Bandwidth Management

and

Fairness

Necdet Uzun

11/13/2001

Agenda

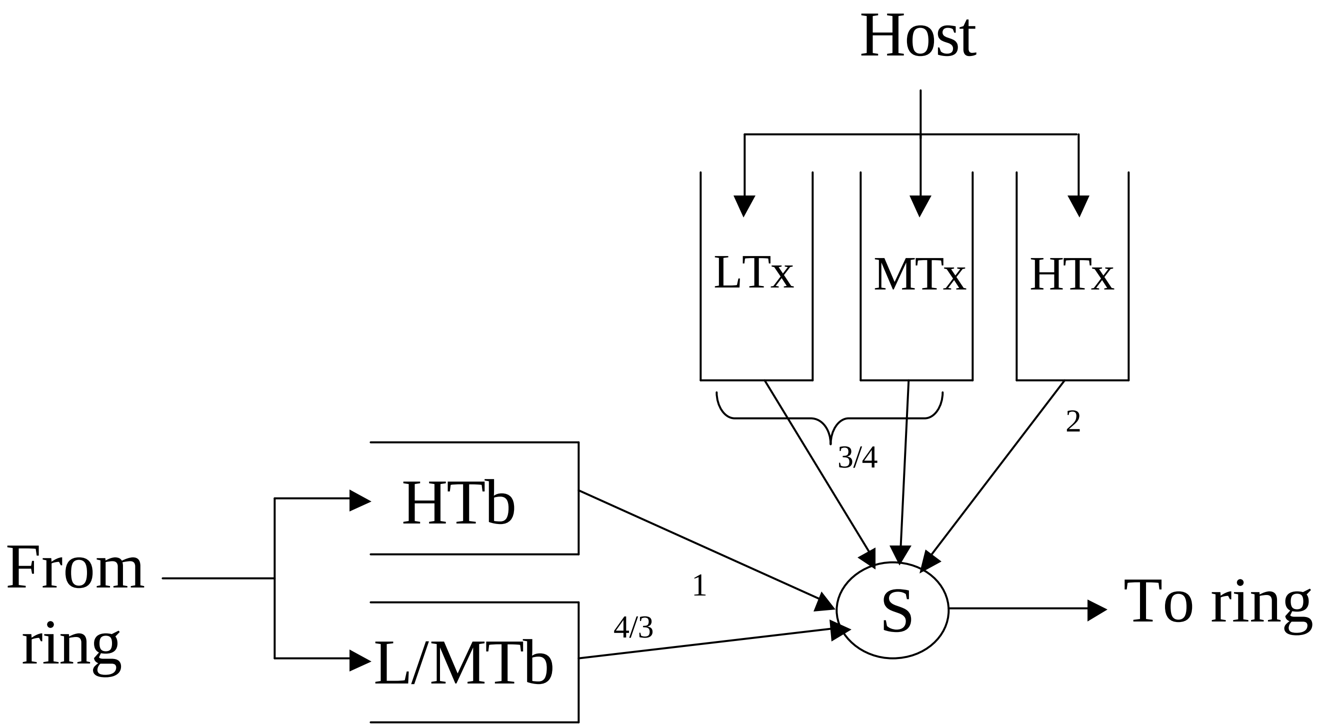
- Introduction
- Requirements
- Node model
- Weighted Fairness Algorithm
  - BW reservation
  - 3 Priority Support
  - VDQ Support
  - Fairness Message Handling
- Comparison
- Conclusion

Requirements

- Weighted Fairness
  - Each node has an assigned weight (1 to 32)
  - Advertise fair_rate value scaled by weight
  - Only shifts and adds are needed
- HP bandwidth reservation
- Three priority support
- 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

