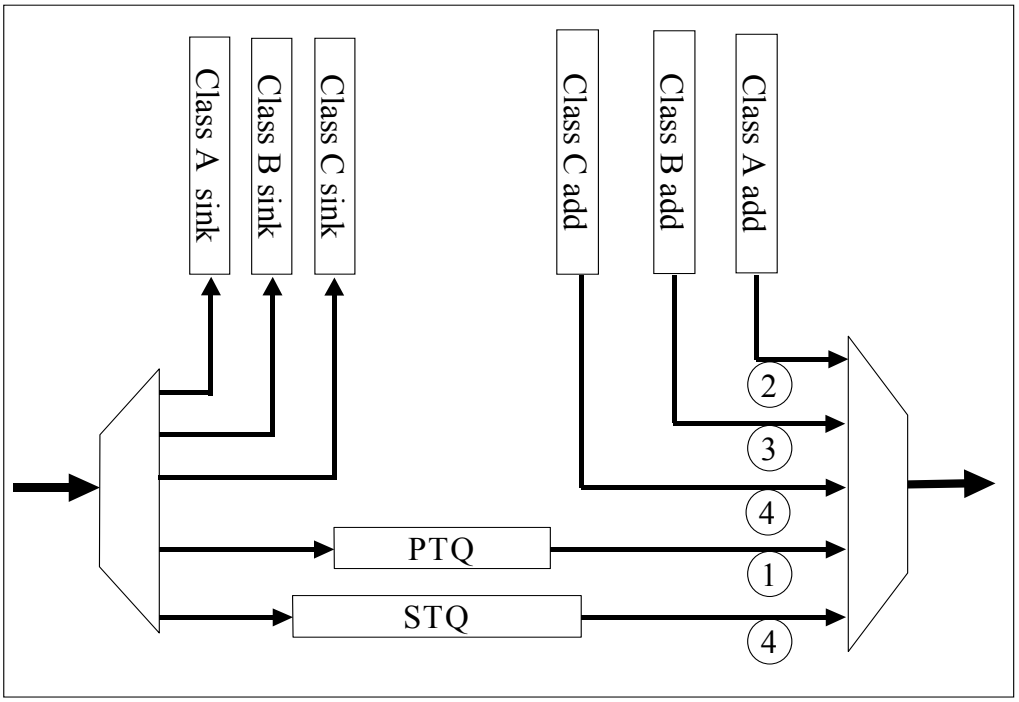

Figure 3. The attachment to one ring by a Dual Transit Queue Station. The numbers in the circles give a very crude indication of output link priority.

An RPR station may, however, have one transit queue only (the PTQ). In order for class A traffic to move quickly around the ring, the transit queues in all single transit queue stations should then be almost empty. This is achieved by letting transit traffic have priority over all add traffic, and by requiring all class A traffic to be reserved (class A0). Hence there will always be room for class A traffic and class B and C traffic are competing for the remaining bandwidth, just like in the two transit queue stations.

Figure 4 shows part of an example run where the latency of frames sent between two

given stations on an RPR ring is measured. The ring is overloaded with random background, class C, traffic. Latency is measured from the time a packet is ready to enter the ring (i.e. first in the ingress queue), until it arrives at the receiver. Notice how class A traffic keeps its low delay even when the ring is congested. Also notice how class B traffic still have low jitter under high load, while class C traffic experiences some very high delays. This example was run with two transit queue stations.

An RPR ring may consist of both one and two transit queue stations. The rules for adding and scheduling traffic are local to the station, and the fairness algorithm described below works for both station designs.

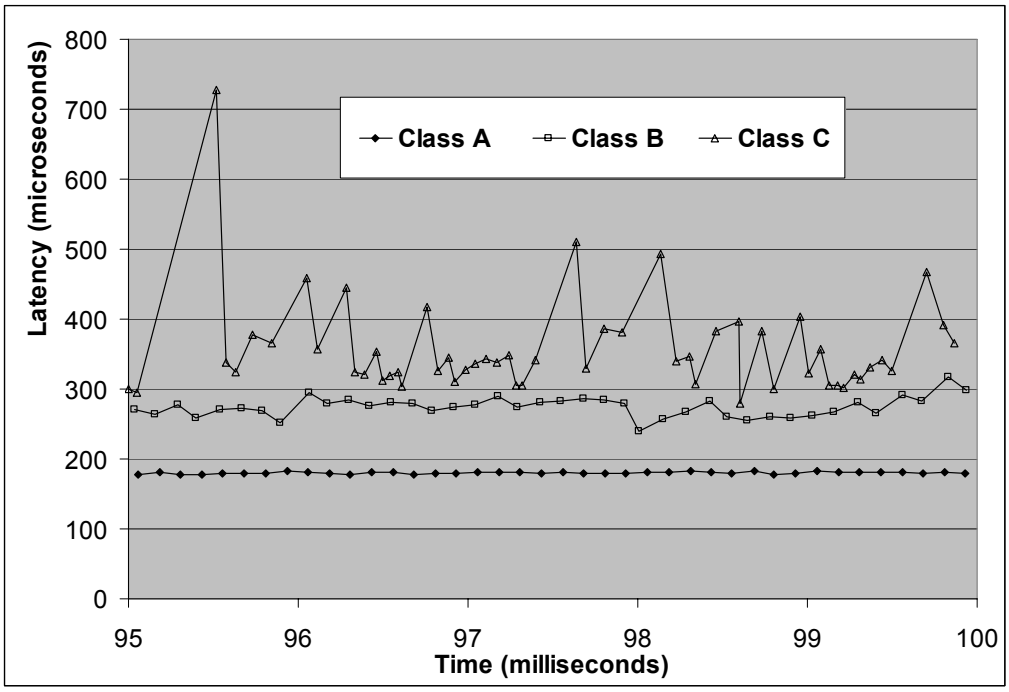

Figure 4. Frame latency from station 1 to station 7 on a 16 station overloaded ring. The propagation and minimum frame latency is 180 microseconds.

