*i*PT

Control Access Protocol (iPT-CAP)

Harry Peng and Allan Pepper

July 11, 2000

iPT Control Access Protocol
Fair Access with QoS

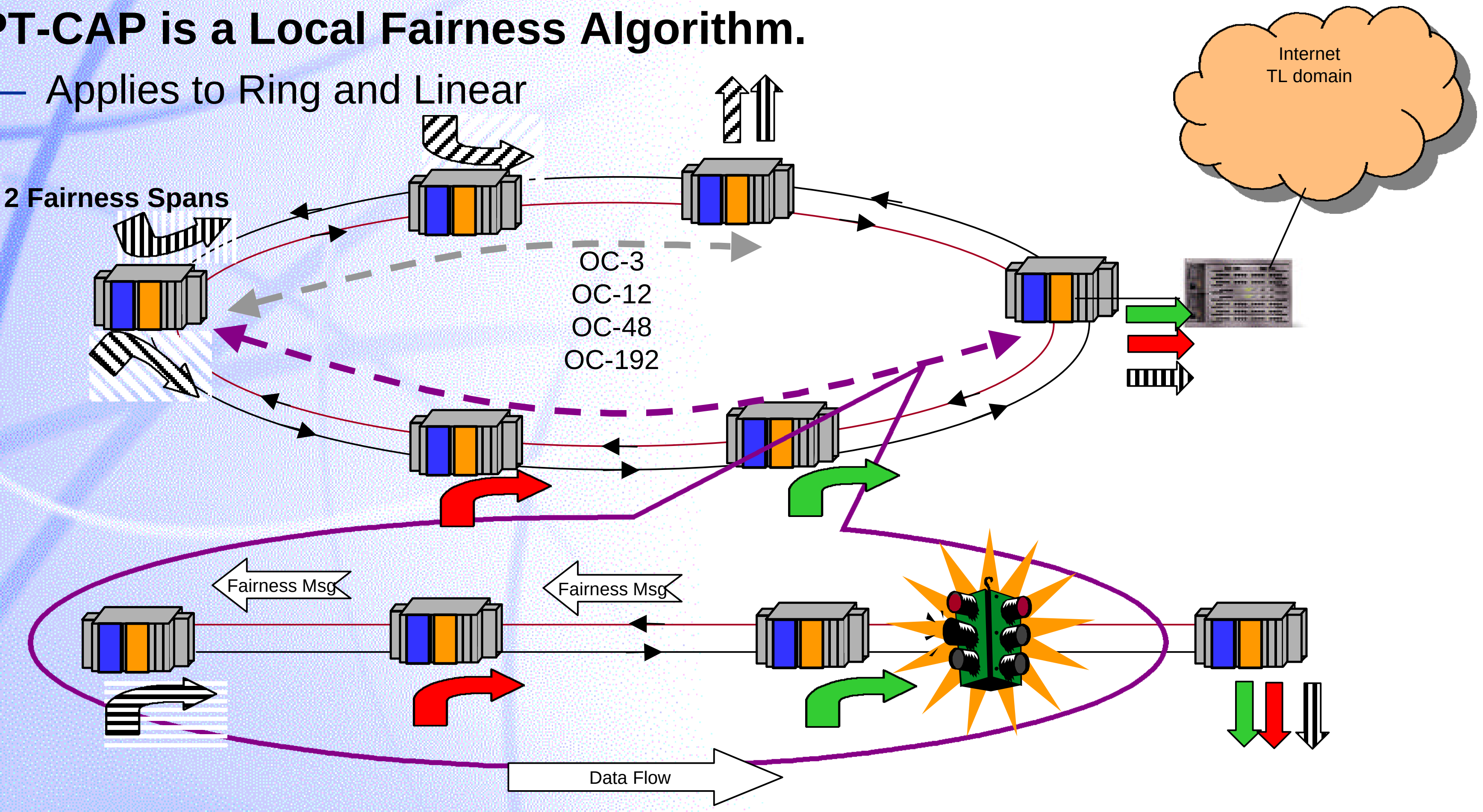
- **Prevents Starvation under Congestion**
- **Provide fair access to shared to WAN BW for same Class traffic**
  - “WAN traffic scheduling”
  - Ingress Queue management
- **Provide QoS for iPT Network**
  - Allows high priority packets to be delivered before low priority packets
  - Provide differential treatment between different packet classes
  - Supports 4 CoS
- **Enabler for over subscribed Networks**

iPT-Control Access Protocol
Efficient, Flexible, and Robust

- **A Backpressure Mechanism**
  - Advertise credits
- **It is a Local Fairness as oppose to a Global Fairness scheme.**
  - Allows Spatial Reuse
  - Responds within Span Round Trip Delay
- **Provides maximum BW availability under fault scenarios (non-wrap)**
- **Fast response and convergence for optimal BW utilization**
  - Event triggered and specific target rate advertising
  - Optimized algorithm triggers on packet delay performance
  - Stable algorithm prevents oscillation. Applies to bursty and steady state traffic patterns.
- **Control messages are designed for flexibility and it's scalable**

iPT-Control Access Protocol
Local Fairness

- **iPT-CAP is a Local Fairness Algorithm.**
  — Applies to Ring and Linear



Local Fairness applies to a Congested Span; Degenerates problem to a Linear Problem