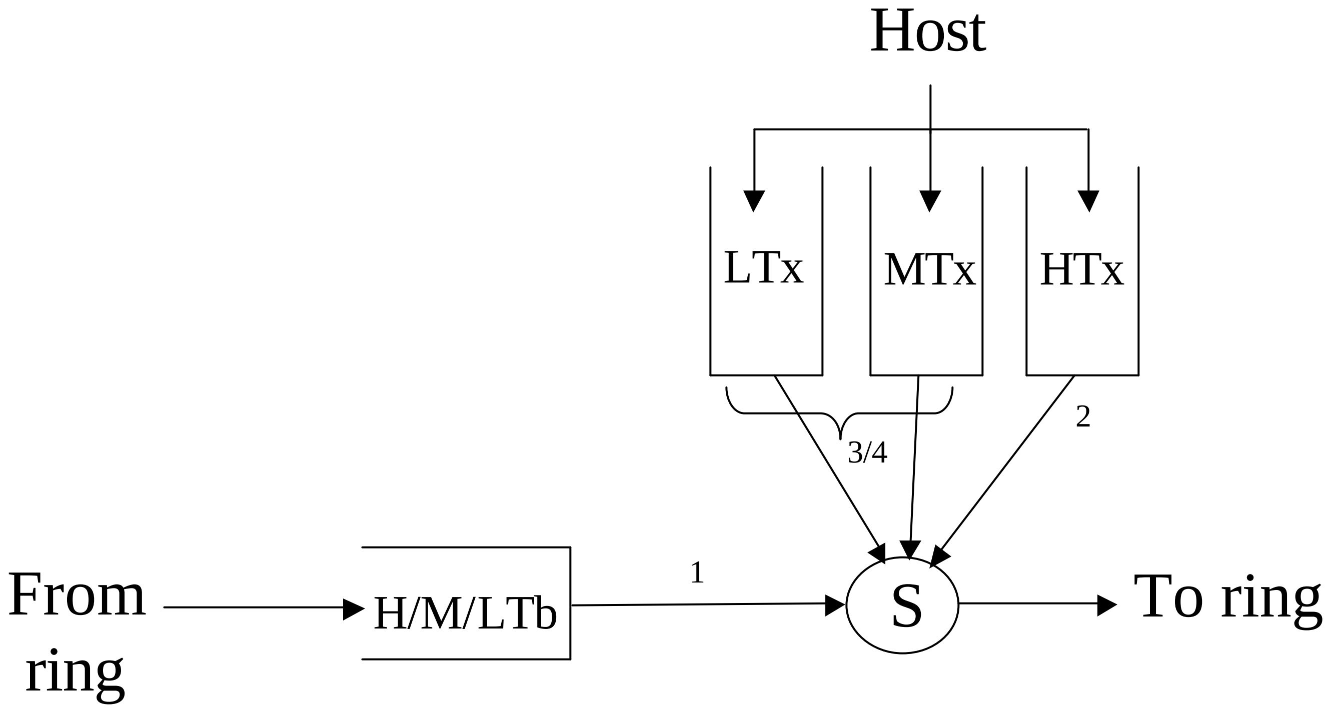
Node Model 2TB

- 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



Bandwidth reservation

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

$$C - \sum r_i$$

3 Priority Support

- Provide 3 priority classes in 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