IV. THE RPR FAIRNESS ALGORITHM

In the basic "buffer insertion" access method, a station may only send a frame if the transit queue is empty. Hence it is very easy for a downstream station to be starved by upstream stations. In RPR, the resolution to the starvation problem is to enforce all stations to behave according to a specified "fairness" algorithm.

The objective of the fairness algorithm is to distribute unallocated and unused reclaimable bandwidth fairly among the contending stations and use this bandwidth to send class B-EIR and class C traffic, i.e. the fairness eligible (FE) traffic. Class A0 traffic is obviously not affected, since bandwidth is reserved for this class exclusively. Classes A1 and B-CIR are indirectly affected, as will be explained below.

When defining fair distribution of bandwidth, the P802.17 working group decided to enforce the principle that when the demand for bandwidth on a span of the ring is greater

than the supply, the available bandwidth should be fairly distributed between the contending source stations. This is the same principle as outlined by the so called RIAS fairness [4]. The working group also decided that a weight is assigned to each station so that a fair distribution of bandwidth need not be an equal one.

There are several ways to ensure fair allocation of bandwidth around the ring. One way is an algorithm that enquires around the ring how much bandwidth each of the stations need, and then notifies them afterwards how much they were allocated [18]. RPR takes another approach. When the bandwidth on the output link of a station is exhausted (the link is congested), the fairness algorithm starts working to distribute this bandwidth fairly. The most probable cause of congestion is the station itself and its immediate upstream neighbors. Hence by sending a so called fairness message upstream (on the opposite ring) the probable cause of the congestion is reached faster than by sending the fairness message downstream over the congested link. Figure 5 shows how the attachment to one ring asks the other attachment to queue and send a fairness message. In the sequel we focus on fairness on one ring. The fairness algorithm on the other ring works exactly the same way.

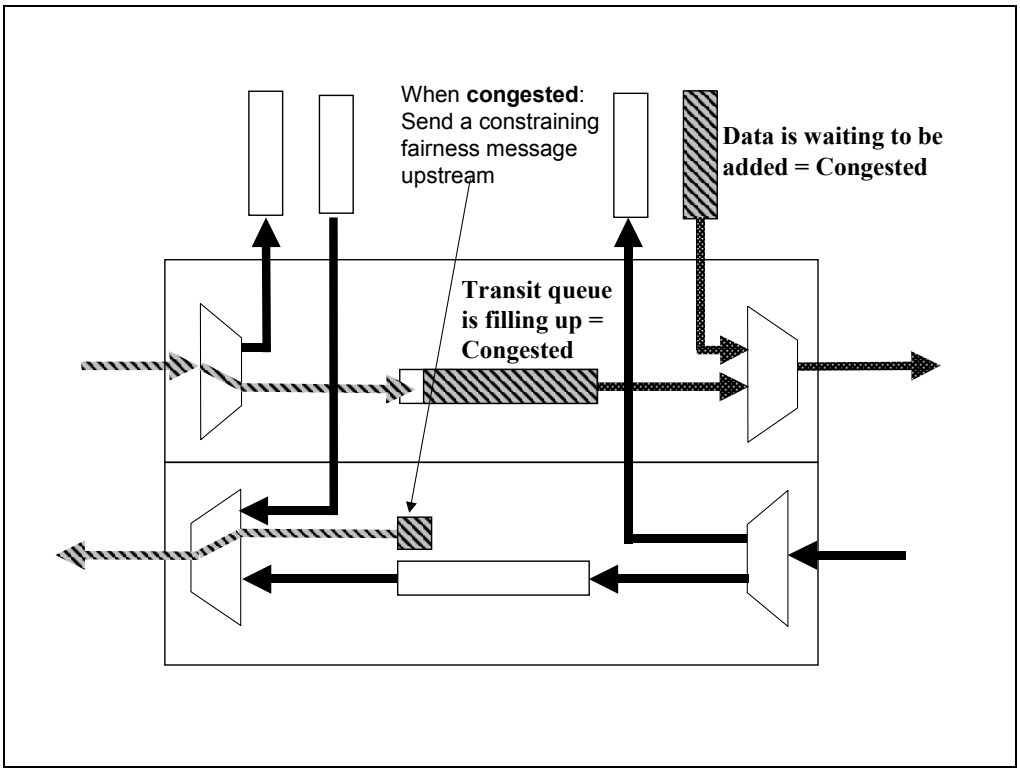

Figure 5. When a station becomes congested it sends a fairness message upstream.

That a link is congested (or that it is very close to being congested) may be identified in several ways by the attached upstream station: a) Frames that are to be added have to wait a long time before they are forwarded, b) the transit buffer is filling up (and hence transit frames have to wait a long time before they are sent on), and c) if the capacity of the

transmit link is (almost) fully used. The first method is used in a single transit queue design, the second method is used in the dual transit queue design, and the third method is used in both designs.