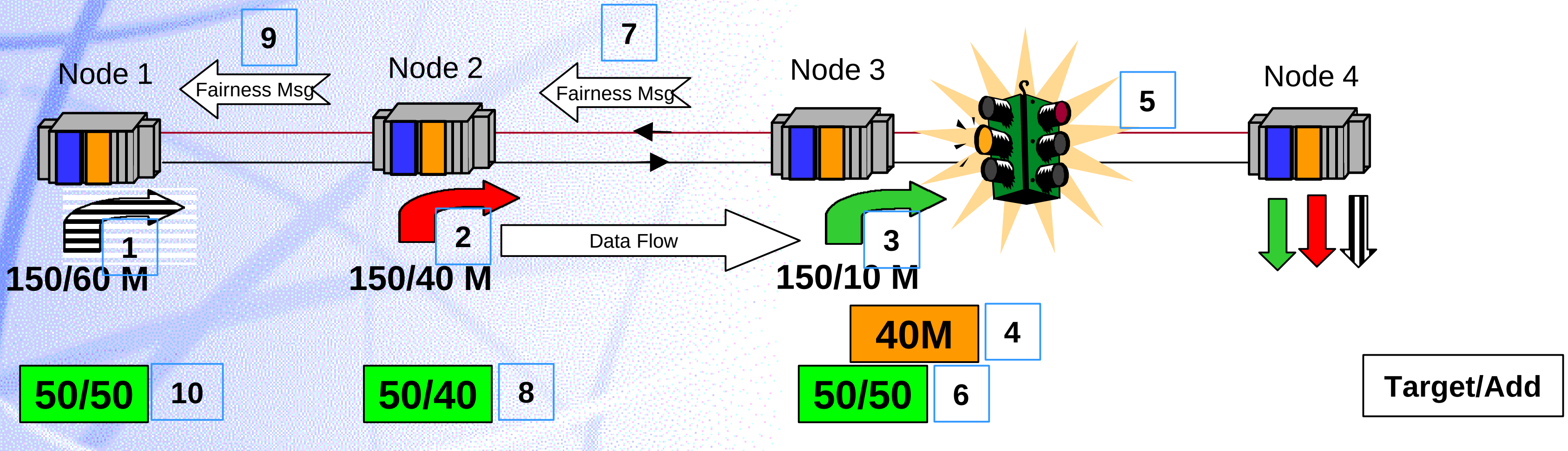
iPT-Control Access Protocol
Goals

- **Normal state, every node is allowed to burst to line rate**
- **CAP is activated when Congestion is Detected:**
  - HOL timer expires
  - Output Link BW utilization exceeds threshold
- **Sends Fair rate Message to upstream node to back-off**
- **Maximizes link utilization by continuously adjusting advertised rate**
- **Returns to normal state when congestion disappears**
- **Protocol protects against multiple failure scenarios**

