Congestion Management: Requirements and Proposals – A TEM's View

Robert Brunner, Ericsson
Shashank Merchant, Nokia

Robert.Brunner@ericsson.com,
shashank.merchant@nokia.com

Ethernet as a Backplane Transport

(focus on Congestion Management)

- Primary goal is to utilize Ethernet as a silicon-to-silicon inter-connect, with cross-bar like functionality
- Need to transport “anything” over Ethernet, such as data content from SONET, FR, POS, ATM, PPP, & Ethernet, over a lightweight Ethernet tunnel preserving channel information
- Packets should not be discarded in the switching sub-system.
  - Intelligence on the edge
- Rate mis-match between the blades, as well as multi-chassis situation should be considered
- Congestion Management implementations should be in Hardware to maintain low latency
  - Software involvement for configuration and monitoring purpose only
- CM mechanism should support multiple priority classes
- Proposed solution should accommodate asymmetric traffic
- Fabric Interface Chip should provide congestion feedback to Traffic Manager to take appropriate Rate Control/Drop action
  - Congestion feedback needs to be standardized

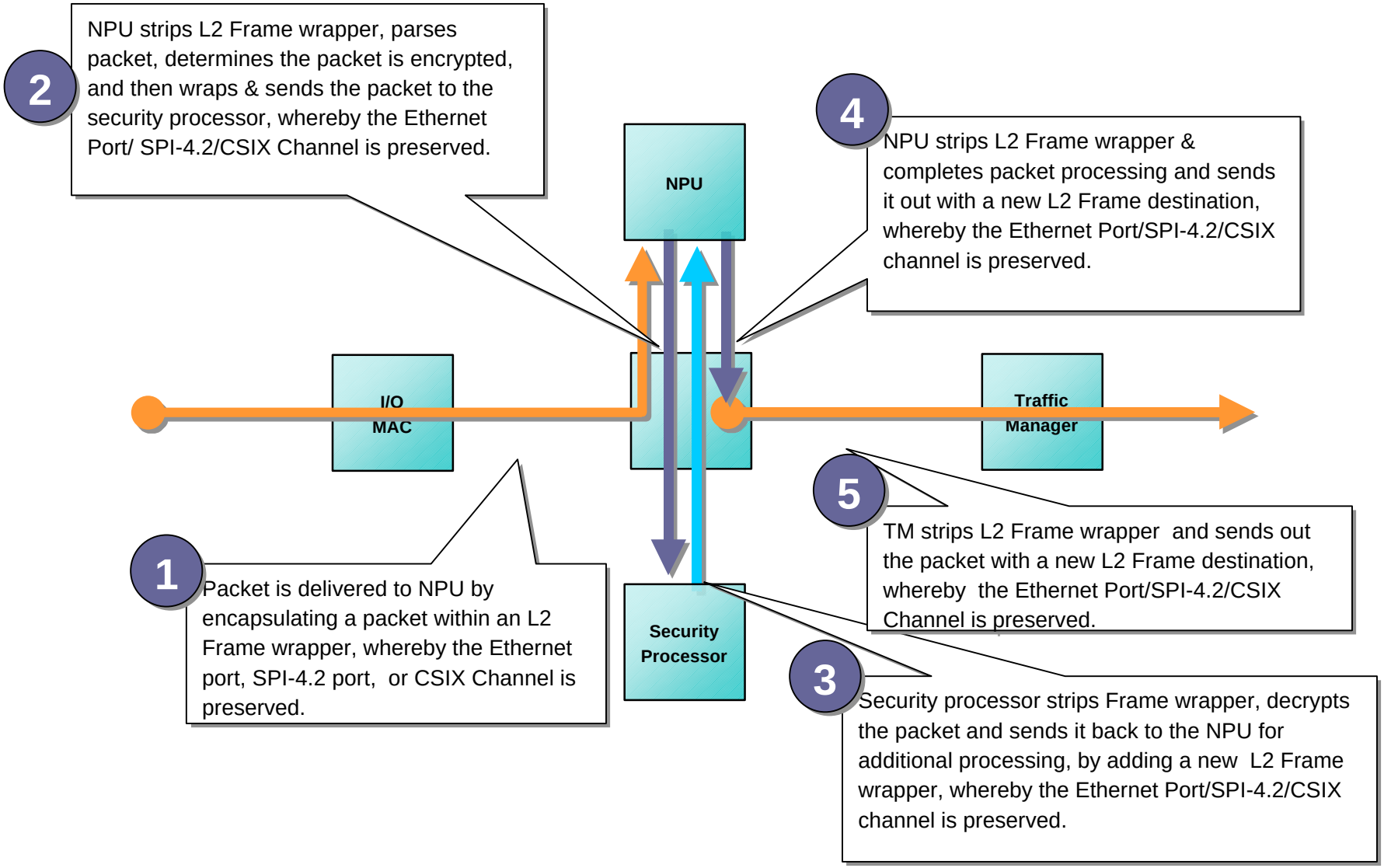
Ethernet as a Backplane Transport (contd)