The congested station tries to make a first estimate at what the fair rate should be. After an estimate is calculated, this value is sent upstream in a fairness message. The station receiving the fairness message will decrease it's add rate to this value. If this station also transits more traffic than this value, the fairness message is transmitted further upstream, in order to tell more stations to decrease their add rate. In this way all stations upstream from the congestion point, that are contributing to the congestion, are notified and have to throttle their add traffic to the rate given in the fairness message originating from the congestion point. This segment of the ring (that receives a fairness message with the same value) is called a congestion domain.

There are two methods specified for fair rate estimation and adjustment: 1) The Aggressive method: 2) The Conservative method. The goal of the Aggressive fairness method is to quickly respond to changes in the traffic load. The Conservative option performs rate adjustments more restrictively. It makes an adjustment and waits to observe the effect before making a new decision. Since the station design with one transit queue needs a more restrictive rate control, the conservative method of rate estimation and adjustment may be a good match.

The main difference between conservative and aggressive fairness is the way the fair rate is initially estimated, and how it is adjusted towards the real fair rate. In the conservative mode, the congested station calculates the initial fair rate either by 1) dividing the available bandwidth between all upstream stations that are currently sending frames through this station or by 2) use its own current add rate. A timer is used to ensure that additional rate changes are made only when the congested station have had time to see how this new fair rate affects the congestion (i.e., gets better or worse). The period of this timer is referred to as the Fairness Round Trip Time (FRTT). FRTT is an estimate of the time it takes for a congested station to see the full effect of the fairness message it sent to upstream stations. FRTT consists of two parts: 1) the propagation delay for a class A frame when transmitted from the congestion domain head (i.e. the congested station) to the congestion domain tail (the station at the other end of the congestion domain) and back (LRTT – Loop Round Trip Time). 2) The difference between the propagation delay for a class C and a class A frame sent from the tail to the head (FDD – Fairness Differential Delay). LRTT needs to be computed on initialization of the ring and when the topology changes, while FDD is computed when a station becomes tail of a congestion domain and thereafter at configurable intervals. FDD reflects the congestion situation, i.e. the STQ fill levels on the transit path from head to tail. As the congestion domain changes, so does the FRTT. LRTT and FDD frames are special types of control frames.

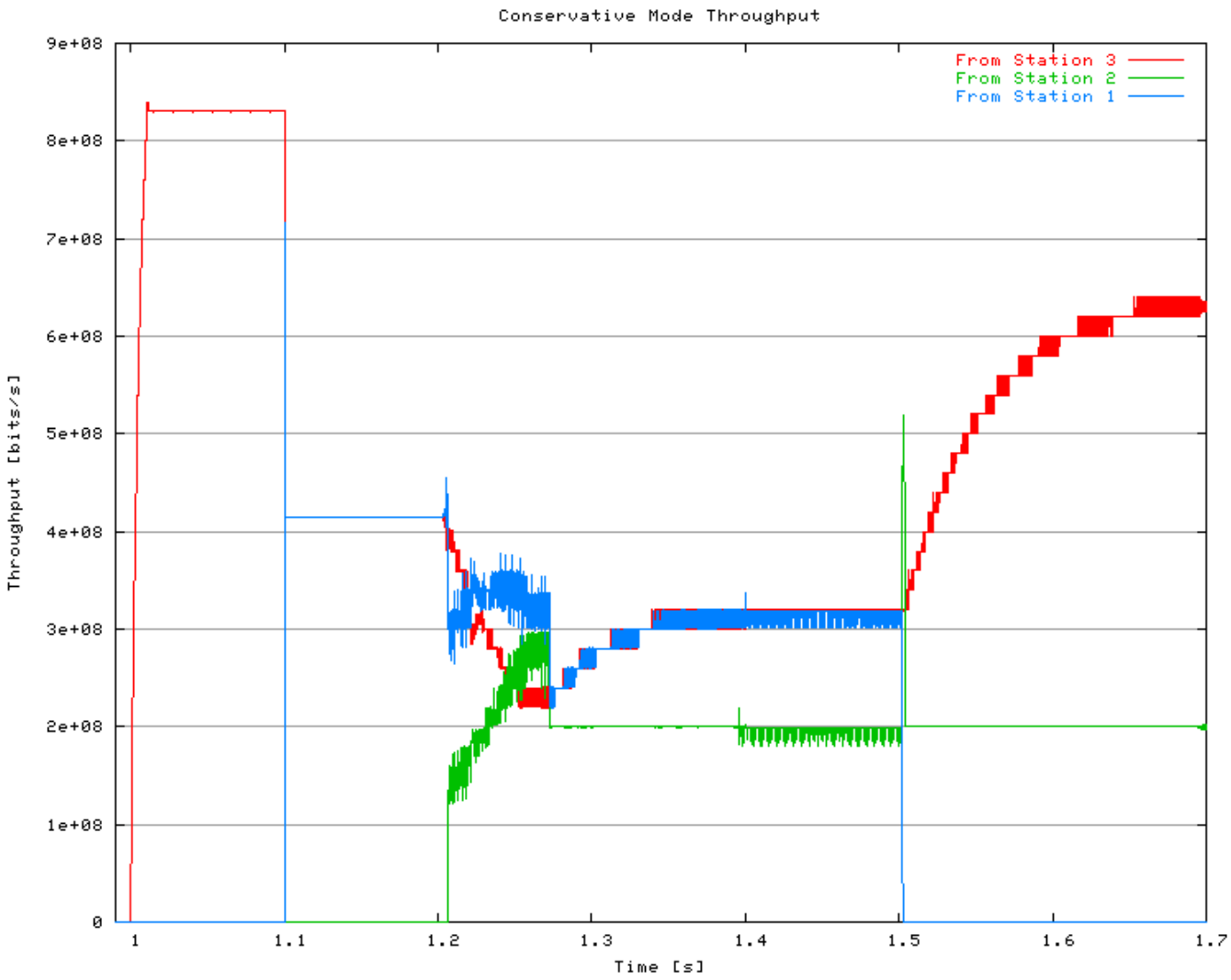

Figure 6. Dynamic traffic handled by the conservative fairness algorithm. Number of bits/sec. as received by station 4.