iPT-Control Access Protocol Example

Animated Slide

- 3 Node Example: Congestion on 150 M Pipe; 1 traffic class

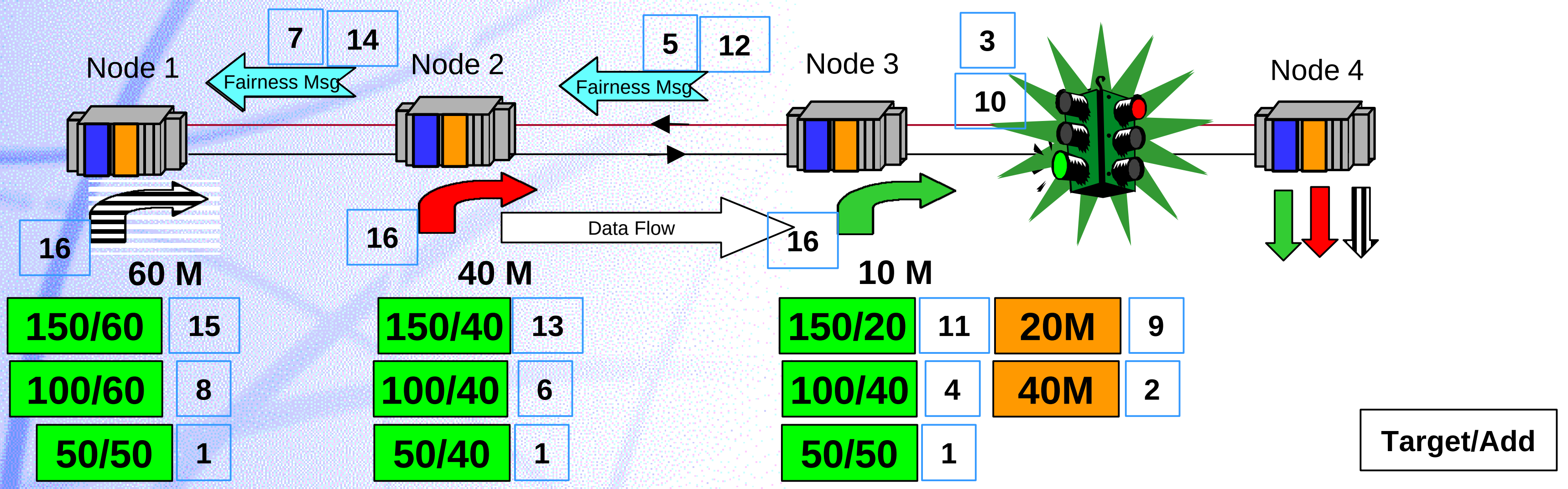


1. Node 1 sends 70 Mb/s to Node 4
2. Node 2 sends 40 Mb/s to Node 4
3. Node 3 sends 10 Mb/s to Node 4
4. Node 3 increases to 40 Mb/s to Node 4 and climbing to 50M
5. Node 3 declares congestion
   when Node 3 add traffic reaches 40M
   a. HOL timer expires
   b. Aggregate BW on output link > congestion threshold
   Node 3 detects 3 sources
6. Node 3 set its target add rate to 50 M
7. Node 3 send Fairness Message to Node 2
8. Node 2 sets its target add rate to 50 M
9. Node 2 send fairness message to Node 1
10. Node 1 sets its target add rate to 50 M

If spare capacity is large enough, a higher rate will be advertised
Trade-off between stability and maximize utilization

iPT-Control Access Protocol Example Cont.

