RPR Bandwidth Management

and

Fairness

Necdet Uzun

Harry Peng

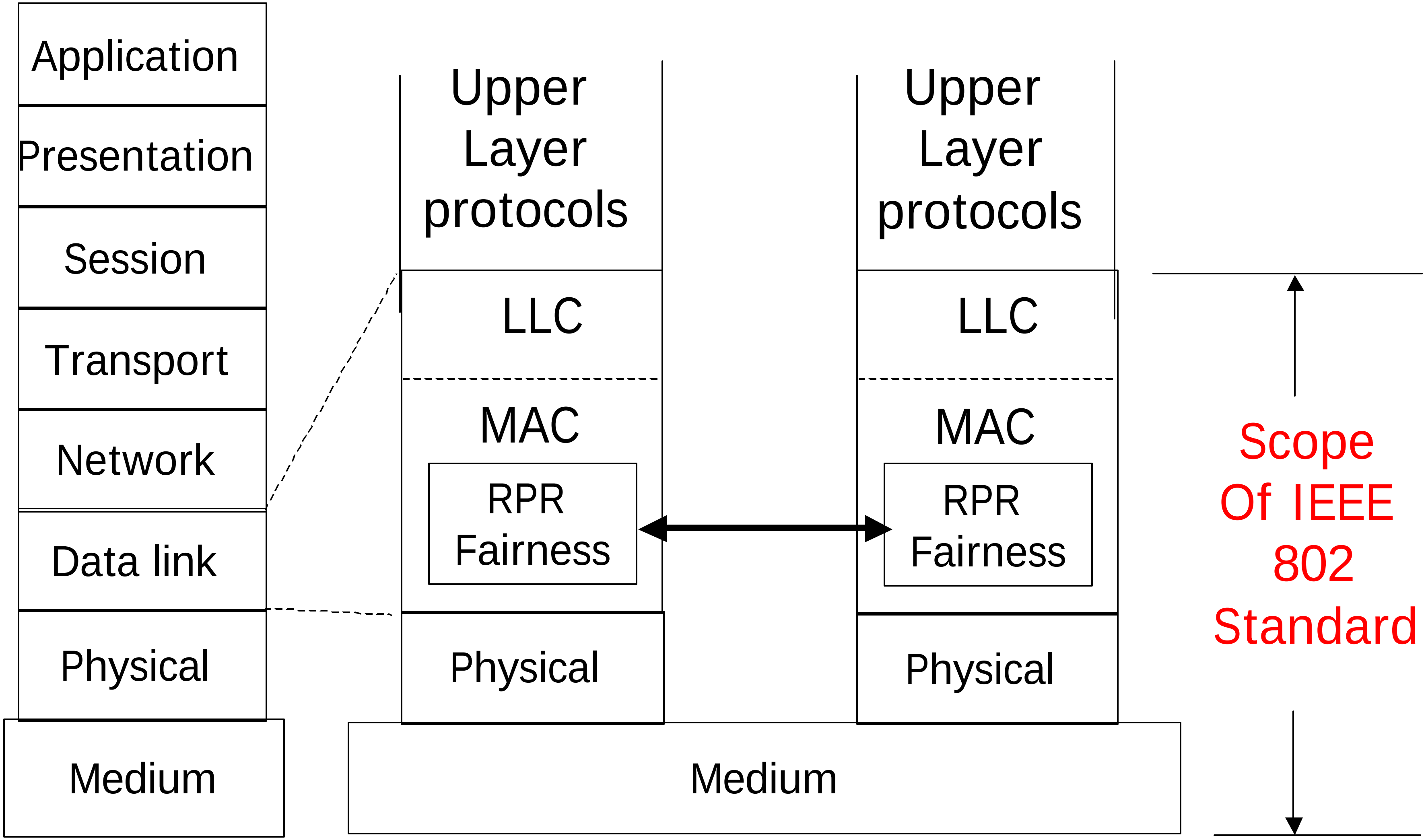
Agenda

- Requirements
- Node model
- Fairness Algorithm
  - Weighted
  - BW reservation
  - 3 Priority Support
  - VDQ Support
- Fairness Message Handling
- Conclusions

