Priority Pause support for CN (e.g. BCN) Mechanism
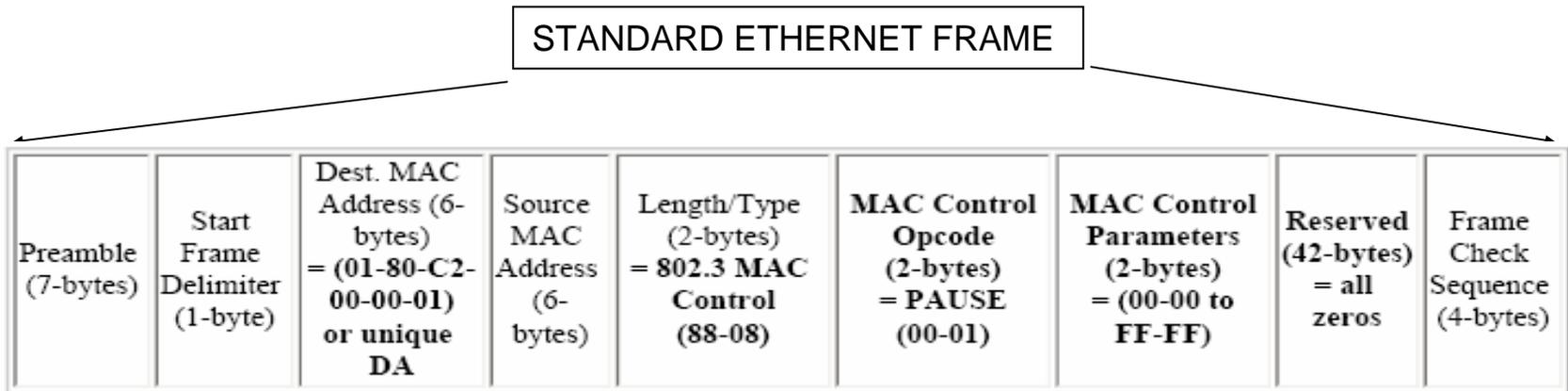
Asif Hazarika, Fujitsu
Bob Brunner, Ericsson

ahazarik@fma.fujitsu.com
**robert.brunner@ericsson.com**

Agenda

- **Recap of Pause and its issues**

- **Review of suggested Priority Pause**

- **Review of simulation results**

- **Need for a Priority Pause mechanism**

- **Some suggested techniques**

PAUSE Message

STANDARD ETHERNET FRAME

Preamble (7-bytes)	Start Frame Delimiter (1-byte)	Dest. MAC Address (6-bytes) = (01-80-C2-00-00-01) or unique DA	Source MAC Address (6-bytes)	Length/Type (2-bytes) = 802.3 MAC Control (88-08)	MAC Control Opcode (2-bytes) = PAUSE (00-01)	MAC Control Parameters (2-bytes) = (00-00 to FF-FF)	Reserved (42-bytes) = all zeros	Frame Check Sequence (4-bytes)