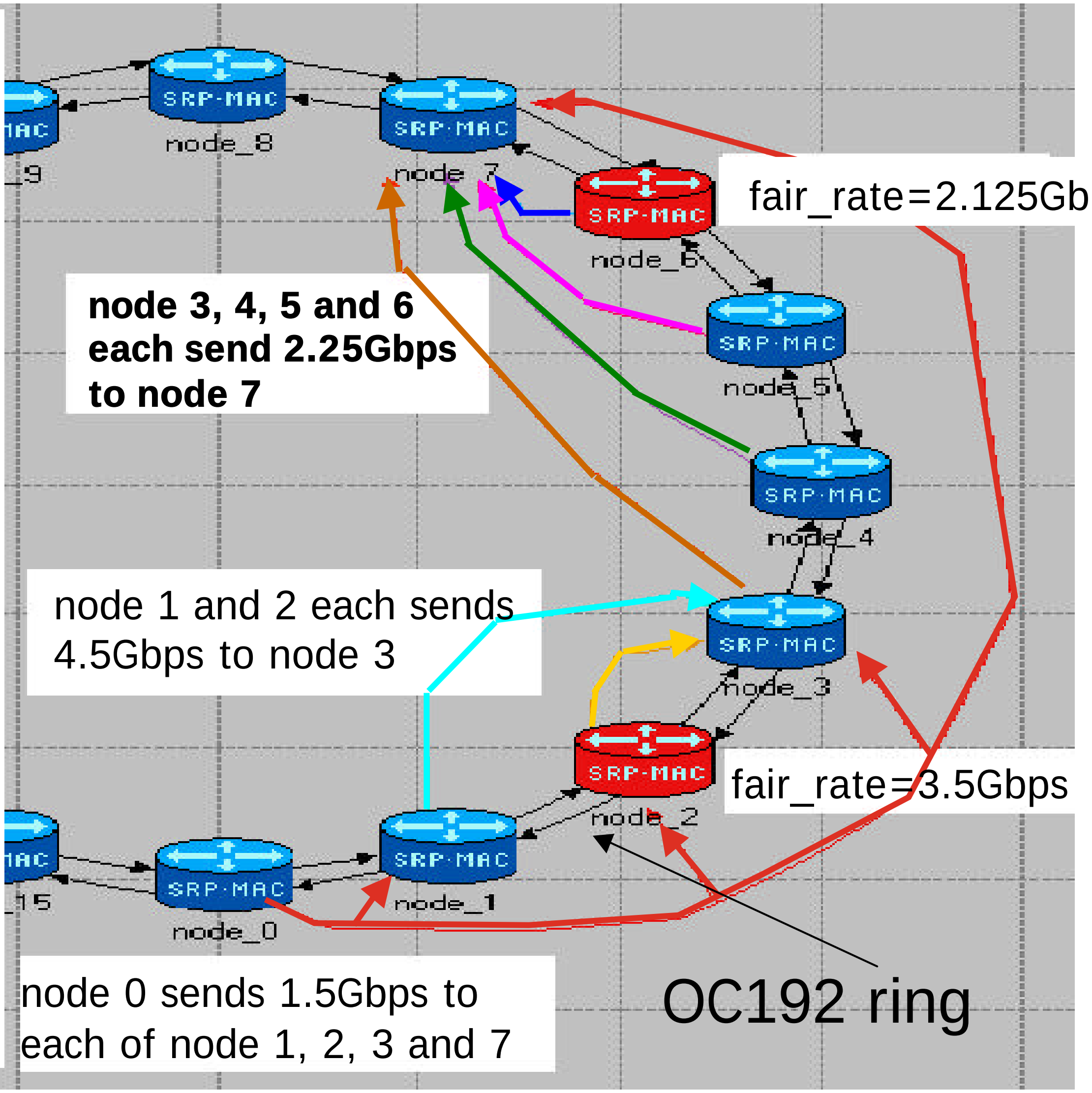
Multiple Congestion Domains

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



fair_rate=2.125Gbps

node 3, 4, 5 and 6 each send 2.25Gbps to node 7

node 1 and 2 each sends 4.5Gbps to node 3

fair_rate=3.5Gbps

node 0 sends 1.5Gbps to each of node 1, 2, 3 and 7

OC192 ring

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

- 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.
- This approach may require as many queues as the number of nodes on the ring.
- Clients limit the amount of insertion traffic sent through each choke points to appropriate allow_rate.

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 $fair\_rate_6$.

Fairness Message Handling

- Type 1 fairness messages are generated in every fairness message interval and passed hub by hub
  - 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 10 fairness message intervals and may be broadcast hub by hub
  - Fair_rate is passed to each MAC client along the way and stripped by the source
  - Used by VDQ Clients only

Comparison

	Alladin	**DVJ**	**Gandalf**
Fairness	Global knowledge	Local/global knowledge	Local knowledge
Policing	N choke - Complex	Single - Simple	Single Choke - Simple
Weighted fairness	N weights info in each node - complex	Local weight info only - simple	Local weight info only - simple
HP strict BW reservation	Yes (can't be disabled)	Yes(can be disabled)	Yes (can be disabled)
1 add queue support	Not clear	Yes	Yes
VDQ support	Yes	Yes	Yes
MAC ETE jitter	Round trip delay	2N * MTU	2N * MTU
# of ring priorities	1	2	2
# of TB supported	Single only	Single and dual	Both single and dual
Need for # of active source monitoring	Yes - complex	No/maybe - simple	No - simple
Need for per source traffic monitoring	Yes - complex	No/maybe - simple	No - simple
Unused BW reclamation	Not clear	Yes	Yes
Throughput	91% to 97%	95%-100%	~100%

Conclusion

- RPR fairness algorithm shall/can be simple
  - No per source information is needed in fair rate calculation
- RPR fairness algorithm shall work with both single and dual transit buffers
- 3 Priority classes and weighted fairness algorithms shall be implemented
- RPR MAC shall support VDQ implementations
- RPR MAC shall police traffic based on most congested fairness domain that its client is contenting for
  - No per destination policing is needed