Requirements

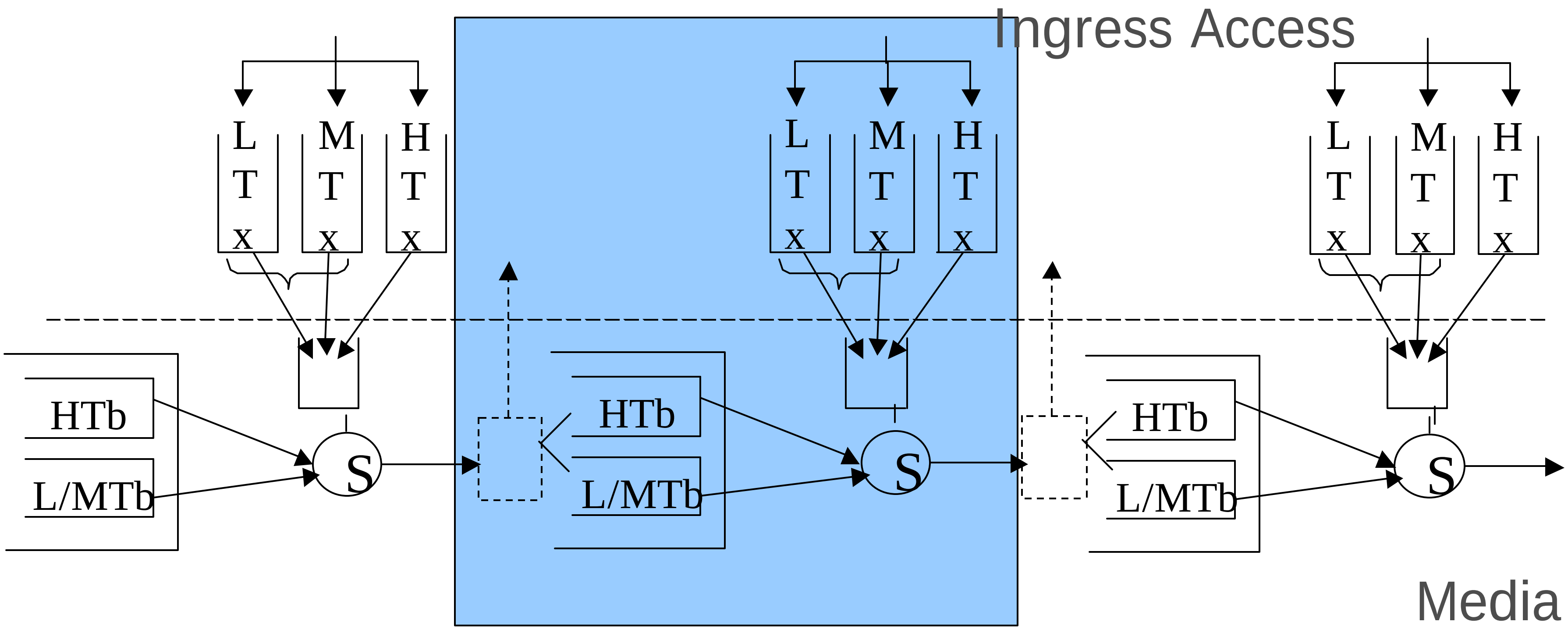
- Target MAN application (LAN/MAN/RAN/WAN)
  - Ring size: MAN: circumference < 1000km, stations<128
  - SLA: BW, delay, loss, jitter
- Shared Media Access
  - Unlike 802.3 CSMA/CD where collision is detected by all stations on the wire
  - Spatial Reuse
- Three priority support
- HP bandwidth reservation
- Weighted Fairness
  - Each node has an assigned weight
  - Advertise fair_rate value scaled by weight