- **PAUSE controls adjacent Link-Layer-Device's Transmitter**

- **PAUSE is time based (PAUSE Quantum = 512 bits of time)**

- **PAUSE time can be extended / timed-out (zero value)**

PAUSE Mechanism exhibits coarse grained Flow Control
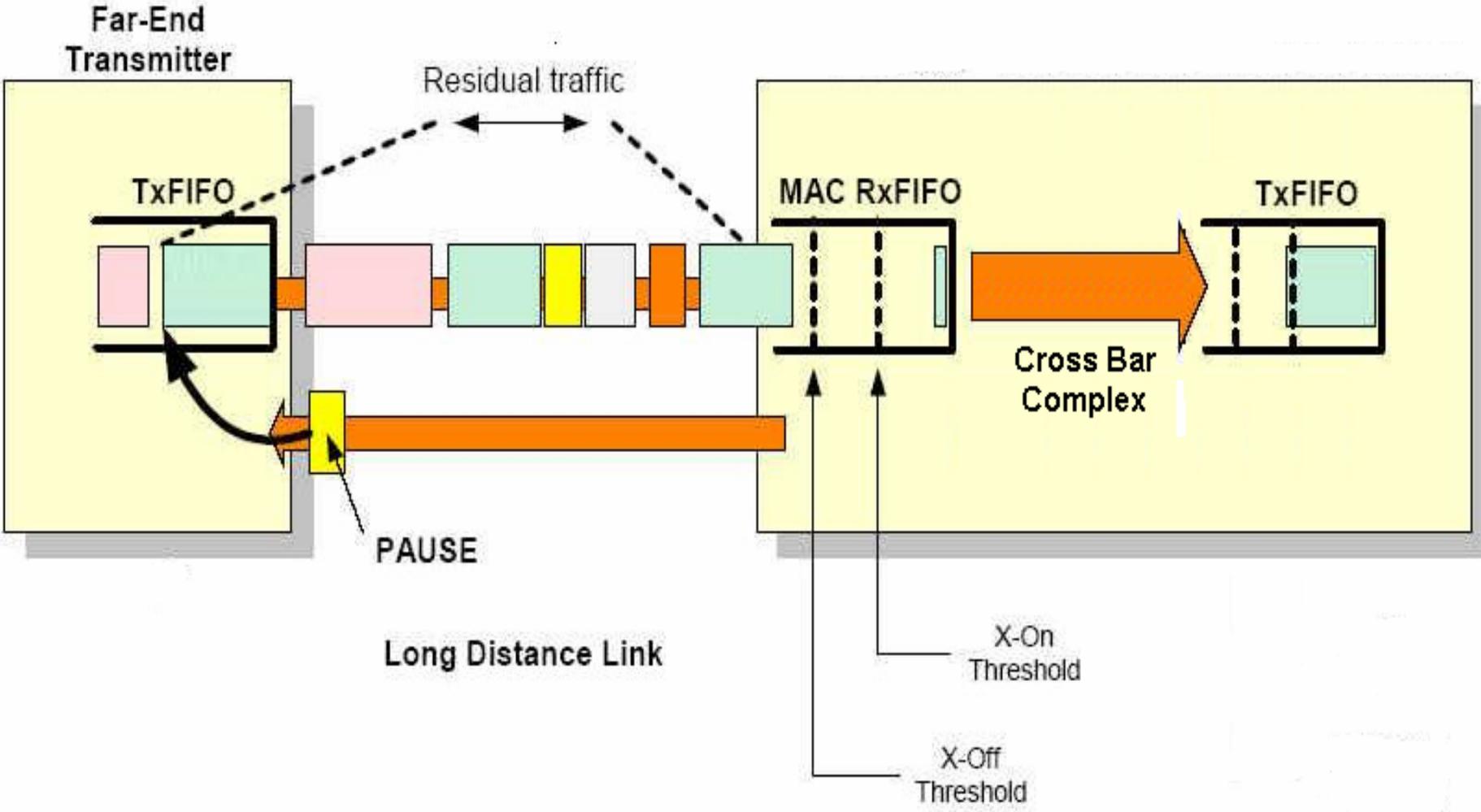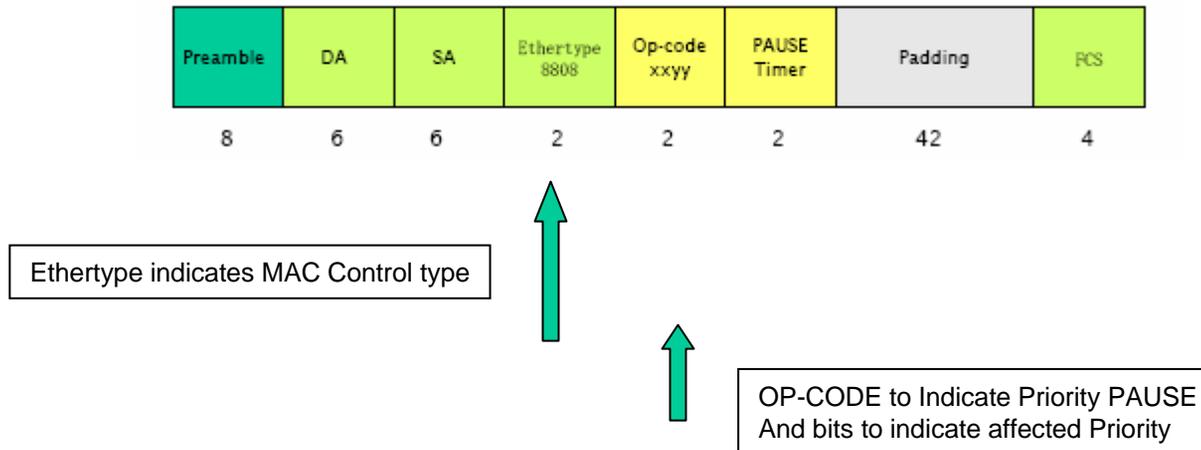
PAUSE INEFFICIENCIES