(focus on Congestion Management)

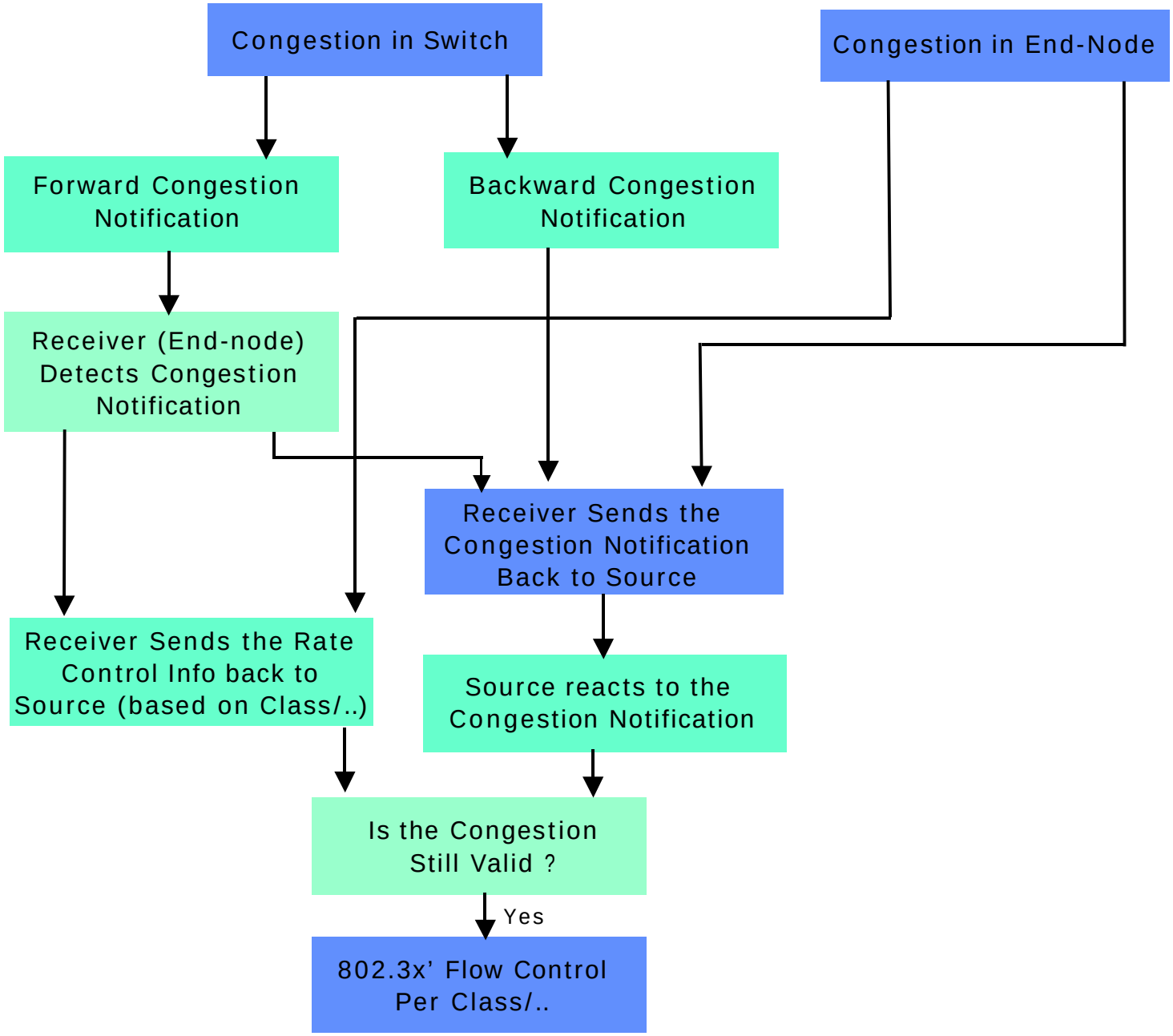
- 802.3x flow control mechanism as its defined is not sufficient
  - Causes head of line blocking
- Need to tunnel back-pressure (flow/rate control equivalent of) NPSI/ CSIX/ SPI signals on a per “channel basis” over Ethernet
- Need Ethernet to support an equivalent “channel based” flow control paradigm as NPSI/CSIX/SPI, in order to allow inter-operability with Ethernet
- CM feedback latency defines buffer requirements in Traffic Manager
  - Feedback latency needs to be as small as possible