Reference Model



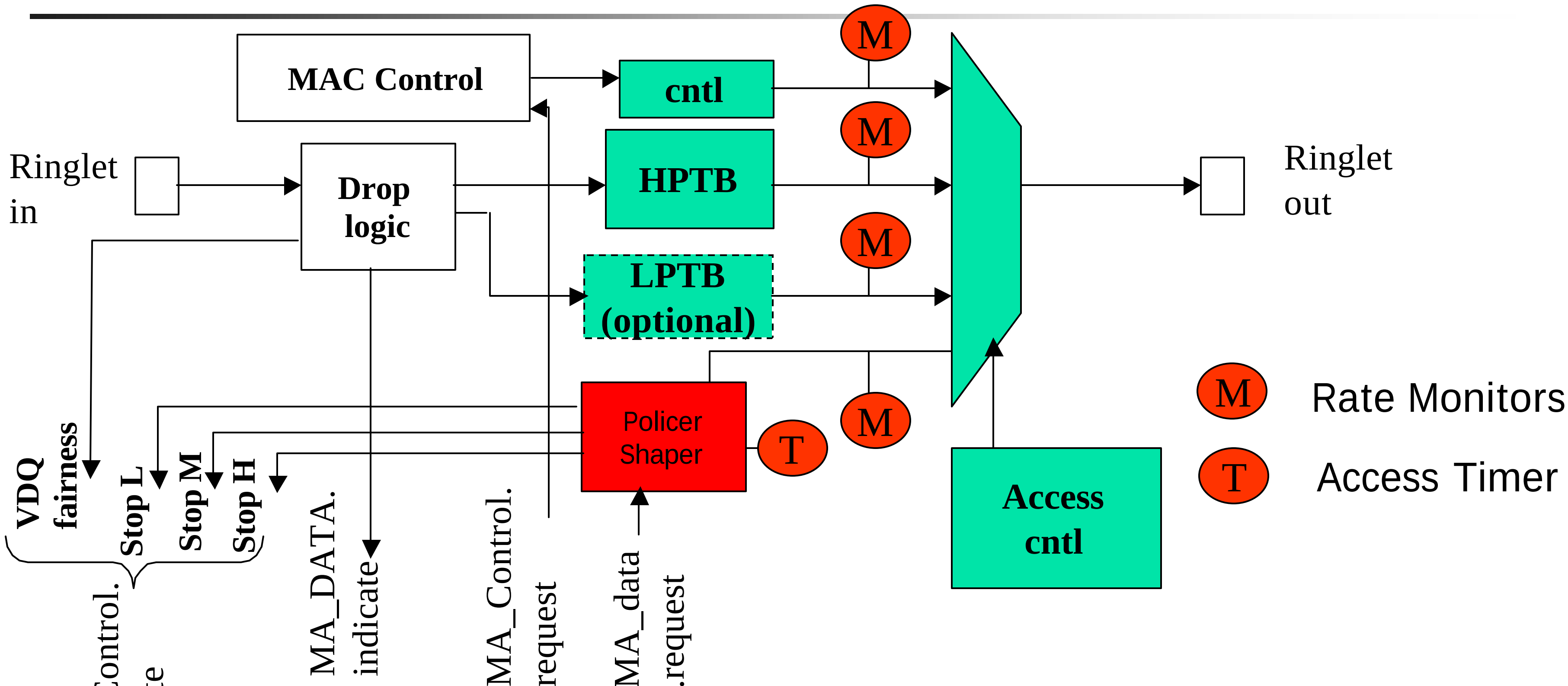
- RPR-fa is a MAC peer to peer function

Ring Fairness Model



- MAC peer to peer function
- Transit path is an extension of the physical medium
- Stations are connected in series
  - Ingress and transit
- Fairness Messages