Animated Slide



1. Node 1-3 schedules 50M add traffic

2. Node 3 traffic drops to 40M

3. Spare threshold crossed

4. Node 3 increases target add rate to 100M

5. Node 3 advertises 100M to Node 2

6. Node 2 increase target add rate 100M

7. Node 2 advertises 100M to Node 1

8. Node 1 increase target add rate 100M

9. Node 3 traffic drops to 20M

10. Node 3 detects spare BW cross another threshold

11. Node 3 increases target add rate to 150M

12. Node 3 advertises 150M to Node 2

13. Node 2 increase target add rate 150M

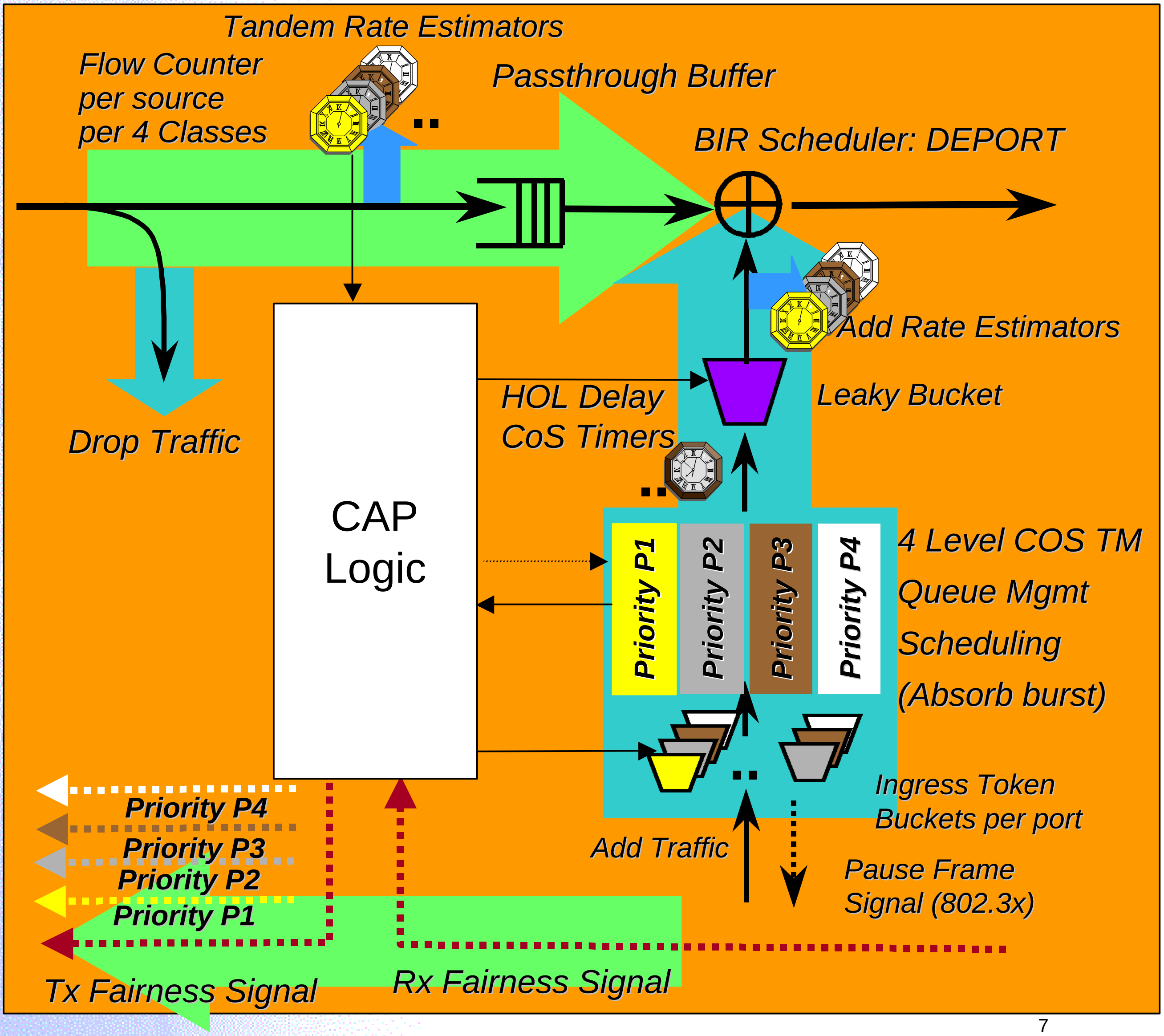
14. Node 2 advertises 150M to Node 1

15. Node 1 increase target add rate 100M

16. All nodes reaches un-congested steady state transmission

iPT- Control Access Protocol
Detailed Functional Blocks

1. **Tandem Rate Estimators**
2. **Add Rate Estimators**
3. **Scheduler**
   1. Control Messages
   2. Add Traffic Leaky Bucket
4. **HOL Delay Timers**
5. **Ingress Traffic Scheduler**
6. **Ingress Queue management with intelligent discard**
7. **Ingress Token Buckets per class for policing**
8. **Control Access Protocol Logic**
9. **Ring utilization statistics collection support**

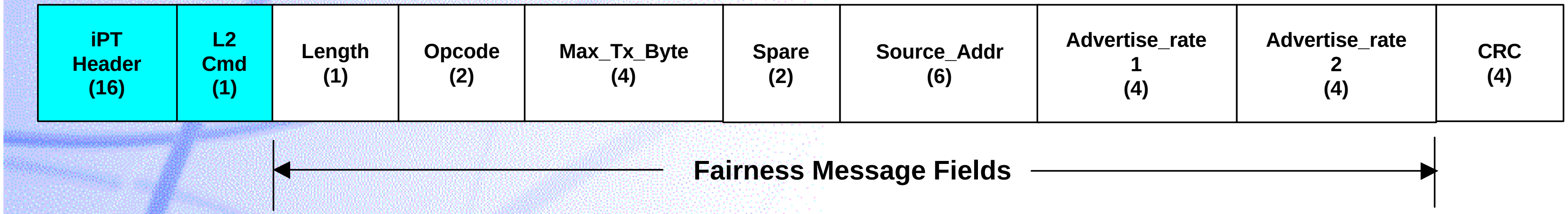


iPT-Control Access Protocol
Fairness Message Protocol

- **Message format**
  - 44 bytes, transmitted every “n” milliseconds (n = programmable)
- **Soft-state protocol**
  - source periodic retransmit message
  - closed loop control system
  - Very Robust
- **Compatible with L2 Protection Protocol**
  - Efficient BW utilization
  - high availability with single fault