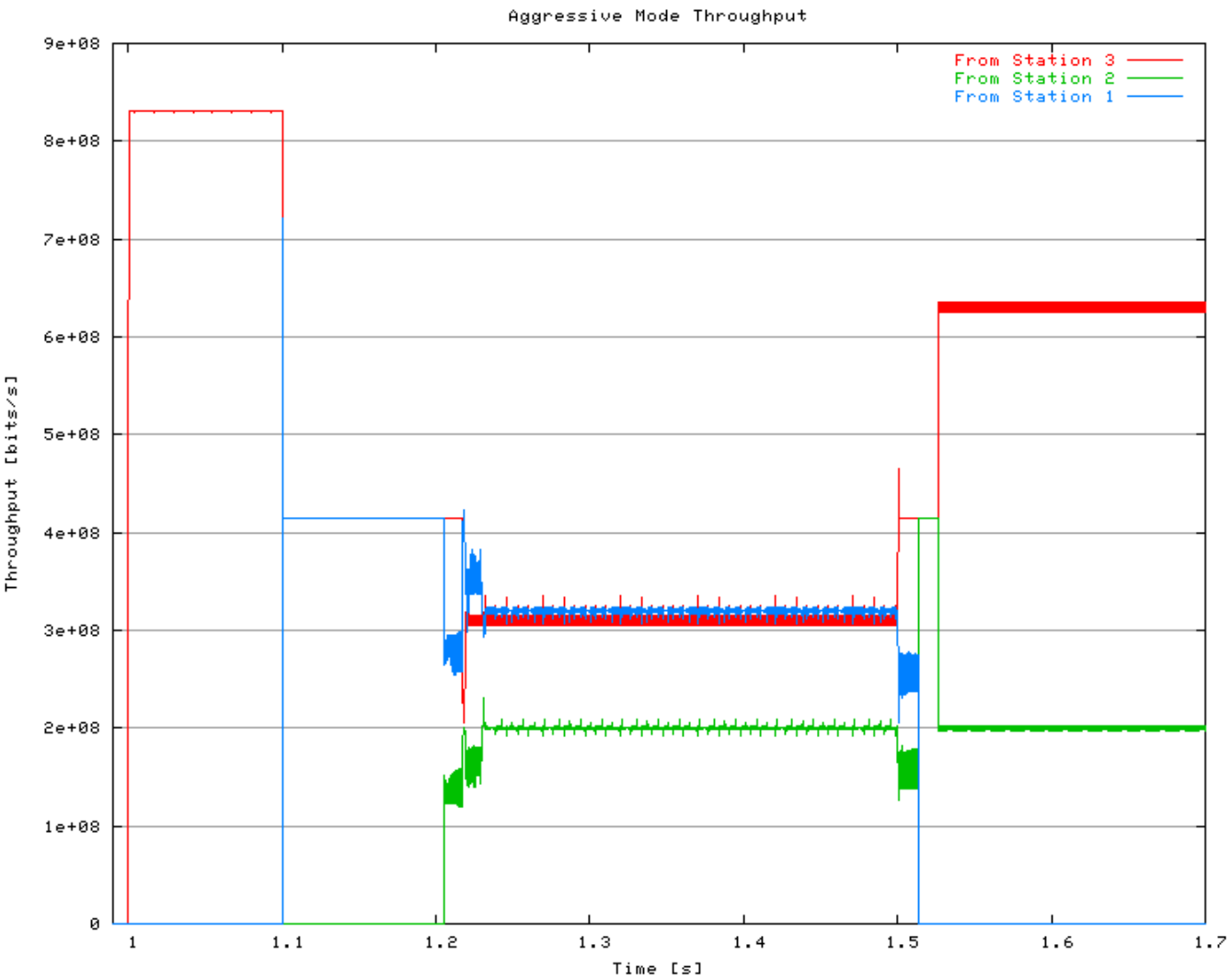

Figure 7. Dynamic traffic handled by the aggressive fairness algorithm. Number of bits/sec. as received by station 4.

In the Aggressive mode, the congested station makes a first estimate of the fair rate equal to the rate the station itself lately have been able to add to the ring. Since the station is congested, this means that it has been able to send very little traffic onto the ring recently. Hence this estimate is probably too low, but it is used as a starting point and a way to alleviate congestion. When congestion is indeed removed, the (previously congested) station will not send any more fairness messages upstream, or more correctly it will send fairness messages with a default fair value representing the full link rate (such frames are sent all the time with preset intervals as heart beats.) A station receiving a fairness message indicating no congestion (i.e., full link rate) will increase its add traffic (assuming the station's demand is greater than what it is currently adding). In this way (if the traffic pattern is stable) the same station will become congested again after a while, but this time the estimated fair rate will be closer to the real fair rate, and hence the upstream stations do not have to decrease their traffic rate as much as previously.