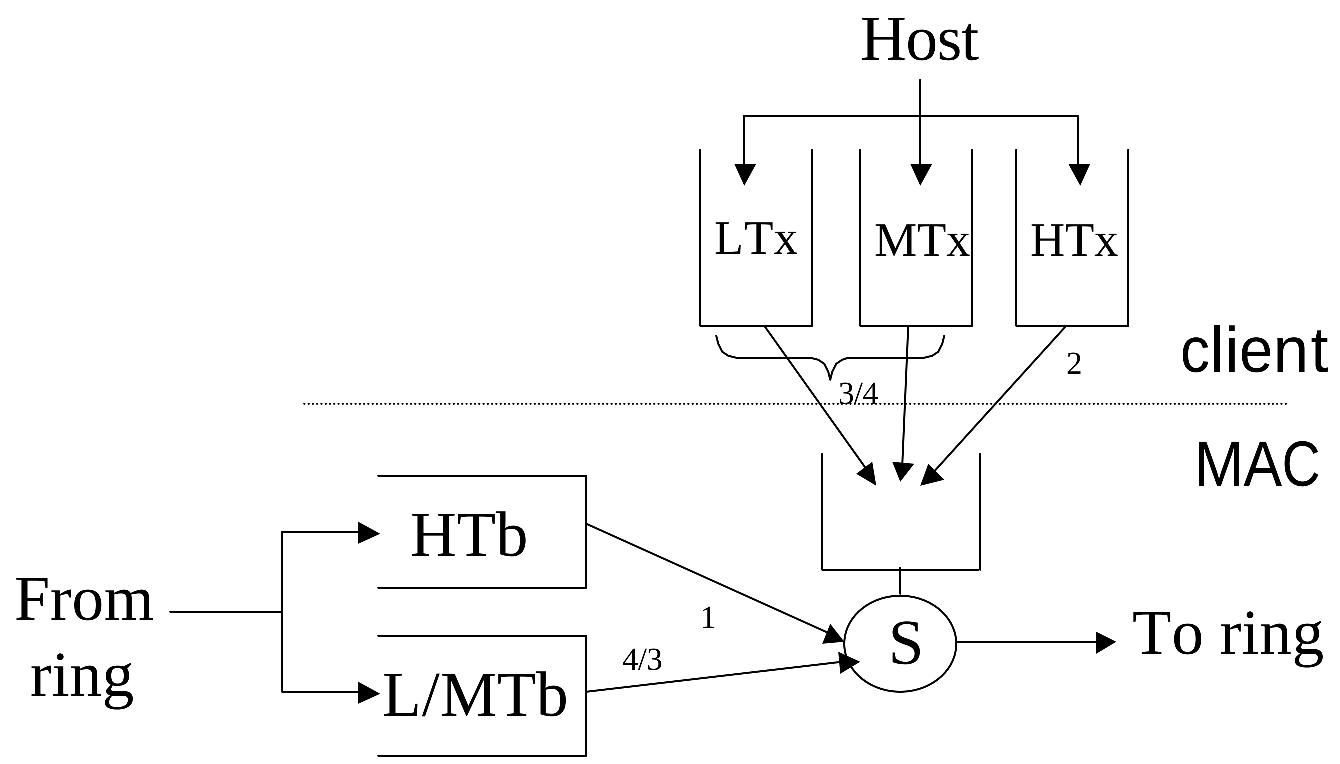
Fairness: Implementation Model



- HPTB is required, LPTB is optional
- Transit, add, output, rate monitors
- Control insert fairness, Protection, and topology messages
- Access Control: Ring egress scheduler: simple priority
  - Strict priority for small transit buffer design
  - Conditional forwarding rules when LPTB exists
- Access delay timer

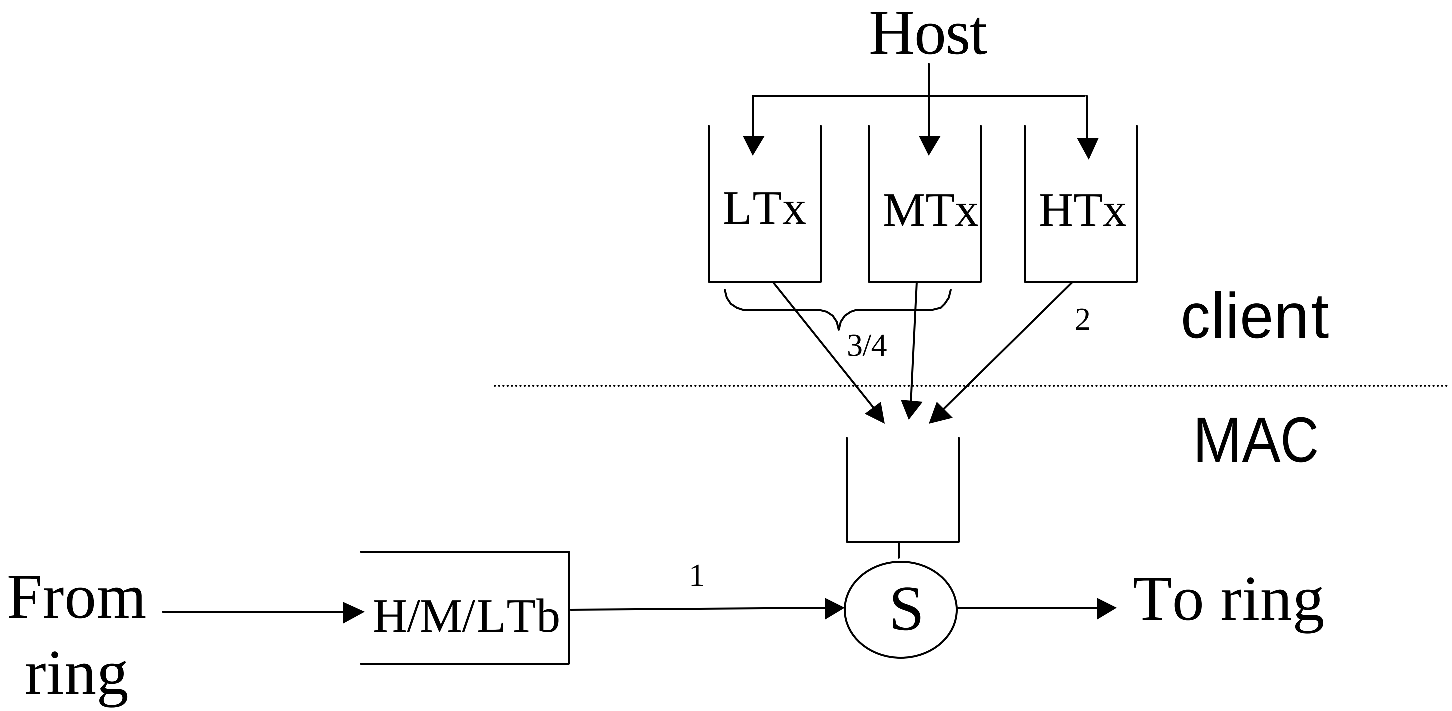
Node Model 2TB

- Two transit buffers
- Three transmit buffers
  - 3 token bucket counter for HP, cMP, eMP+LP