iPT-Control Access Protocol

Fairness Message Detail Description

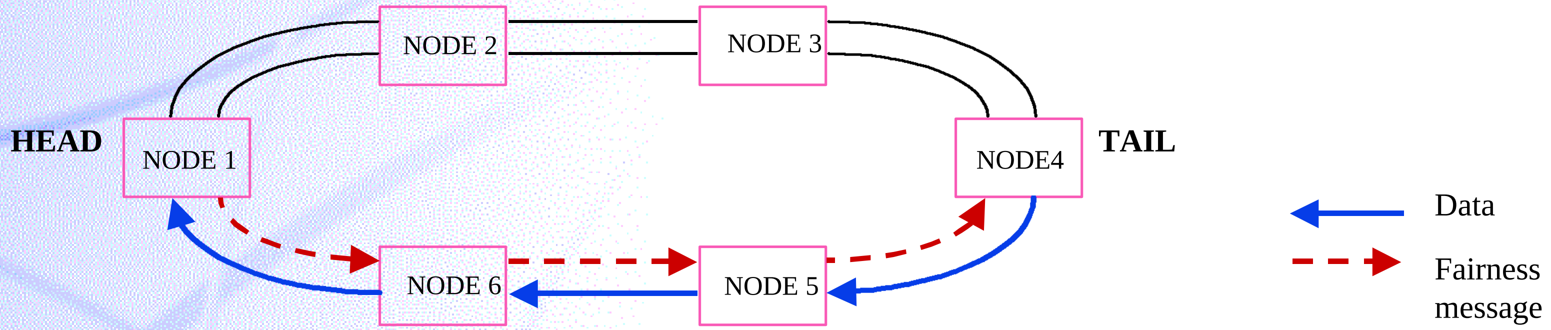


iPT Header (16)	L2 Cmd (1)	Length (1)	Opcode (2)	Max_Tx_Byte (4)	Spare (2)	Source_Addr (6)	Advertise_rate 1 (4)	Advertise_rate 2 (4)	CRC (4)

Fairness Message Fields

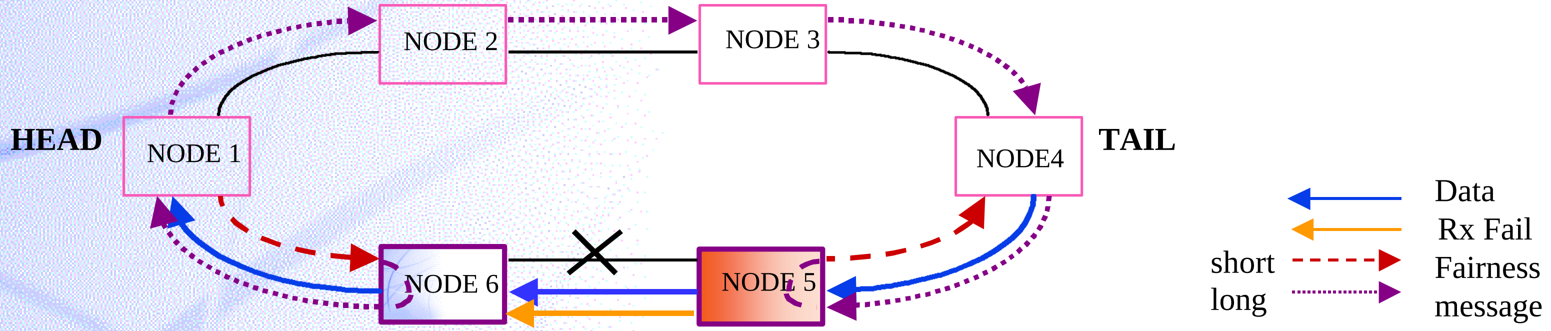
- Length [7:0]: Length in bytes of fairness message. Covers Fairness message fields
- OPCODE[15:0]: [15] 0=invalid message, 1=valid message
  [14] 0= not loop back message, 1= loop back message
  [13] 0= not direct, 1= direct
  [12] 0= forward, 1=not forward
  [11] 0=down stream Rx failed, 1=not failed
  [10] 0= version
  [9:4] hop count to congestion
  [3:1] last HOL packet priority
  [0] 0= no HOL congestion, 1 HOL timer congestion
- Max_Tx_Byte maximum link BW in bytes.
- Source_Addr Message Source address, used for source removal
- Advertised_rate*n* Advertised rate to upstream node, 2 classes defined
- CRC CRC-32 for message integrity

iPT-Control Access Protocol
Case1: Normal State of Operation



- **Node_1 is the HEAD node and it sends a fairness message to Node_6**
  - DS_rate, loopback=0, forward=0, RxFail=0
- **Node_6 is a CHAIN node. It receives DS_rate and applies to its leaky bucket. It forwards the same message to Node_5**
  - DS_rate, loopback=0, forward=0, RxFail=0
- **Node_5 is another CHAIN node. It receives DS_rate and applies to its leaky bucket. It forwards the same message to Node_4**
  - DS_rate, loopback=0, forward=0, RxFail=0
- **Node_4 is the TAIL node. It receives DS_rate and applies to its leaky bucket.It is the Tail node. It does not forward the message.**