Figure 6 and Figure 7 shows how respectively the conservative and the aggressive fairness algorithm work for a given scenario. Both scenarios are simulated on a 16-station ring, with 50 km one Gbit/sec links and 500 Byte packets, of which each station uses 1 % for A0 traffic. Stations 1, 2 and 3 are sending to station 4. The traffic starts at time 1.0 sec., and initially only station 3 is sending. At time 1.05 sec. station 1 starts sending. Both of these flows are greedy class C flows, and both fairness methods are quick to share the bandwidth on the congested link (from station 3 to station 4) equally. At time 1.2 sec. station 2 starts sending a 200 Mbit/sec flow (also class C frames to station 4). We see that the aggressive method very quickly (but after some high oscillations) changes to the new fair distribution of bandwidth. The conservative method, however, waiting longer between every rate adjustment, uses more time to converge to the new fair division of add rates.

At time 1.5 sec., the traffic from station 1 stops. For both methods we see that some traffic from station 2 that has been queued, now are being released, and hence there are an added number of station 2 packets received at station 4. The aggressive method has some additional oscillations, but otherwise adjusts quickly the new traffic pattern. The conservative method adjusts with fewer oscillations, but more slowly.

The fairness algorithm affects class A1 and B-CIR traffic indirectly. When stations have less such traffic to send than they have pre-allocated, this bandwidth may be used to send FE traffic. When a station, at a later time, wants to reclaim this bandwidth (now it has more A1 or B-CIR traffic to send), this bandwidth is taken away from the FE traffic, and a link may now be (even worse) congested. Because class A1 and B-CIR add traffic have priority over traffic from the STQ, the STQ occupancy increases while the fairness algorithm tries to throttle upstream FE traffic senders. The STQ should be sized so that it does not overflow when this situation occurs (discussed further below).

In the single transit queue station design, the transit traffic has priority over all add traffic, hence no frame will ever be lost on the way around the ring. In a dual transit queue architecture, class A1 and B-CIR add traffic has priority over traffic transiting in the STQ. This may cause the STQ to fill up, and if the fairness algorithm does not work fast enough (is not able to stall upstream traffic fast enough), and the STQ is small, then the STQ may fill completely. If the STQ is full, and there is class A or B-CIR traffic to be added, there is a design choice between throwing away frames in the STQ, or stalling the class A and B-CIR add traffic. The latter choice would cause priority inversion. Because the STQ should never be completely full if the system is properly configured (the STQ is large enough) and the stations do not send excess traffic, the RPR working group decided that RPR will not throw frames away. A short priority inversion period is what will happen in the unlikely event that an STQ becomes full.

As noted above, the way to avoid priority inversion in the dual transit queue case, is to have a large STQ that is able to store class B-EIR and class C traffic while the station adds class A1 and B-CIR traffic and the fairness algorithm tries to alleviate congestion. The negative factors in having a large STQ is obviously that traffic transiting a very full STQ will experience severe delay. However, if the STQ's were not that large, the delay would be experienced in the ingress buffers instead. The fairness algorithm is triggered when the STQ becomes partially full. This threshold is named "stqLowThreshold" and the default value is 1/8 of the size of the STQ. By having a large STQ, some variation in traffic load

may be experienced and smoothed by the STQs even without the fairness algorithm kicking in.

In the fairness algorithm, as explained above, a station sees at most one congestion point. There is an optional fairness message in RPR called a multi choke fairness message. Each station on the ring puts its own congestion status (that tell how much its output link is used presently) in such a message and sends it to all the other stations on the ring. A receiving station may collect these messages and build a global image of the congestion situation on the ring, and schedule its add traffic accordingly. How this is done is however outside the scope of the standard.

v. TOPOLOGY DISCOVERY

Topology discovery determines connectivity and the ordering of the stations around the ring. This is accomplished by collecting information about the stations and the links, via the topology discovery protocol. The collected information is stored in the topology databases of each station.

At system initialization, all stations send control frames, called topology discovery messages, containing their own status, around the ring. Topology messages are always sent all the way around the ring, on both ringlets, with an initial TTL equal to 255 (the maximum number of stations). All other stations on the ring receive these frames. Upon reception of such a frame, the station has two pieces of information, namely 1) the distance, measured in hops, to the origination station (calculated from 255 – frame.ttl + 1) 2) the capabilities of the originating stations as described by the frame contents. Having received such frames from all other stations on the ring, each station has enough information to compute a complete topology image.

When a new station is inserted into a ring, or when a station detects a link failure, it will immediately transmit a topology discovery message. If any station receives a topology message inconsistent with its current topology image, it will also immediately transmit a new topology message (always containing only the stations own status). Hence the first station that notices a change starts a ripple effect, resulting in all stations transmitting their updated status information, and all stations rebuilding their topology image.

The topology database includes not only the ordering of the stations around the ring and the protection status of the stations (describing its connected links, with status signal fail, signal degrade, or idle), but also the attributes of the stations and the round trip times to all the other stations on the ring.

Once the topology information has become stable, meaning that the topology image does not change during a specified time period, a consistency check will be performed. For example the station will make sure that the information collected on one ringlet matches the other.

Even under stable and consistent conditions, stations will continue to periodically transmit topology discovery messages, in order to provide robustness to the correct operation of the ring.

When the client submits a frame to the MAC, without specifying which ringlet to be used, the MAC uses the topology database to find the shortest path. Information in the topology database is also used to calculate the Fairness Round Trip Time in the

conservative mode of the fairness algorithm.

VI. RESILIENCE

As described in the previous section, as soon as a station recognizes that one of its links or a neighbor station has failed, it sends out topology messages. When a station receives such a message telling that the ring is broken, it starts to send frames in the only viable direction to the receiver. This behavior, which is mandatory in RPR, is called steering.