Node Model 1TB

- Single transit buffer
- Three transmit buffers
  - 3 token bucket counter for HP, cMP, eMP+LP



Fairness Algorithm

- 3 Entities/state machines
  1. Congestion Machine: detects when fa is required
  2. Rate Advertisement machine: generate and advertised rate
  3. Rate Conformance machine: apply advertised message to ingress
- Station in congested point controls (advertises) a rate for the upstream nodes to conform to once it enters congestion
  - entering and exiting congestion could have hysterisis
  - one set of knobs to control entry and exit of congestion
- a station is in “congestion” determines a rate which is advertised
  - the rate needs to be modified as the node moves around the congestion zone
  - possibilities are: up or down or hold
- upstream nodes need to respond to the fa messages
  - always follow the rate the station receives unless
  - when a NULL message is received (go to full rate) the ramp up rate is configurable

3 Priority Support

- Provide 3 priority classes on the ring
- High Priority
  - Guaranteed bandwidth (provisioned)
  - Bounded delay and bounded jitter
- Medium Priority
  - Committed Access Rate (CAR) for MP (cMP)
  - MP Traffic exceeding CAR (eMP) is subject to fairness algorithm control in the transmit path
  - Committed bandwidth (provisioned), best effort for excess traffic
  - Bounded delay and (loosely) bounded jitter
- Low Priority
  - No guarantees
  - Best effort for bandwidth, delay and jitter

Bandwidth Reservation:

- Optionally a certain amount of bandwidth on each span can be reserved
  - For use of HP or guaranteed traffic
  - This bandwidth can not be reclaimed by fairness algorithm (it is wasted if not used)
- Reserving bandwidth on a span is simple
  - limit forward rate + add rate of MP+LP to

$$C - \sum r_i$$

Virtual Destination Queues

- Supported to provide certain network applications

- Multiple node congestion information for Virtual Destination Queuing (VDQ)
  - Use of more detailed choke (congested) point information in the client provides better utilization of network resources
  - A scheduling policy in the client may utilize multi-choke information
- Choke point and corresponding fair_rate information is passed to MAC client and MAC client does the scheduling of VDQ's.
  - Upon reception of fair_rate info, client updates allow_rate info for the appropriate choke point.
  - Client can keep up to N number of choke points.
  - Clients limit the amount of insertion traffic sent through each choke points to appropriate allow_rate

Message Format

Field	bits
Ring Header (type =0x06)	16
SA	48
Ver | Length | Reserved	16
Control Value	16
FCS32	32

- Header as in standard frame
  - TTL, TYPE, RI, PRI, IOP
- Fairness Control Header
  - Version = 3 bits.
  - Length= 8bits.
    - Optional not all nodes have to support it
    - Topology will be used synchronize version defaults to the lowest
    - RESERVED = 4 bits
- Control Value provides the rate information for the fairness algorithm
- Packet integrity FCS protected

- Optimized for MAC peer to peer messaging
  - Reduces control BW requirement

Message Format (Cont'd)

- Type 1 fairness messages are generated in every fairness message interval and passed hop by hop
  - Type 1 fairness messages can not cross fairness domain boundaries (**isolation of congestion / fairness domains**)
  - Fair_rate is processed by each MAC and passed to VDQ MAC client
  - A new fair_rate is determined by intermediate MAC and either originators SA or the current node's SA is used depending on whichever is more congested is sent to upstream
- Type 2 messages are generated by each MAC in every N=10 fairness message intervals and may be broadcast hop by hop
  - Fair_rate is passed to each MAC client along the way and stripped by the source
  - Used by VDQ Clients only

Conclusions

- RPR-fa fairness algorithm is simple
  - No per source information is needed in fair rate calculation
- RPR-fa algorithm works with both single and dual transit buffers
- RPR-fa supports up to 3 Transport Priority classes and
- RPR-fa supports weighted fairness algorithms
- RPR-fa supports efficient VDQ implementations
  - RPR-fa polices traffic based on most congested fairness domain that
  - No per destination policing is needed

Backup

Fairness Algorithm Detail

Motivation:

