InfiniBand Credit-Based Link-Layer Flow-Control

802.1 DCB TG - IEEE 802 Plenary

March 2014
Introduction to InfiniBand Credit Based Flow Control

- “Credit” Represents Receiver Commitment
- In-band Delivery of Flow Control Credits
- Requires Accurate Accounting “in-sync” with Data Transmission
  - vs. pause/xon-xoff schemes

- Primary Challenge is Resiliency
  - Loss of Flow Control Updates
  - Loss of Data Packets

- “Absolute” Credits
  - vs. “Incremental”
  - “Credit Limit” – total allowed since initialization of the link

- InfiniBand Specification Vol.1 Section 7.9
Principles of Operation (Simplified)

- **Receiver (per VL)**
  - Tracks **ABR** – (Adjusted) Blocks Received
    - Counts total blocks received since initialization of the link
    - Updated (incremented) for every received packet
  - Calculates “Credit Limit” (**FCCL**)
    - FCCL is ABR + "Available Buffer Space"
    - FCCL Sent to Transmitter via Credit Packets

- **Transmitter**
  - Tracks **FCTBS** – Total Blocks Sent
    - Counts total blocks sent since initialization of the link
    - Updated (incremented) for every sent packet
  - Receives FCCL in Credit Packets
  - Packet Transmission is Allowed if FCTBS+"Packet Size" is smaller than or equal last received FCCL
Principles of Operation (cont’d)

**Receiver**

- **Initialization of the Link**
- **Blocks already delivered onwards** ...
- **Blocks in Receiver’s Buffer**
- **Available Buffer Size**
- **Receiver’s Total Buffer Size**
- **Blocks on the wire**
- **“Credits”** (FCCL - FCTBS)
- **FCTBS** (Total Blocks Sent)
- **ABR** (Total Blocks Received)
- **FCCL** (Credit Limit)

**Transmitter**
Resiliency to Transient Failures

- **Absolute Credits**
  - Inherently Resilient to Loss of Credit Packet

- **Algorithm Relies on Consistent View of “Total Bytes Sent”**
  - FCTBS (on transmitter) must remain equal to ABR+Blocks_on_the_wire (on the receiver).
  - Disrupted by Loss of Data Packet
    - …ABR falls behind
  - Solution: Periodically Force Sync to Guarantee Consistent View
    - FCTBS Piggybacked in (reverse) Credit Packet
InfiniBand Credit Based Flow Control (with Resiliency)

- **Receiver (per VL)**
  - Tracks **ABR** – (Adjusted) Blocks Received
    - Counts total blocks received since initialization of the link
    - Updated (incremented) for every received packet
    - **Override with FCTBS value reported by Transmitter**
      - piggybacked in received (reverse) Credit Packet
  - Calculates “Credit Limit” (**FCCL**)
    - FCCL is ABR + ”Available Buffer Space”
    - FCCL Sent to Transmitter via Credit Packets

- **Transmitter**
  - Tracks **FCTBS** – Total Blocks Sent
    - Counts total blocks sent since initialization of the link
    - Updated (incremented) for every sent packet
    - **FCTBS is sent to Receiver**
      - piggybacked in (reverse) Credit Packet
  - Receives FCCL in Credit Packets
  - Packet Transmission is Allowed if FCTBS+”Packet Size” is smaller than or equal last received FCCL
InfiniBand Credit Based Flow Control (cont’d)

- **Flow Control Blocks**
  - 64B (working towards configurable size)
  - Packets “consume” an integer number of blocks

- **Credit Updates (per VL)**
  - Every 64KB or before

- **12 bit fields**

- **Modulo Arithmetic**
  - Max 2048 Credits
    - 128KB at 64B blocks
Failsafe Mechanisms (non-transient failures)

- **Receiver Detected**
  - Buffer Overrun Threshold Exceeded

- **Transmitter Detected**
  - Flow Control Update Monitor

- **Lync Resync**
  - Triggers Initialization of the Credit Accounting
Priorities, Traffic Classes, Queues and Buffers

- **InfiniBand Service Levels (SLs) and Virtual Lanes (VLs)**

  - SL is conceptually equivalent to 802.1 Priority
    - Indicates requested level of service across the InfiniBand L2
    - 16 SLs (15 for data, 1 for management traffic)

  - VL is somewhat equivalent to 802.1 Traffic Class (i.e. Transmit Queue)
    - Number of VLs supported is an implementation choice
    - SL to VL mapping on Transmit
    - Credit Based Flow Control is per VL
      - Prevents HOL Blocking
      - InfiniBand Mandates Separate Receive Buffering Resources per VL
        - Transmitter Queue -> Dedicated Buffer on Receiver
      - Both SL and VL are carried on the packet
        - Required for proper credit accounting
        - 802.1 is different on this regard - Receiver is unaware of Transmitter Queue
Priorities, Traffic Classes, Queues and Buffers (cont’d)

- Ethernet Receiver doesn’t know which Transmitter Queue a frame is coming from
  - No way to allocate dedicated Receiver Buffer for Transmitter Queue
  - Independent Mapping at Receiver of Prio to Buffer

- Can’t do “per-Transmitter-Queue” Credit Based Flow Control

- Per- Priority (Credit Based) Flow Control
  - Much different than Per-Priority Pause (802.1Qbb)
  - May result in Transmitter Queue HOL Blocking

- Solution Space
  - to be discussed on a separate presentation
Thank You