iPT-Control Access Protocol
Case 2: Single Link Failure

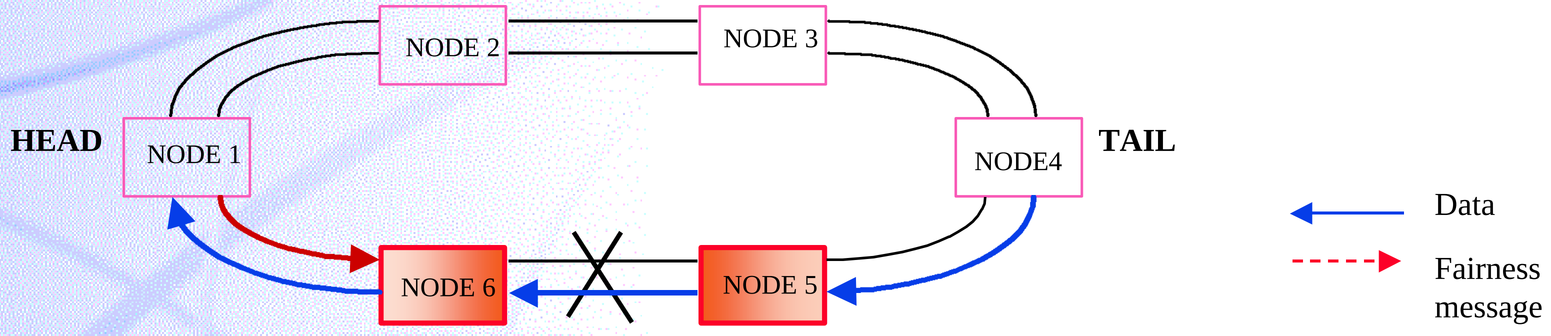


Failure occurs between Node_5 and 6 in the Counter Clockwise Ring.

Node_5 detects failure: RX_FAIL set. Node_5 sends status to Node_6

- **Node_1 is the HEAD node and it sends fairness message to Node_6**
  - DS_rate, loopback=0, forward=0, RxFail=0
- **Node_6 receives DS_rate and applies to its leaky bucket, and forwards the message to Node_5. But, Node_6 has received RX_FAIL message and loopbacks message to Node_5 via long path**
  - DS_rate, loopback=1, forward=0, RxFail=0
- **Node_5 receives fairness message on long path and applies to its leaky bucket, and forwards the message to Node_4.**
  - DS_rate, loopback=0, forward=0, RxFail=0
- **Node_4 receives DS_rate and applies to its leaky bucket. It is the Tail node.**

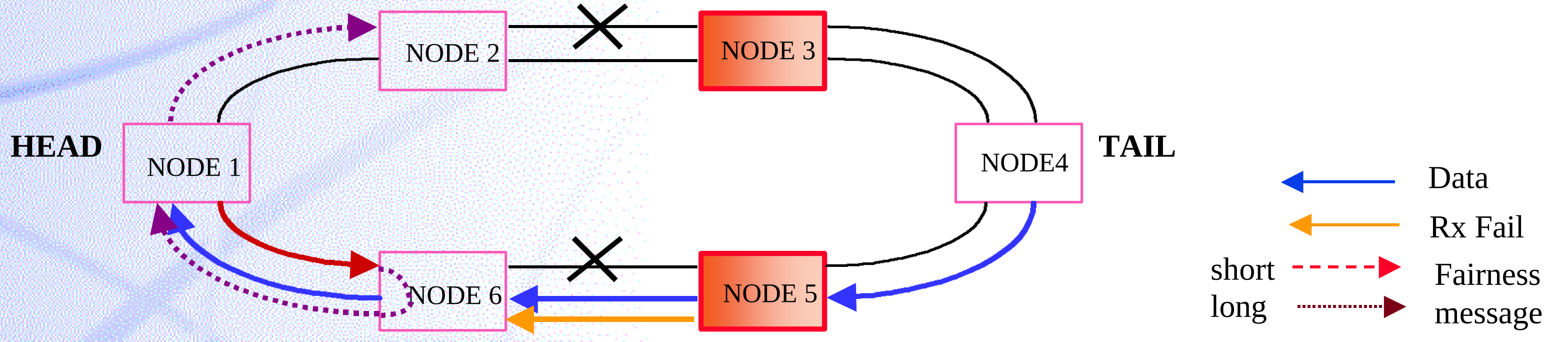
iPT-Control Access Protocol
Case 3: Double Link Failure; Same Span



- **Both sides detect failure and do not loopback Fairness messages.**
- **Node_6 does not loopback Fairness messages. It becomes the tail for counter-clockwise ring.**
- **Node_5 detects failure and becomes tail node for clockwise ring.**
  - Node_5 times out in receiving fairness message in long path.
  - L2 protection detects failure and re-routes packets away from failure. Node_5 detects no link utilization in clockwise ring.