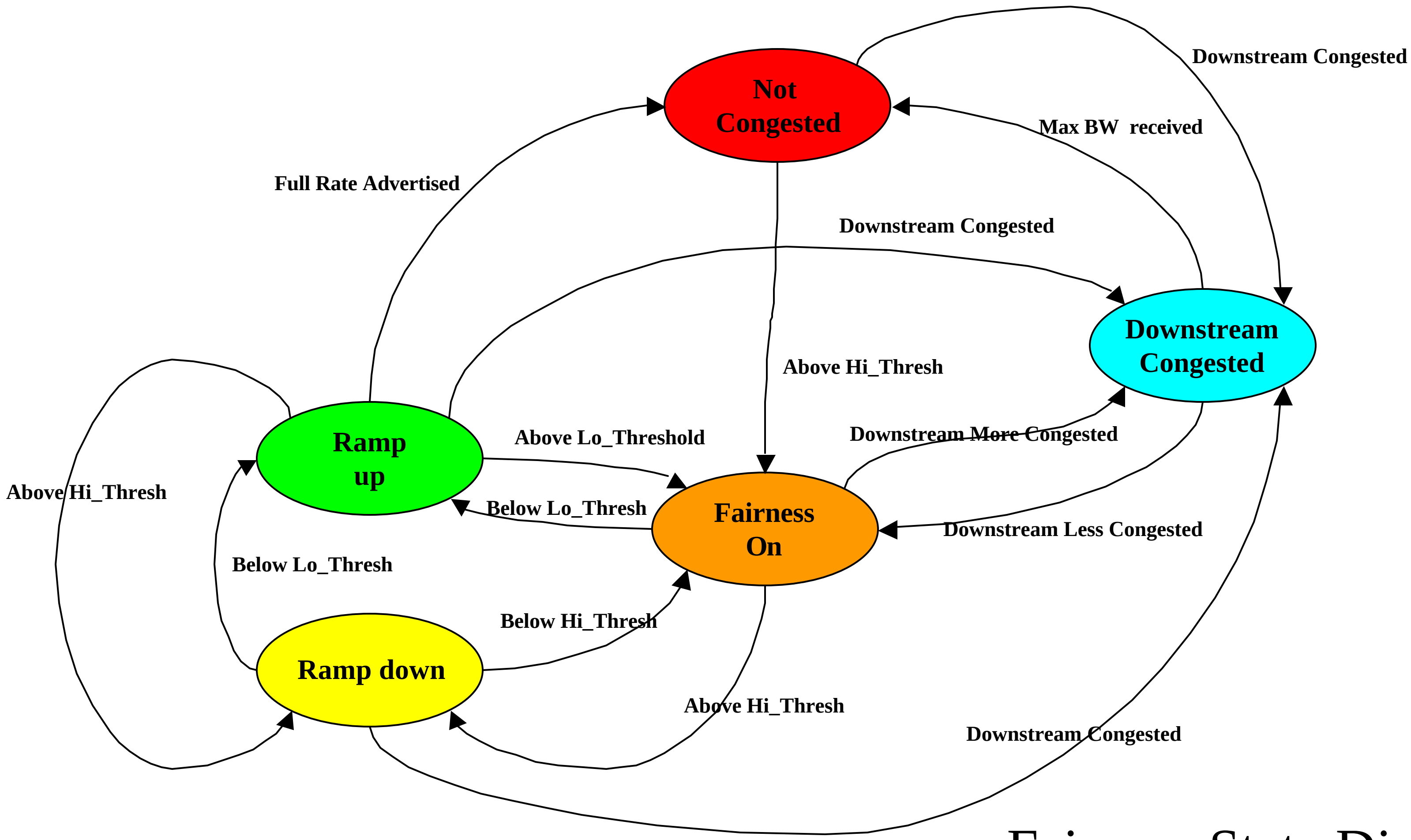
- 50,000 ft view is that the control of the congestion domain can be done with a basic algorithm with a few knobs. Complements requirements in many applications.
  - 2 sets of controls at the congestion point
  - 1 set of controls at the upstream nodes
- allows different behaviors to be configured and the algorithm adjusted to things like rings size, number of nodes, applications, performance.
- has more degrees of control than Gandalf or Aladdin

Fairness Algorithm Detail (Cont'd)