- **Response to Pause**
  - PAUSE from end-node will shutdown an adjacent LLD's transmitter
  - In turn, when a switch-node's resources are exceeded, PAUSE's are propagated towards all egress ports

- **Inefficiency**
  - PAUSE mechanism inevitably causes congestion spreading
  - PAUSE mechanism can result in a high degree of service degradation
  - shutdown of a whole link instead a particular flow causes system bandwidth to diminish
  - Effect on high priority traffic
    - Higher latency and jitter for high-priority traffic
    - Delay for high priority traffic
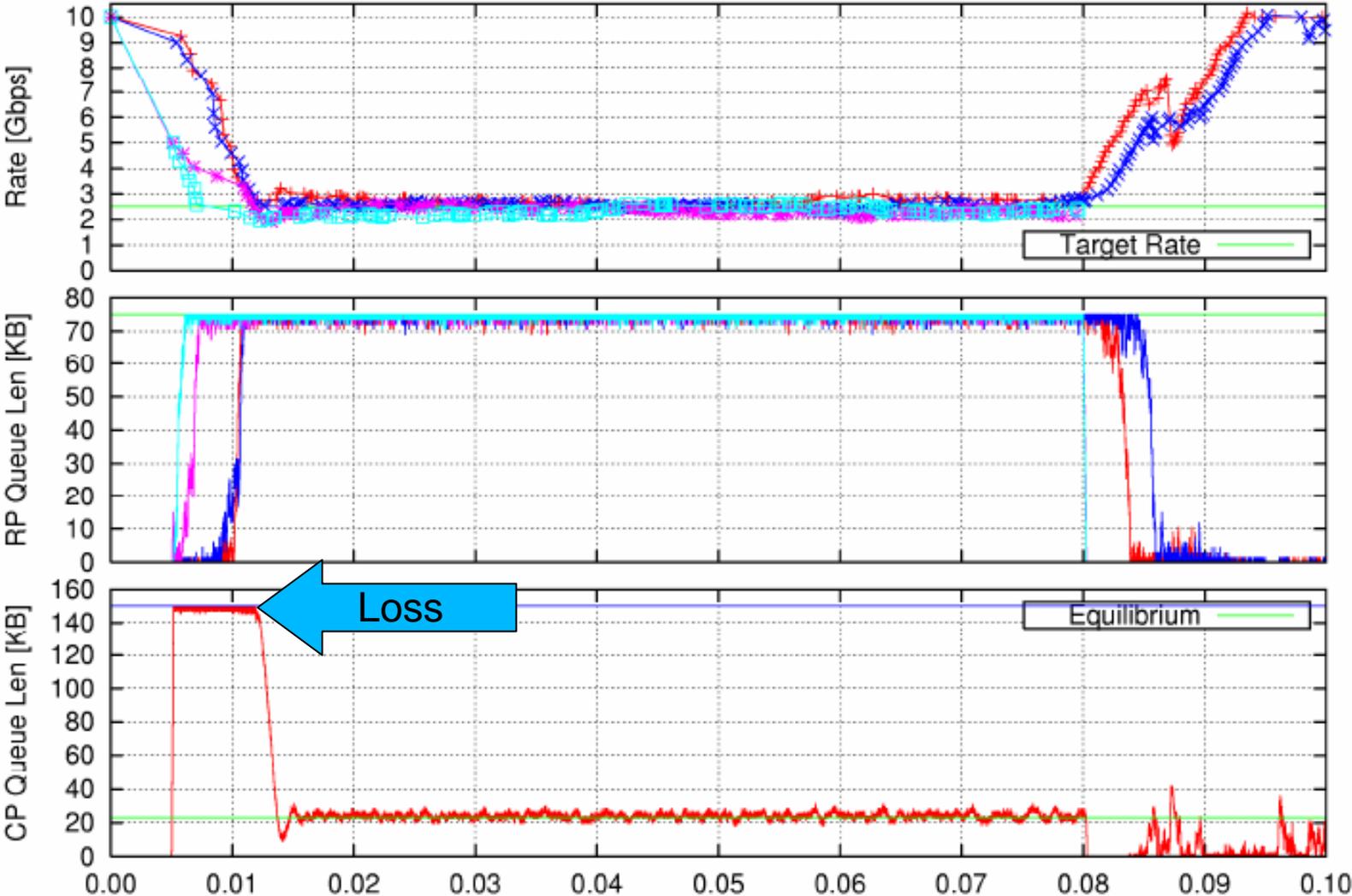   - No guarantees