The IEEE 802 family of networks have a default packet mode called "strict" in RPR. This means that packets should arrive in the same order as they are sent. To achieve this after a link or station failure, all stations stop adding packets and discard all transit frames until their new topology image is stable and consistent. Only then will stations start to steer packets onto the ring. Even on a 2000 km ring, it will take a maximum of 50 ms for this algorithm to converge, that is from the failure is observed by one station, until all stations have consistent topology databases and can steer new frames.

RPR optionally defines a packet mode/attribute called relaxed, meaning that it is tolerable that these packets arrive out of order. Such packets may be steered immediately after the failure has been detected and before the database is consistent. Relaxed frames will not be discarded from the transit queues either.

When a station detects that a link or its adjacent neighbor has failed, the station may optionally wrap the ring at the break point (called "wrapping") and immediately send frames back in the other direction (on the other ringlet) instead of discarding them. Frames not marked as wrap eligible (via the *we* frame field) are always discarded at a wrap point.

VII. BRIDGING

The RPR standard specifies methods for networks interconnection by bridging to other network protocols in the IEEE 802 family. An application of this can be transport of Ethernet frames over RPR to provide resilience, class of service support and better utilization of the network.

Figure 8 shows an example, where an RPR ring is bridged to many Ethernet in the First Mile access networks. Station 1 may be the station that includes the interface between the 802-networks (the access and MAN network), and the backbone network.

RPR uses 48-bit source and destination MAC addresses in the same format as Ethernet (see section VIII). When an Ethernet frame is bridged into an RPR ring, the bridge inserts some information into the Ethernet frame in order to transform it into an RPR frame. Similarly this information will be removed if the frame moves from RPR (back) to Ethernet.

When participating in the spanning tree protocol, RPR is viewed as one broadcast enabled subnet, exactly like any other broadcast LAN. The ring structure is then not visible, and incurs no problem for the spanning tree protocol. The spanning tree protocol may not break the ring, but may disable the bridges connected to the RPR stations on the ring.

RPR implements broadcast by sending the frame all around the ring, or by sending the frame half way on both ringlets. In the latter case the TTL field is initially set to a value so that it becomes zero, and the packet is stripped, when it has traveled half the ring. Using broadcast, obviously, no spatial reuse is achieved.

Since RPR can bridge to any other Ethernet, for example Ethernet in the First Mile, we can envision Ethernets spanning all the way from the customer into the Metropolitan or even Wide Area Network. Whether such large and long ranging Ethernets will be feasible or practical in the future, is to be seen.

An extended frame format is also defined in the standard to transport Ethernet frames. In this format an RPR header encapsulates Ethernet frames.

Another way to connect RPR to other data networks is to implement IP or layer 3 routers on top of the MAC clients. In this way RPR behaves exactly like any other Ethernet connected to one or more IP routers. In Figure 8, station 1 may implement an IP-router on top of the MAC client interface. IP routers connected to RPR should in the future also take advantage of the class based packet priority scheme defined by RPR when they send Quality of Service constrained traffic over RPR.

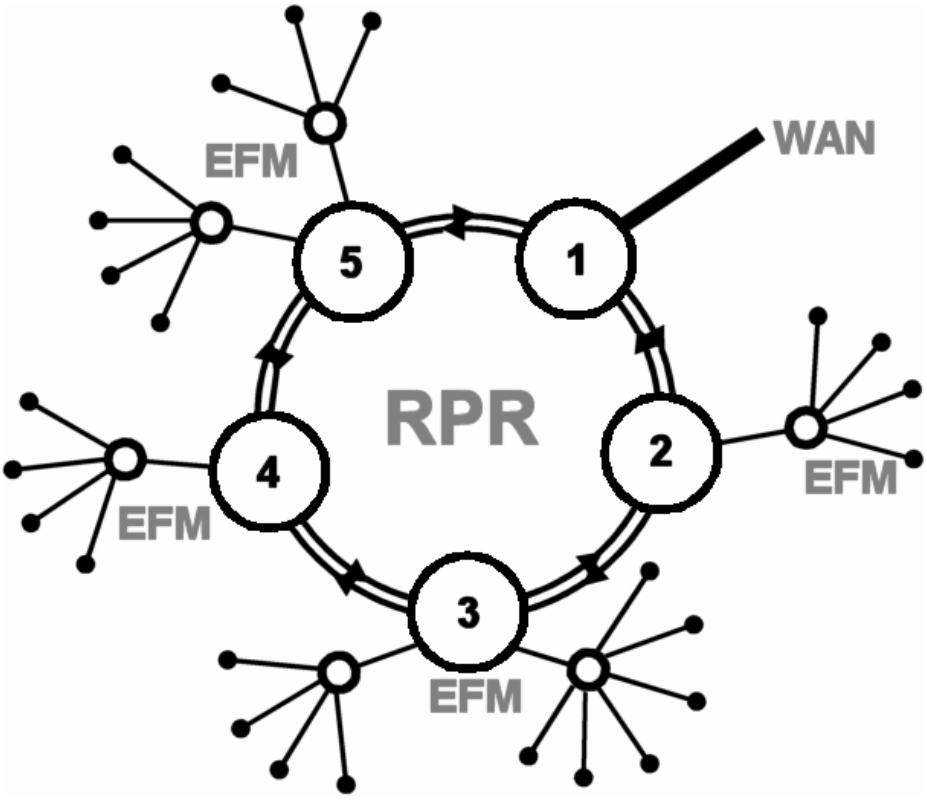

Figure 8. Example RPR-EFM network.