Three state machines per MAC ringlet

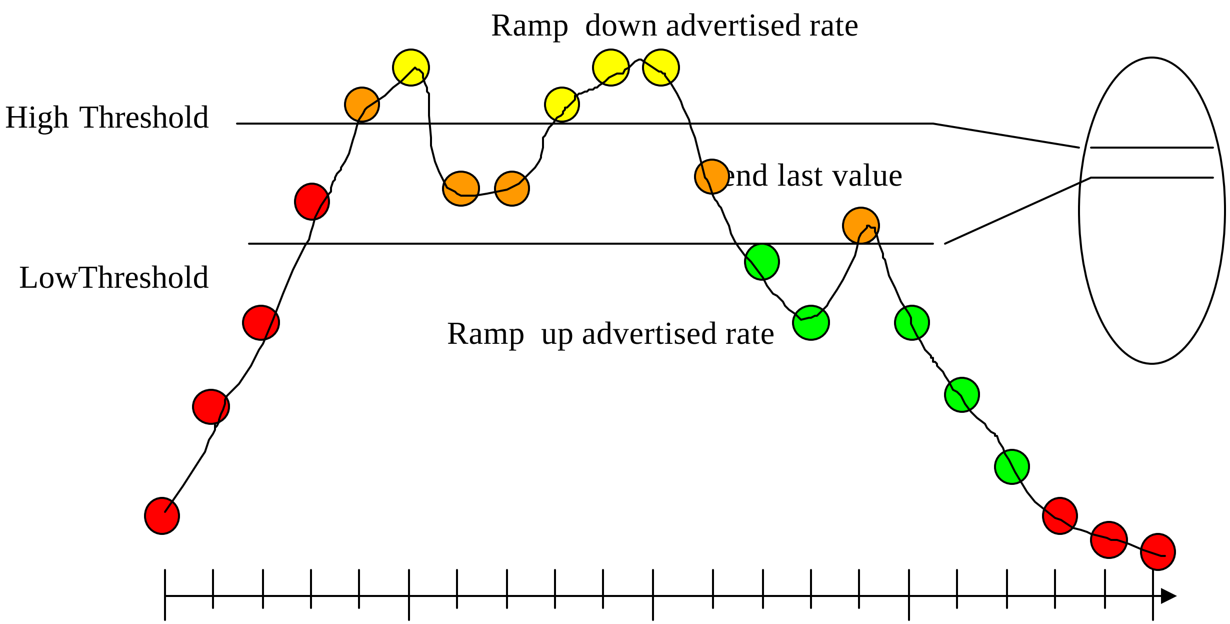
- Congestion machine – determine whether ringlet is congested and how to adjust the advertised rate
  - HI_THRESH is the threshold that marks the entry into congestions, it is not the same as the LPTB HI_THRESHOLD.
  - LO_THRESH is the threshold that marks the exit of congestion, and may be the same value as HI_THRESH (no hysteresis) or smaller
- Rate Advertisement machine – adjust the advertised rate
- Rate Conformance machine – determine what rate we are allowed to transmit when receiving upstream fairness messages

Congestion Machine



Fairness State Diagram

- This chart shows the use of hysteresis in a single TB node



- Utilization measured and compares with threshold
- Two thresholds provide hysteresis for stability
  - Utilization crosses high threshold ramp down advertised rate to decrease upstream station usage
  - Utilization crosses low threshold ramp up advertised rate to increase upstream station usage

Rate Advertisement Machine

- Initial Advertised Value
  - advertised_rate[t] = ((J-1)*advertised_rate[t-1] + add_rate) / J;
  - optional case: (Max_Lrate -HP_Reserved) / Ns;
    - where Ns is the number of nodes sending
- Ramp Up value
  - advertised_rate[t] = advertised_rate[t-1] * (1+ (K-1)/L)
- Ramp Down value
  - advertised_rate[t] = advertised_rate[t-1] * (1–(K-1)/L) + Filter_Fn()
  - Filter_Fn = (add_rate - advertised_rate[t-1]) * (I-1)/J

Rate Conformance machine

- Stop_hi = 1

(no hi tokens) **or**

LPTB is near full

- Stop_med = 1

no med tokens

Or LPTB is near full

- Stop_lo = 1

(no lo tokens) **or**

(Congestion_Threshold is exceeded) **or**

(allow rate is exceeded) **or**

((fwd rate < add rate)

and (LPTB not empty))