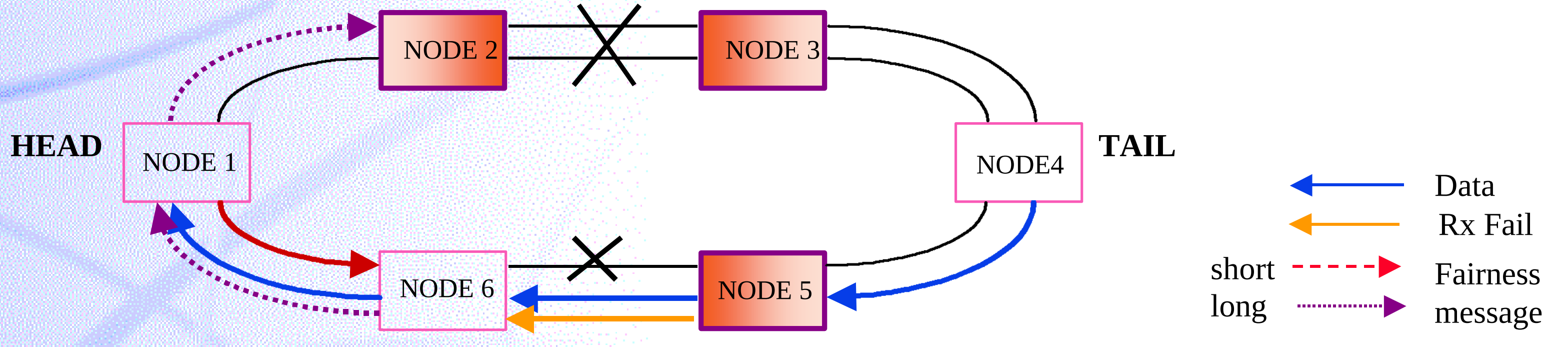
    (FMP cannot distinguish between case 3 and case 4)
- **NO fairness message is generated by Node_5 in the counter-clockwise ring.**

iPT-Control Access Protocol
Case 4: Two Independent Spans with Link Failure



- **Node_1 is the HEAD node and it sends fairness message to Node_6**
- **Fairness Message from Node_6 for Node_5 does not get to destination.**
  - Node_5 times out in receiving fairness message in long path from Node_6.
  - L2 protection still forwards data through clockwise link between Node_6 and Node_5.
    - Set congestion threshold to **NEW** threshold for Node_5.
- **Node_3 will operate in similar mode for counter-clockwise ring, with** NEW **threshold for congestion.**

iPT-Control Access Protocol
Case 5: Multiple Failures; Segmented Ring



- **Node_6 Fairness message RX timer times out.**
- **Fairness Message from Node_6 for Node_5 does not get to destination.**
  - Node_5 times out in receiving fairness message in long path from Node_6.
  - L2 protection still forwards data through clockwise link between Node_6 and Node_5.
    - Set congestion threshold to **NEW** threshold for Node_5.
- **Node_2 clockwise ring output and Node _3 counter-clockwise output do not see congestion due to L2 protection. Operates with normal state parameters.**

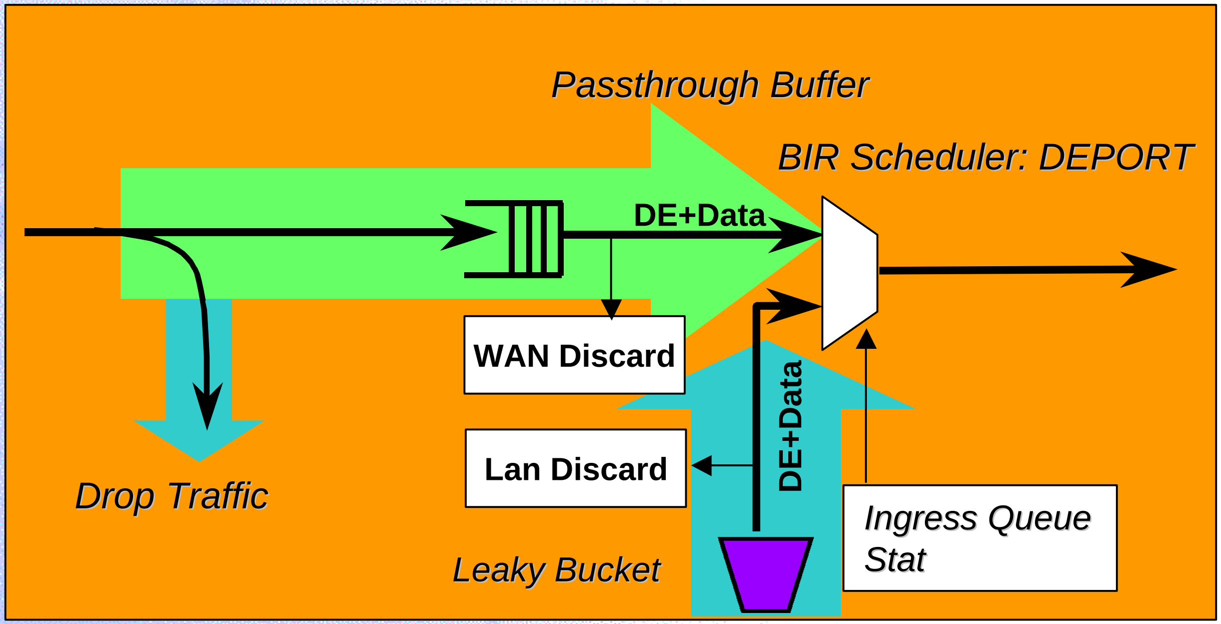
iPT-Control Access Protocol Conclusions

- CAP automatically and efficiently manages the WAN BW with QoS support to maximize its utilization.
- QoS is supported with Intelligent Ingress traffic management, scheduler, and policing.
- Provides statistics for performance monitoring.