Priority Pause - Recap
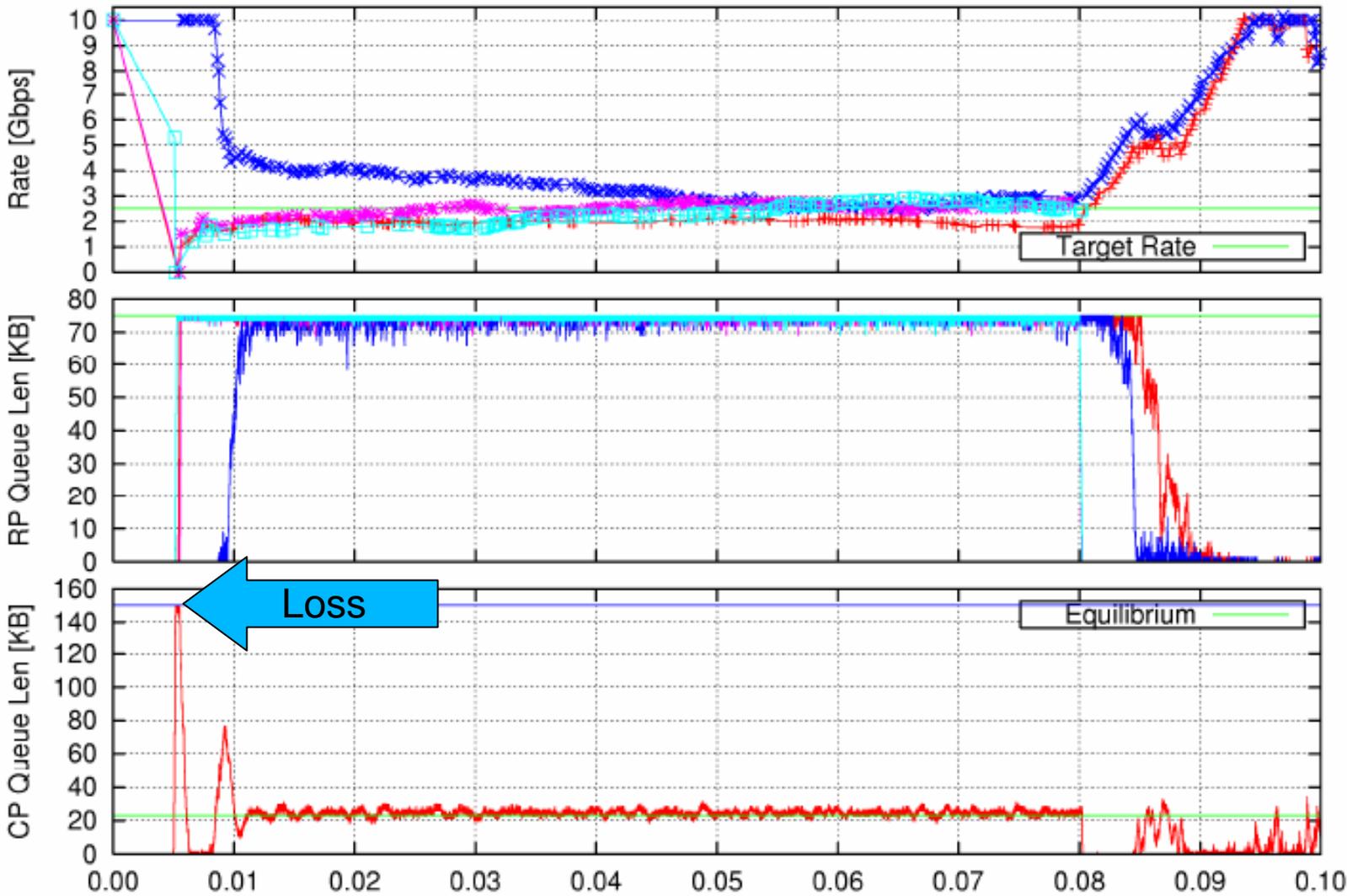
- **Priority Pause**
  - Creation of special PAUSE Frames
    - To address COS-PAUSE which is per class of service
  - Extension to MAC Control Opcodes (an example)
    - New control types that can be defined.
      - PAUSE FOR ALL Classes of service, so switch can pause certain classes
        - » e.g.PAUSE FOR BEST EFFORT PRIORITY FRAME (802.1p Class=0)