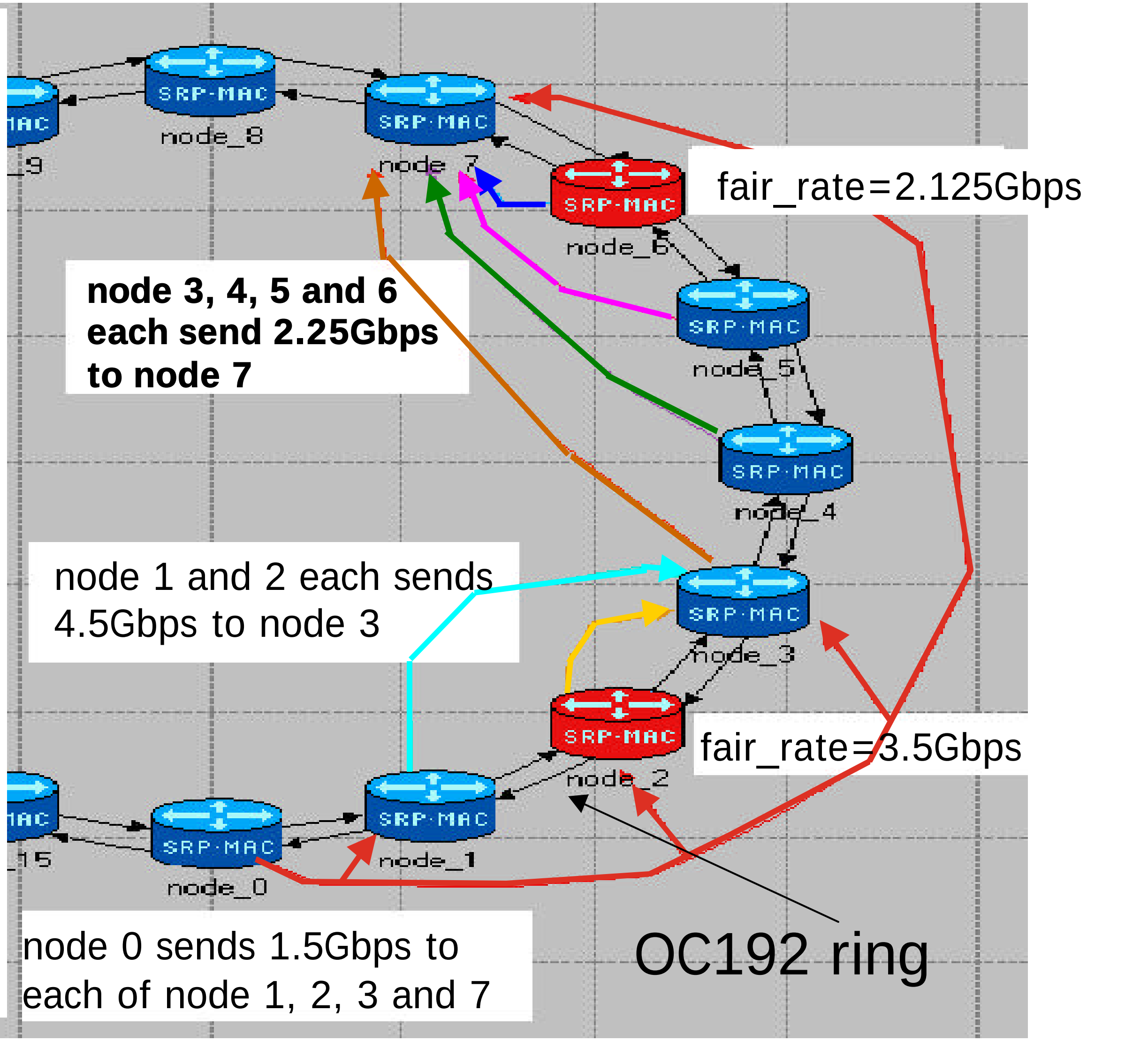
Multiple Congestion Domains

Node 3 to 6 are in the 1st congestion domain

Node 0 to 2 are in the second congestion domains

Type 1 fairness messages from domain 1 should not be propagated to domain 2 by node 3 (fairness domain isolation)

As node 0 to node 7 traffic increases to 2Gbps, 2 fairness domains collapse



Congestion Domains

- Node 0 (if VDQ) is aware of 3 congestion domains:
  - 3rd fairness domain: node 0, node 1 and node 2
  - 2st fairness domain : nodes between node 3 and node 6 (inclusive)
  - 1st fairness domain: nodes beyond node 6
- Node 0 (if simple client) is aware of 2 congestion domains:
  - Before congestion domains collapse:
    - 2nd fairness domain: node 0, node 1 and node 2
    - 1st fairness domain: nodes beyond node 3
  - After congestion domains collapse:
    - 2nd fairness domain: node 0 to node 6 (inclusive)
    - 1st fairness domain: nodes beyond node 6

Congestion Domains

- Node 0 (if VDQ) is aware of 3 congestion domains:
  - 3nd fairness domain: node 0, node 1 and node 2
  - 2st fairness domain : nodes between node 3 and node 6 (inclusive)
  - 1st fairness domain: nodes beyond node 6
- Node 0 (if simple client) is aware of 2 congestion domains:
  - Before congestion domains collapse:
    - 2nd fairness domain: node 0, node 1 and node 2
    - 1st fairness domain: nodes beyond node 3
  - After congestion domains collapse:
    - 2nd fairness domain: node 0 to node 6 (inclusive)
    - 1st fairness domain: nodes beyond node 6

VDQ Details, Cont.

- Node 0 should obey the following constraints while scheduling its virtual destination queues:
  - Up to line rate for traffic destined to node 1 and node 2.
  - Virtual destination queues for nodes 3,4,5, and 6 can be scheduled as long as the total usage beyond VDQ_2 does not exceed $fair\_rate_2$.
  - Virtual destination queues for nodes beyond 6 can be scheduled as long as the total usage beyond VDQ_2 does not exceed $fair\_rate_2$ and the total usage beyond VDQ_6 does not exceed $fair\_rate_6$.