Intra-Blade or Inter-Blade, AMC to AMC/RTM Flows

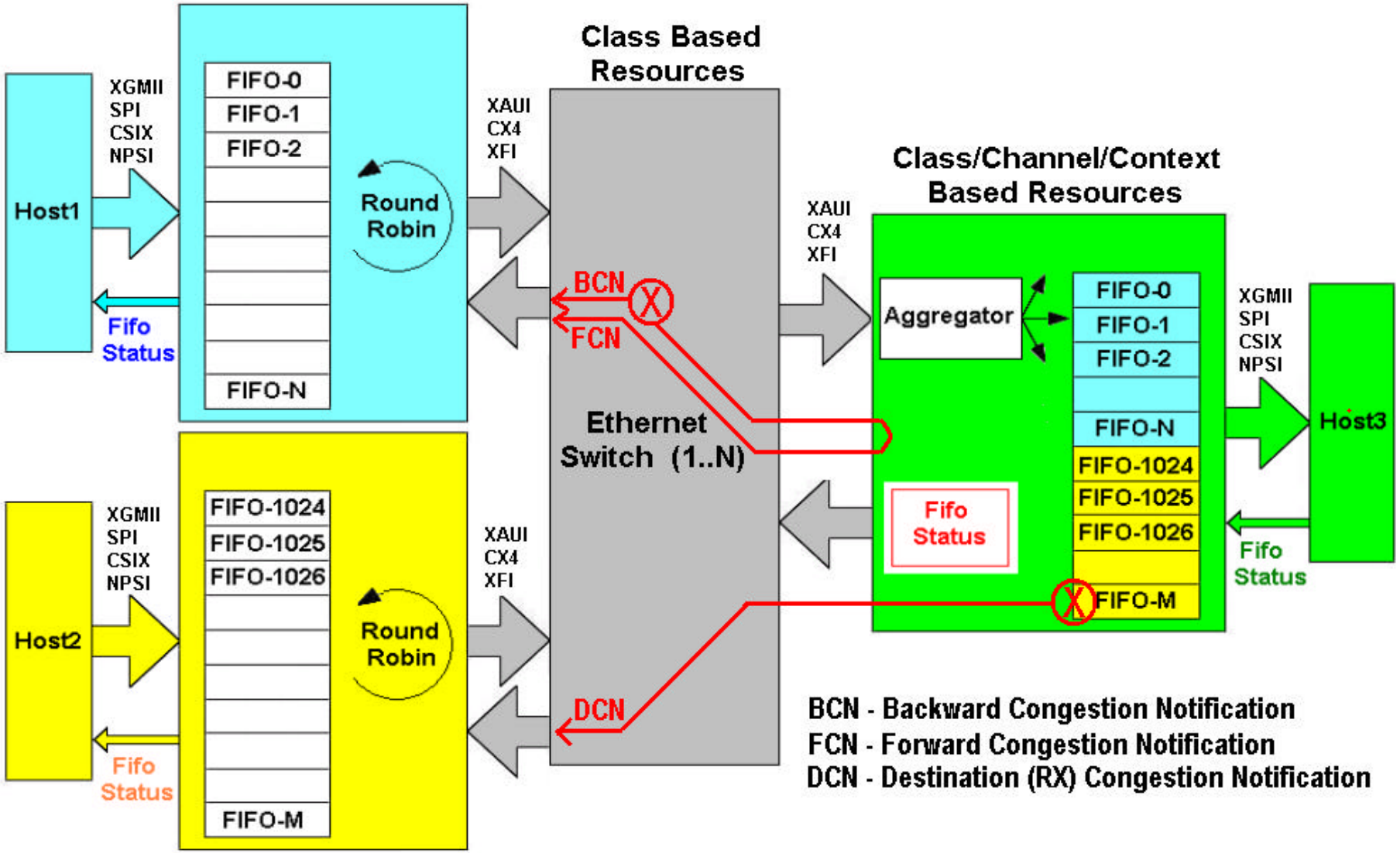
SPI-4.2, CSIX, or Ethernet Flows



Congestion Management



Context Specific Back-Pressure over Ethernet

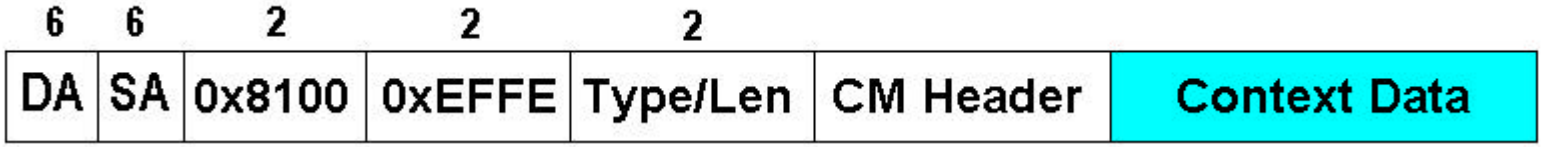


Note: Link-Layer Flow Control provided by "Class" based 802.3x message

CM-INFO for BCN, FCN, and DCN (option 1)

- CM-INFO Packet = Dedicated Packet from Destination (end-node, or intermediate switch) carrying the congestion management related information.
- Option 1:
  - CM-INFO packet is switched based on MAC address
  - Use VLAN priority to carry it as highest priority packet

BCN, FCN, DCN to Source



Note: VLAN with highest priority, and VLAN ID = 0xff3 is reserved for CM-INFO

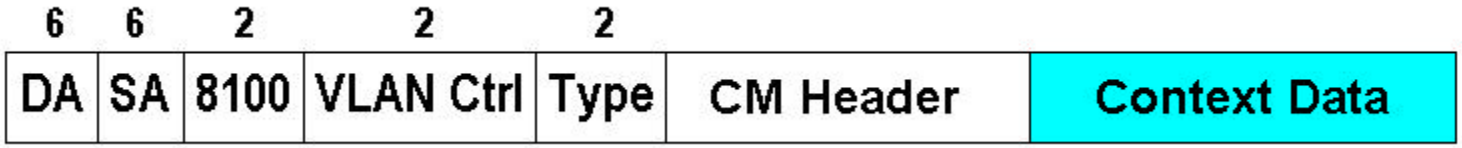
CM-INFO for BCN, FCN, and DCN (option 2)

Option 2:

CM-INFO packet has a well-defined Ether-type

Intermediate switches understand the Ether-type and gives the packet a higher priority. Switching based on MAC address

BCN, FCN, DCN to Source



CONGESTION MANAGEMENT HEADER

- Congestion_Type (Switch, RX; Class, Channel, Context)

- Source_Resource_Control (NIL-XOFF-RATE-XON)

- Class/Channel/Context_Hash

- Context_Length

Observations

- Our proposal is meant to emphasize the requirements. We are open to any solution which meets the requirements in a standardized manner.
  - Example: Ethernet Switch can provide more involved congestion management to reduce end-nodes complexity
- The intelligence for congestion management can be in the transmitter or receiver or even switches.
  - Simulations necessary to verify the schemes
    - Example: Buffering/FIFO sizing in the receiver
- There is a need to expand the 802.3x flow control to provide at least the class-based information back to the source.
  - Overcomes HOL blocking, and enables guaranteed QoS for high priority traffic
- VLAN tag for the CM frames reduces number of classes available for other traffic, and cause mapping issue from DSCP to 802.1p bits