Preamble	DA	SA	Ethertype 8808	Op-code xxyy	PAUSE Timer	Padding	PCS
8	6	6	2	2	2	42	4

Ethertype indicates MAC Control type

OP-CODE to Indicate Priority PAUSE
And bits to indicate affected Priority

P802.1au Congestion Notification

6

Rational for PAUSE: Simulation without BCN(0,0)

Rational for PAUSE: Simulation with BCN(0,0)

P802.1au Congestion Notification

Rational for "Fuzzy-BCN": Simulation with BCN(0,0)

P802.1au Congestion Notification

BCN Concepts

FULL QUEUE Q_{SC} EMPTY QUEUE Q_{EQ}

Severe Congestion Threshold Equilibrium Threshold

IN

OUT

•Sample Frame with Probability P

GENERATE BCN(0,0)

GENERATE BCN(Delta1,Delta2)

•If Sample Frame with Has no RLT >

BCN(0,0) | BCN(Q_{off},Q_{Delta}) | No MSG

•If Sample Frame with Has RLT >

BCN(0,0) | BCN(Q_{off},Q_{Delta})

GENERATE BCN

$$Q_{off} = Q_{len} - Q_{eq}$$

$$Q_{delta} = Q_{len} - Q_{old}$$

P802.1au Congestion Notification

BCN Concepts with Priority Pause

FULL QUEUE Q_{SC}

Q_{EQ} EMPTY QUEUE

Severe Congestion Threshold

Equilibrium Threshold

IN

OUT

GENERATE PRIORITY BASED PAUSE

BCN(0,0)

BCN(Q_{off}, Q_{Delta})

No MSG

GENERATE BCN

BCN(0,0)

BCN(Q_{off}, Q_{Delta})