VIII. FRAME FORMAT

It may seem that the packet format of a protocol is a very simple component, but it allows easier understanding of the protocol and provides clues on how the protocol is implemented. Data, fairness, control and idle frames are the four different frame formats defined in the RPR standard. In the following subsections, the four types of RPR frames and their important fields are introduced.

A. Data Frames

Data frames have two formats, which are basic and extended. Extended frame format is aimed at transparent bridging applications allowing an easy way of egress processing and ingress encapsulation of other medium access control (MAC) frames. Using extended frame format also enables RPR rings to eliminate out of ordering and duplication of bridged packets. The Extended frame format is not described in this article.

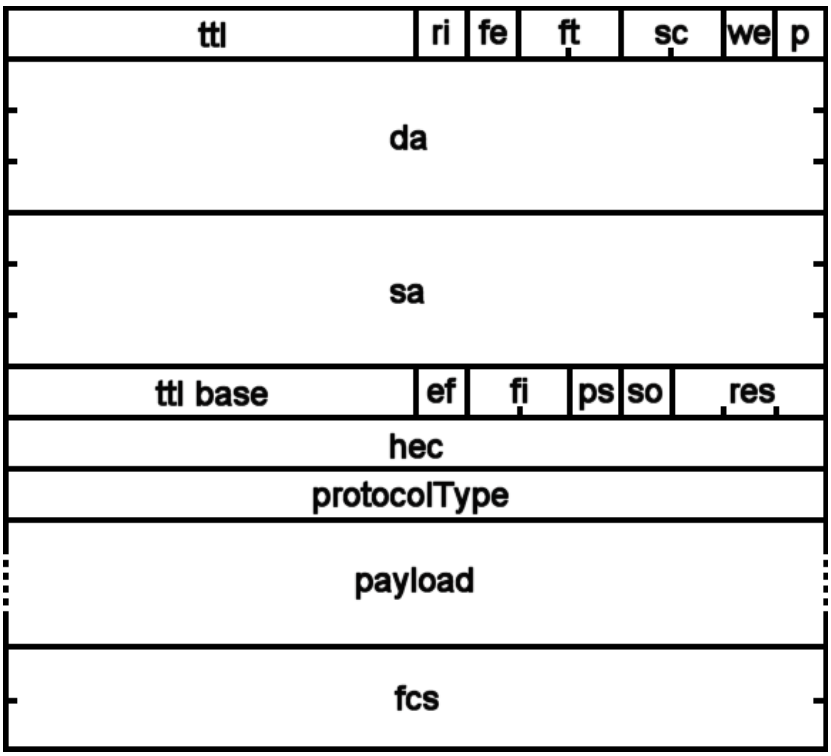

Figure 9. RPR basic data frame format.

Following is a short summary of the functionalities of RPR basic data frame fields:
ttl: The 8 bit "time to live" field is used for detecting packets that should be removed from the ring.
ri: The "ring identifier" bit defines the ring which the packet belongs to (i.e., which ring the packet was inserted into initially).
fe: The "fairness eligible" bit indicates that the packet has to abide by the rules of the fairness algorithm.
ft: The two-bit "frame type" defines the type of the frame.
sc: The two bit "service class" field indicates the class of service (A0, A1, B, C) that the frame is expected to receive.
we: The "wrap eligible" bit defines if the frame can be wrapped at a wrap node.

p: The "parity" bit is reserved for future use in data frames, since the header is being protected by the "*hec*" field.
da: The six-byte "destination address" field defines the destination of a frame.
sa: The six-byte "source address" field shows from which node the frame was sent.
ttl base: This is a fixed field which is set to the initial value of the "*ttl*" field when the packet was initially sourced into the ring. It is used for fast calculation of the number of hops that a packet has traveled.
ef: The "extended frame" bit is set when the frame is an extended frame format.
fi: The two bit "flooding indication" is used to identify if a frame was flooded or not. If the frame was flooded, it also indicates the type of flooding, if the frame is sent on a single ring or on both rings.
ps: The "passed sourced" bit is used as an indication that a frame has passed its source on the opposing ring after a wrap condition. This is useful for stripping packets from the ring under some error conditions.
so: The "strict order" bit enforces the correct order of receive under certain conditions, if there is a chance of receiving a frame with "so" bit set, the frame will be discarded at the receiving station.
res: This is a three-bit reserved field in the header for future expansion.
hec: The two byte "header error correction" field protects the initial 16 bytes of the header.

B. Fairness Frames

The 16-byte fairness frame is a compact control message. It is the communication mechanism of RPR fairness algorithm. The feedback data is the "fairRate" field. The "ffType" (fairness frame format type) field identifies the type of the fairness frame, which can be single-choke or multi-choke message type.

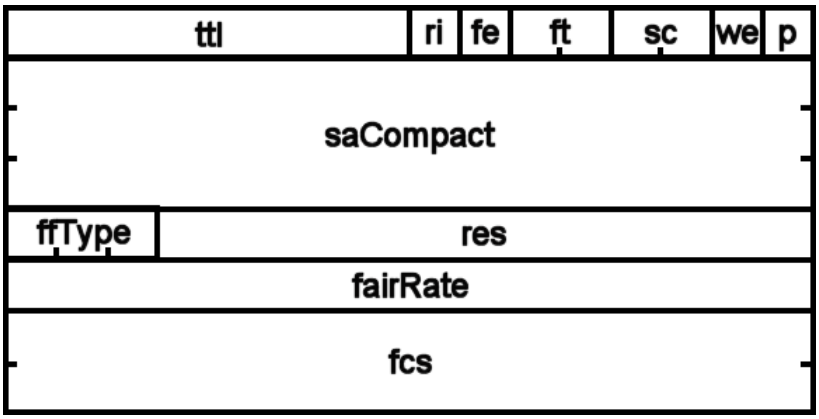

Figure 10. RPR fairness frame format.

