End Station Reaction Points
Which Frames should a Rate Limiter slow?

Caitlin Bestler

Caitlin.bestler@neterion.com

Congestion Notification Message Scope

- **Already limited**
  - Generated based on sampling at CP.
  - Unicast delivery back to a single end station.
- **But the CNM supplies information**
  - It is not a “speeding ticket”
  - Ideally all flows from this end station that reach the congested CP should be throttled
    - But what is realistic?
    - What set of frames should be impacted?

Prior queuing should be Irrelevant

- **End stations have many designs**
  - Specific internal queue structures should neither be rewarded or penalized.
- **Frequently the pre-CNM queue will be too wide**
  - The end station will have had no reason to separate flows based on this destination.
  - Therefore many innocent flows will be slowed.
- **Sometimes the pre-CNM queue will be too narrow**
  - TOE/RDMA per-connection flows that are not the entire output from the end station to the destination.
- **Reaction Points may be created *after* the CNM is received, or it may only identify a *potential* queue.**

Use of Multiple SAs

- **Using Multiple Source Addresses can benefit network utilization when they actually use multiple paths.**
- **But when they hit the same CP, they at best just hog a greater slice of the bandwidth.**
  - The same traffic divided over more flows will be less "dinged" than a single flow would have been.
    - The only escape from this is to make the Source Address irrelevant to the scope of the Rate Limiter created *except* when there is specific reason to believe that Source Address truly will cause the CP to be avoided.
  - Creating an incentive to use *more* Source Addresses in each NIC.

Multiple Queues Can Be Tightly Coupled

- **Multiple source queues can be tightly coupled and have different Source Addresses**
  - Slowing one source will *instantly* cause other flows to increase their output.
  - Within many end stations the scheduler *pulls* “transmit descriptors” or “work requests” to fill the wire capacity.
    - Not the same as independent sources that “push” frames into a set of queues.
    - *Instantly* replacing the output capacity with frames that could be going to the same CP means that the CP will see *no* relief.

Deliberate Cheating Not Required

- **Many legitimate design trade-offs can result in use of more SAs.**
  - QCN should be neutral on these design trade-offs rather than encouraging or forbidding the use of more Source Addresses.
- **Example: Storage Client**
  - VM's use virtual drives. Parent partition is the sole client of the actual storage service.
  - Each VM acts as its own client.
- **Example: HPC**
  - Each rank uses a different VF in a multi-function NIC.
  - All ranks use a single VF.

Which Frames Should be slowed?

- **Ideal would be all frames**
  - From this end station
  - That will hit the same Congestion Point.
- **How close to this ideal be achieved with realistic real-time decision making?**
- **Initial assumptions:**
  - Different Priority, probably a different CP
  - Different VID+DA: probably a different CP
    - But maybe not for “next hop” CPs.
  - Different SA: probably the same CPs
    - Unless the SA selects a different egress port.

L2 Flows that SHOULD NOT be impacted

- **Different Priority**
- **Different Destination End Station**
  - Which should be presumed if VID + DA is unique.
    - Not feasible to know remote VID to FID mapping.
    - Not feasible to know when multiple remote DAs are really the same end station.
  - Different non-aggregated egress port
    - If the first hop is a different non-aggregated port then it is reasonable to assume different CPs will be hit.
      - At least until reaching the final destination.

L2 Flows that SHOULD be impacted

- **Same egress Port**
- **Same priority**
- **Same Destination VID+DA**
- **Rationale:**
  - Other factors such as SA or L3/L4 headers are unlikely to have an impact on whether the same CP will be hit when they do not impact the egress port on the first hop.
  - Merely creating more SAs will *appear* to improve congestion robustness *locally* by *stealing* bandwidth.
  - Require actual knowledge of specific multi-pathing to justify NOT including the flows.

Possible special cases

- **When the CP is the last funnel before the destination then multi-pathing will not avoid it.**
  - Could be inferred by comparing CP's MAC Address with Destination.
  - Could be a boolean flag in the CNM.
- **When the CP is on the first hop**
  - End station could learn first hop on each port, and apply the Rate Limiter more broadly.
  - Alternate: CPs could be explicitly allowed to increase sampling rate on ports they know connect directly to end stations.

Special Cases Unlikely to Justify Special Effort

- **Same Egress Port, Same DA, Same Priority**
  - But interior CPs distribute traffic based on SA or L3/L4 headers.
  - When this happens then *some* false head-of-line blocking will occur for frames that would really have missed the congested CP.
  - But far more often the SA/L3/L4 will not change the CP, but merely evade the Rate Limiter. Traffic will *instantly* divert to the flows that vary of SA/L3/L4 and the CP will see *no* relief.
- **Different Everything, but same internal CP**
  - Using a link-state databases (from Shortest Path Bridging or TRILL) this case *could* be identified.
  - But even if the date exists it is unlikely to be organized to allow a quick test of “would this frame go to this CP”?
  - Why penalize an end station for having a link-state database available?