Back Up Charts

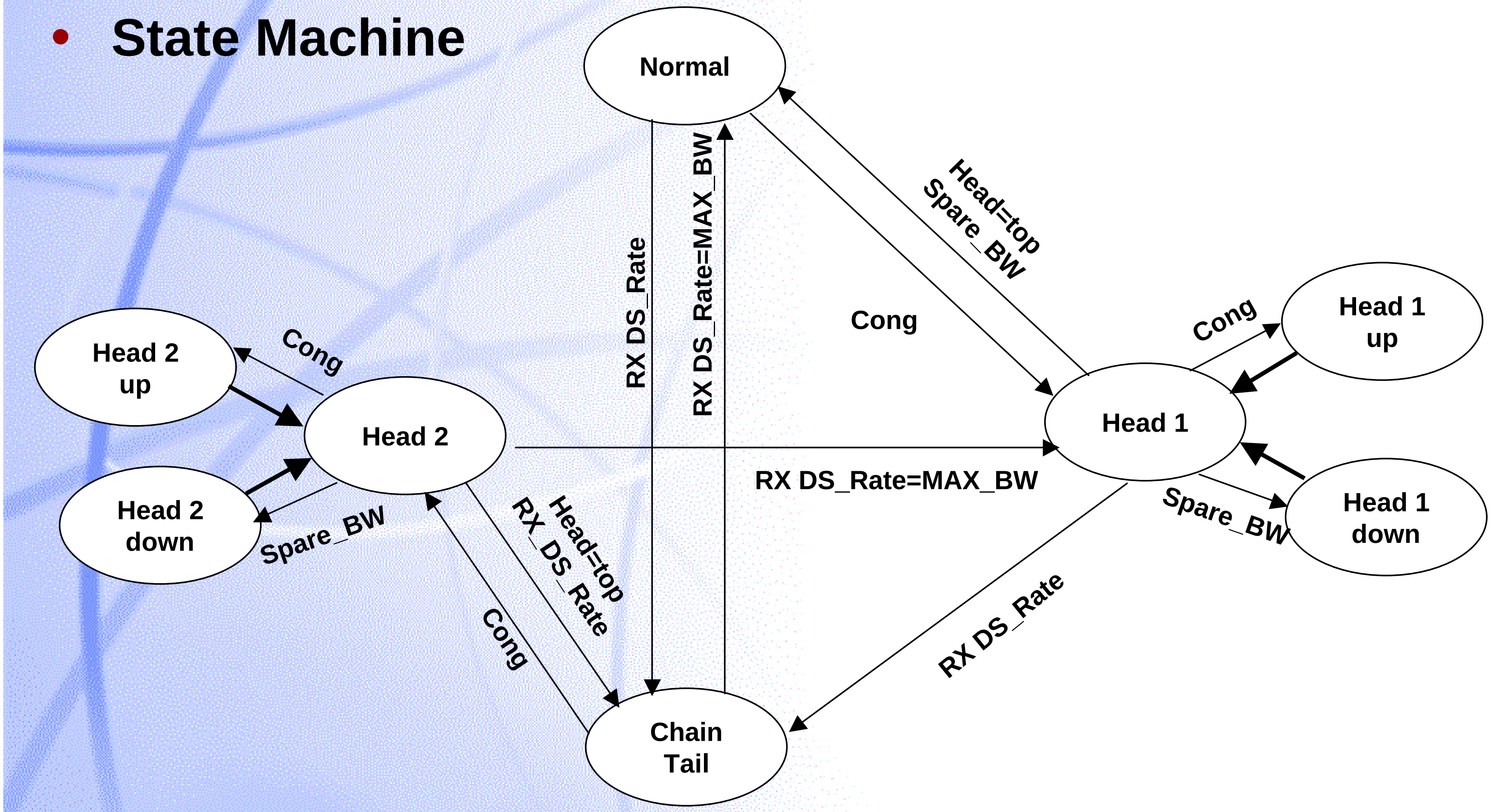
iPT-Control Access Protocol
DEPORT

- **DEPORT (Discard Eligible Packet On Ring Tandem)**
  - Ingress in-profile packet can causes discard on tandem DE (discard eligible) packet if Ingress Queue threshold has crossed.



iPT-Control Access Protocol State Machine

- **State Machine**


Normal
Head 1
Head 1 up
Head 1 down
Head 2
Head 2 up
Head 2 down
Chain Tail
RX DS_Rate
RX DS_Rate=MAX_BW
Head=top
Spare_BW
Cong
Cong
Spare_BW
Cong
Spare_BW
RX DS_Rate=MAX_BW
RX DS_Rate
Head=top
RX_DS_Rate
Cong