C. Control Frames

A control frame has format similar to the data frame but is distinguished by a designated "ft" field value and the type of information carried is specified by its controlType fields. There are different types of control frames in RPR, for example, topology and protection protocol information and OAM (Operations Administration and Maintenance).

D. *Idle Frames*

Idles frames allow strict control of bandwidth allocation around the ring. The frame format is similar to the fairness frame. The “fairnessHeader” and “fairRate” fields are replaced by the four byte “idlePayload”.

IX. CONCLUSION

This paper has discussed and explained the RPR architecture. It has showed how RPR has taken features from earlier ring based protocols, and combined them into a novel and coherent architecture. Important parts that have been covered in this paper include the class based priority scheme, the station design and the fairness algorithm. Performance evaluations using the latest version of the draft standard demonstrate how the protocol behaves using different options. In particular we have demonstrated how the aggressive fairness method is very responsive to change, while the conservative method has a more dampened response under varying load.

RPR is a new MAC-layer technology that may span into the MANs and WANs. RPR can bridged to access networks like EFM, making it possible to perform layer 2 switching far into the backbone network if such large link layer networks turn out to be practical. RPR may also do switching in the backbone network, by letting an RPR ring implement virtual point-to-point links between the routers connected to the stations on the ring.

RPR may differentiate traffic, so when used to implement IP links, it is able to help the IP routers implement the QoS aware communication that is needed in a network that carry multimedia traffic.

REFERENCES

[1] ANSI T1.105.01-2000: Synchronous Optical Network (SONET)- Automatic Protection.

[2] I. Cidon, Y.Ofek, “MetaRing - A Full-Duplex Ring with Fairness and Spatial Reuse”, IEEE Trans on Communications, Vol. 41, No. 1, January 1993.

[3] I. Cidon, L. Georgiadis, R. Guerin, Y. Shavitt: Improved fairness algorithms for rings with spatial reuse. INFOCOM '94. Networking for Global Communications. IEEE, 1994

[4] V. Gambiroza, Y. Liu, P. Yuan, and E. Knightly, “High Performance Fair Bandwidth Allocation for Resilient Packet Rings.”, In Proceedings of the 15th ITC Specialist Seminar on Traffic Engineering and Traffic Management, Wurzburg, Germany, July 2002

[5] E.R. Hafner, Z. Nendal, M. Tschanz, “A Digital Loop Communication System.” , IEEE Transactions on Communications, Volume: 22, Issue: 6, June 1974.

[6] IEEE Draft P802.17, draft 3.0, Resilient Packet Ring, November 2003.

[7] IEEE Standard 802.5–1989, “IEEE standard for token ring”.

[8] IEEE Standard 1596–1990, “IEEE standard for a Scalable Coherent Interface (SCI)”

[9] I. Kessler, A. Krishna: On the cost of fairness in ring networks. IEEE/ACM Transactions on Networking, Vol. 1 No. 3, June 1993

[10] R.M. Needham, A.J. Herbert: The Cambridge Distributed Computing System. Addison-Wesley, London, 1982.

[11] D. Picker, R.D. Fellman: Enhancing SCI’s fairness protocol for increased throughput. IEEE Int. Conf. On Network Protocols. October, 1993.

[12] 1.F.E. Ross, “Overview of FDDI: The Fiber Distributed Data Interface”, IEEE J. on Selected Areas in Communications, Vol. 7, No. 7, September 1989.

[13] ISO/IECJTC1SC6 N7873, “Specification of the ATMR Protocol (V. 2.0)”, January 1993.
[14] W.W. Lemppenau, H.R.van As, H.R.Schindler, “ Prototyping a 2.4 Gbit/s CRMA-II Dual-Ring ATM LAN and MAN”, Proceedings of the 6th IEEE Workshop on Local and Metropolitan Area Networks, 1993.
[15] OPNET Modeler. http://www.opnet.com.
[16] Nortel Networks OPTera Technology, http://www.nortelnetworks.com/products/family/optera.html
[17] Cecil C. Reames and Ming T. Liu, “A Loop Network for Simultaneous Transmission of Variable-length Messages”,.ACM SIGARCH Computer Architecture News , Proceedings of the 2nd annual symposium on Computer architecture, Volume 3 Issue 4, December 1974.Cecil C. Reames and Ming T. Liu, “A Loop Network for Simultaneous Transmission of Variable-length Messages”,.ACM SIGARCH Computer Architecture News , Proceedings of the 2nd annual symposium on Computer architecture, Volume 3 Issue 4, December 1974.
[18] . J.H. Schuringa, G. Remsak, H.R. van As, A. Lila, “Cyclic Queueing Multiple Access (CQMA) for RPR Networks”, 7th European Conference on Networks, & Optical Communications (NOC2002), Darmstadt, Germany, pages 285-292, June 18-21, 2002.
[19] D. Tsiang, G. Suwala, “The Cisco SRP MAC Layer Protocol”, IETF Networking Group, RFC 2892, Aug. 2000
[20] Cisco DPT Technology, http://www.cisco.com/warp/public/779/servpro/solutions/optical/dpt.html