$Q_{off} = Q_{len} - Q_{eq}$

P802.1au Congestion Notification

$Q_{delta} = Q_{len} - Q_{old}$
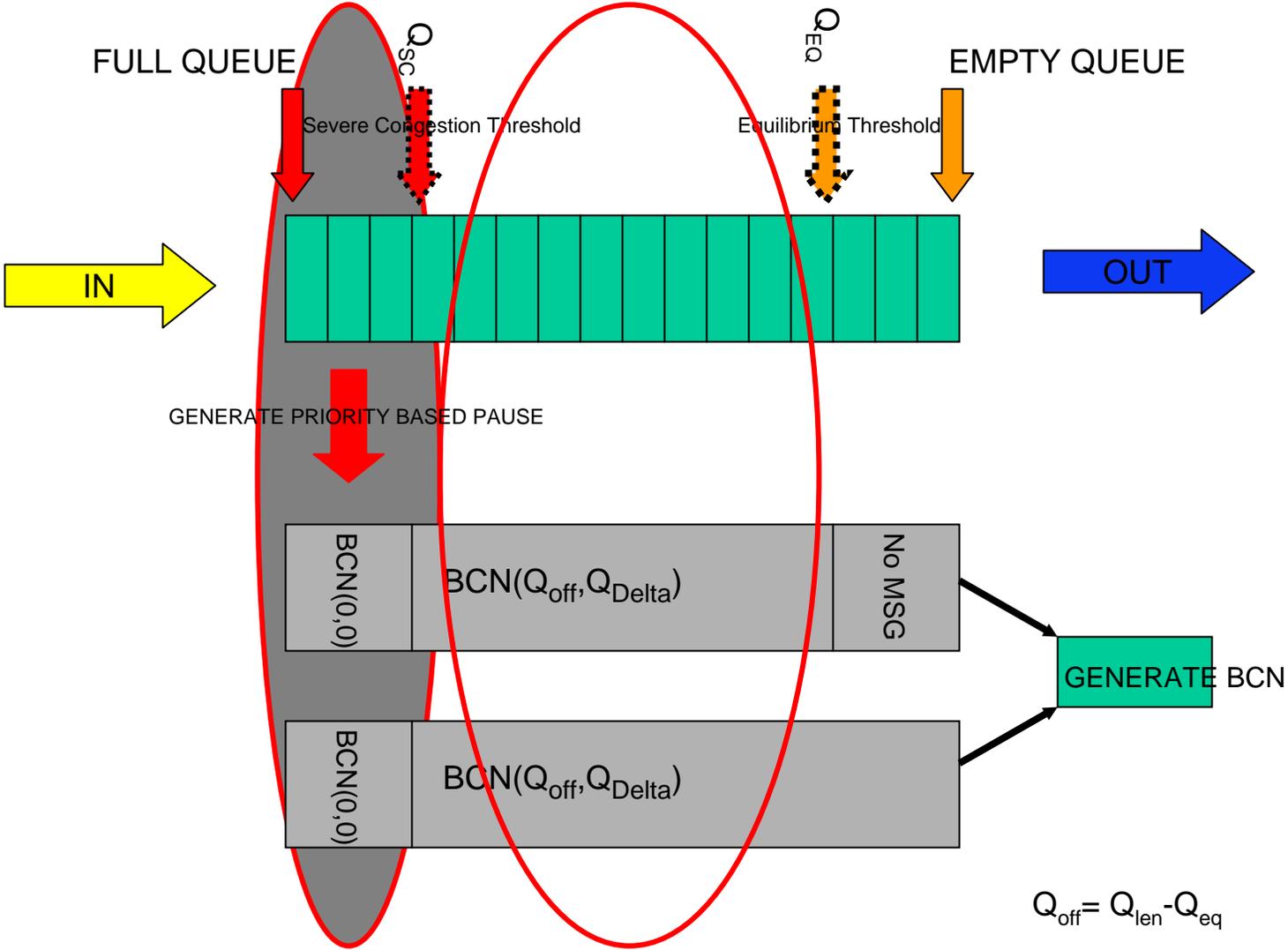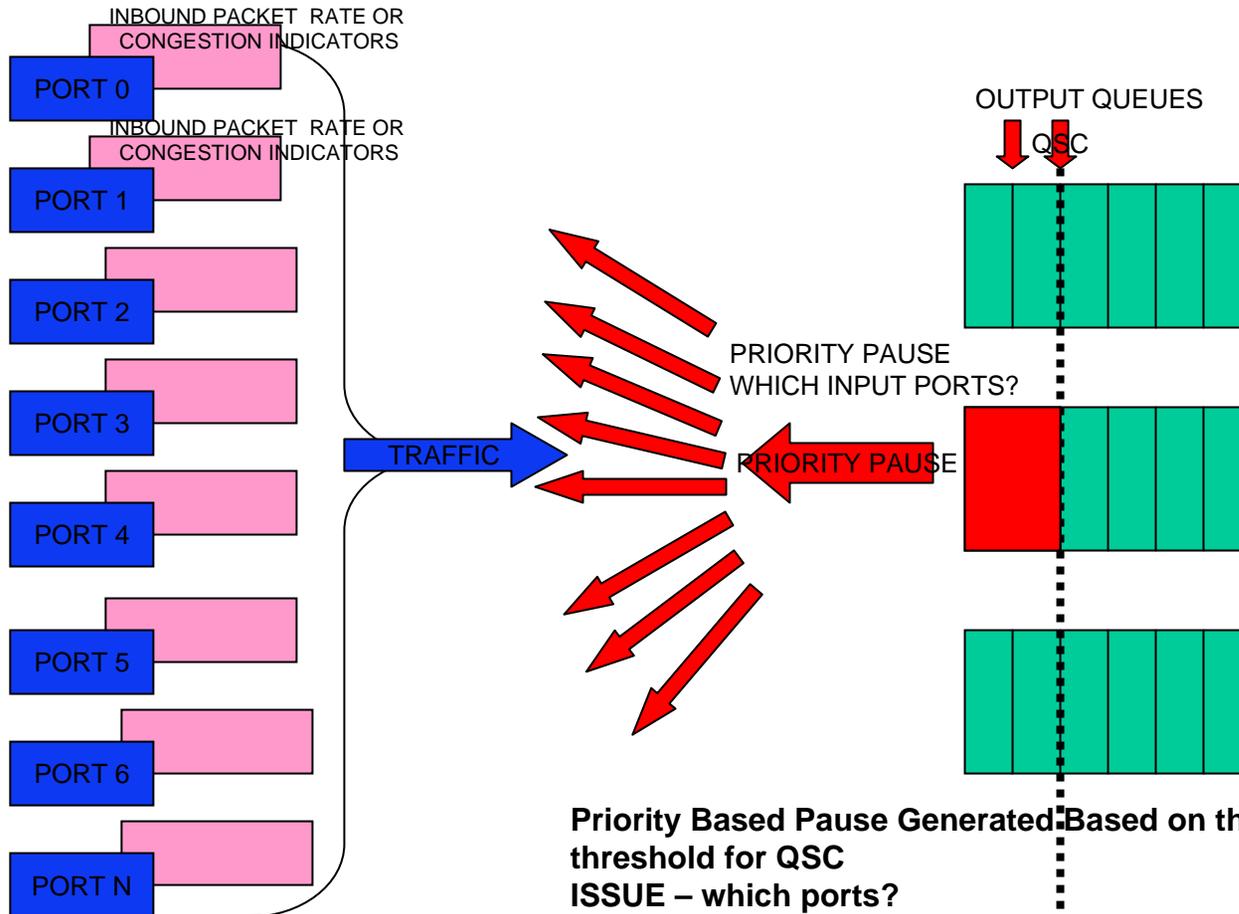
ISSUES WITH PRIORITY PAUSE

- **During Severe Congestion**
  - Priority Pause needs to be generated to relieve the congestion
  - To provide interim relief before BCN mechanism sets in
  - However the mechanism has to be selective to avoid unnecessary back-pressure messaging

- **Suggested approach**
  - Mechanism should in place to direct the PAUSE to the related port that contributed to Congestion

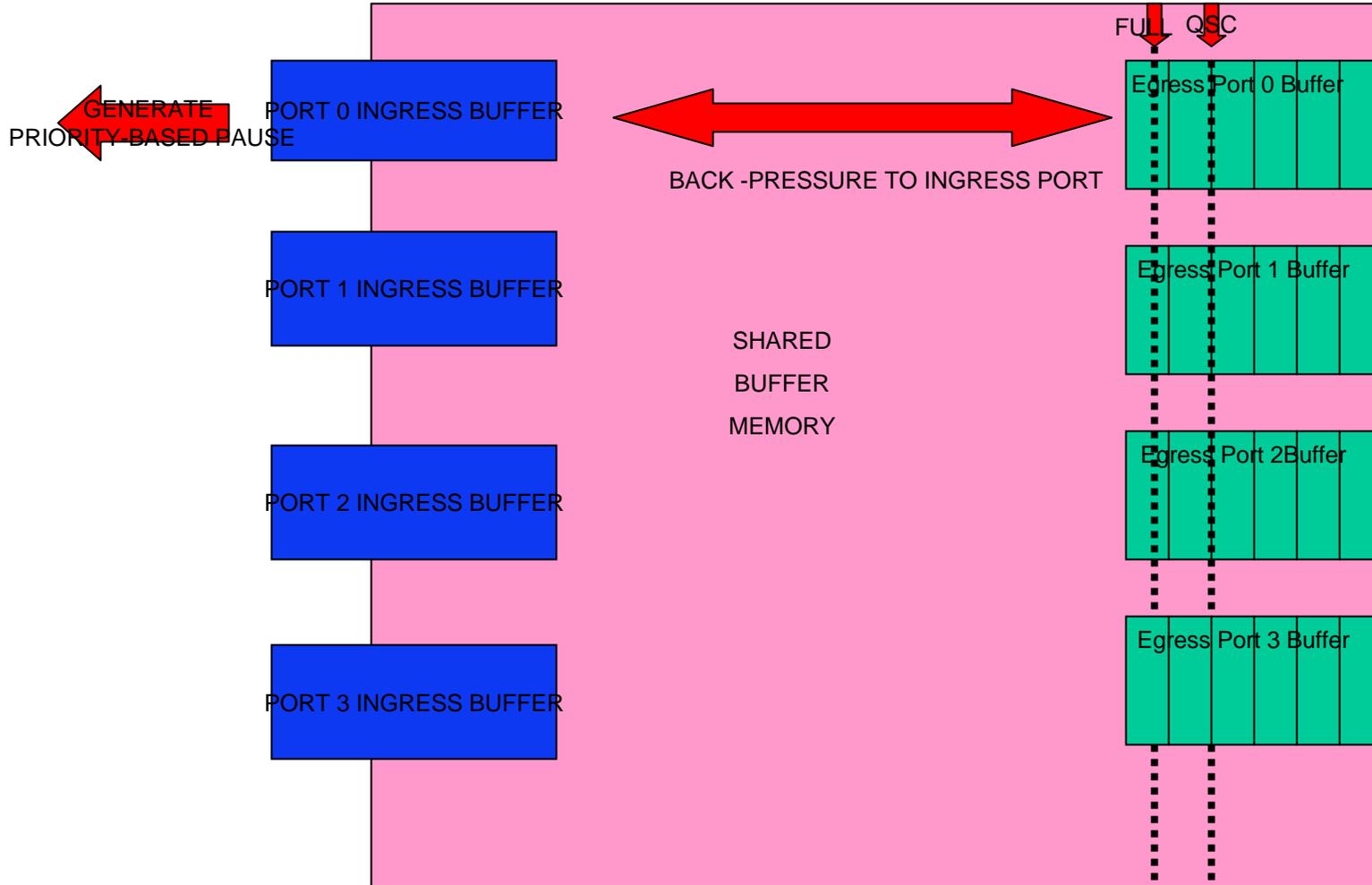# Generation of Priority-based of Pause (Output Queue Model)

PRIORITY PAUSE WHICH INPUT PORTS?

PRIORITY PAUSE

TRAFFIC

INBOUND PACKET RATE OR CONGESTION INDICATORS

OUTPUT QUEUES

QSC

PORT 0
PORT 1
PORT 2
PORT 3
PORT 4
PORT 5
PORT 6
PORT N

Priority Based Pause Generated Based on the threshold for QSC
ISSUE – which ports?
PP to all ports – will cause inefficiency
Solution :
• Measure of per port contribution to congestion
• Generate PPP for the input port causing most congestion

Generation of Priority-based of Pause (Input Queue Based Model)

- CONGESTION IN EGRESS BUFFER TRANSLATES TO THE RELATED INGRESS BUFFER
- INGRESS BUFFER TRIGGERS THE PRIORITY PAUSE BASED ON THE STATUS OF THE OUTPUT QUEUES

FULL   QSC

Egress Port 0 Buffer

GENERATE
PRIORITY-BASED PAUSE

PORT 0 INGRESS BUFFER

BACK -PRESSURE TO INGRESS PORT

PORT 1 INGRESS BUFFER

Egress Port 1 Buffer

SHARED

BUFFER

MEMORY

PORT 2 INGRESS BUFFER

Egress Port 2Buffer

PORT 3 INGRESS BUFFER

Egress Port 3 Buffer

P802.1au Congestion Notification

Thank You