End Station Reaction Points
Which Frames should a Rate Limiter slow?
June 2008 – Rev 3

Caitlin Bestler

Caitlin.bestler@neterion.com

End Station Output Queues

- **End Station Output Queues reflect many different design approaches:**
  - L2-only service, Offload/L4-L5 service
    - And mixtures thereof
  - Multiple physical and/or virtual ports
  - Where memory lives
    - On-chip
    - External memory owned by chip
    - Host memory
  - Throughput versus latency tradeoff
- **QCN models must be vetted for a large range of potential end station designs.**

First Issue: Congestion Notification Message Scope

- **When an end station gets a CNM, which L2 flows should be rate limited?**
- **The CNM is already limited in scope**
  - 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.
- **Rate limited queues may be created *after* the CNM is received, the pre-CNM queue may fix relevant and irrelevant flows.**

Reaction Point IDs (au-nfinn-RPID)

- **A Reaction Point ID (as proposed in au-nfinn-RPID)**
  - could identify an queue (or set of queues)
  - or merely flows that *could* be queued separately.
- **Multiple Queues for one Reaction Point ID**
  - Multiple offloaded connections with the same L2 source/destinations.
  - Separate queue may only last for duration of offload, and therefore should not have a distinct RPID.
- **Multiple RPIDs single queue**
  - RPIDs are not yet rate limited, and a single queue simplifies the host/NIC interface.
  - RPIDs are rate limited, but co-mingled with similar Rate Limiters to minimize resources with minimal head-of-line blocking.

Use of Multiple SAs

- **RPIDs allow full utilization of fabric multi-pathing without artificially creating new Sources.**
- **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.
  - We should avoid 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. The “round trip time” is zero.
  - 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 that:**
  - Are from this end station
  - 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 different RPID is used.

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

- **Full match on:**
  - Egress Port
  - Priority
  - 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.

Reasonable Number of RPIDs

- **Explicit RPIDs would allow limiting each end station to a reasonable quota of RPIDs for flows targeted to a given DA at a given VLAN Priority**
  - A modest number of RPIDs is enough to take advantage of fabric provided multipathing.
  - The only use for more RPIDs is to evade CNMs by micro-fragmenting the end station's traffic.
    - This should be explicitly forbidden.
    - But CPs would not be expected to enforce this.
- **To be done: define what a "reasonable" number is**
  - And whether it is a constant or a result of fabric discovery.

Split Reaction Points

- **End Station may have special purpose Output Queues that have a narrower scope than desired for a Rate Limiter.**
  - Primary example: Send Queues for TOE/RDMA.
- **For some designs the output from these queues would not naturally flow past general purpose Rate Limiters.**
- **Proposed solution: allow "split Rate Limiters" to be created on multiple internal queues in response to a single Congestion Notification Message**

End Station Congestion Points

- **Not the topic of this presentation, but…**
- **End Station Congestion Points are *NOT* necessarily the inverse of its Reaction Points.**
- **For multi-function devices, the CPs are likely VF (Virtual Function) dependent.**
  - VID + DA determines VF, but multiple indexes could yield the same VF.
  - This is frequently a “default” VF for unknown addresses.
- **Having VF sensitive QCN triggers is desirable to limit inbound traffic based on VF.**

End Station Host Stack Participation

End Station Stack Must Participate

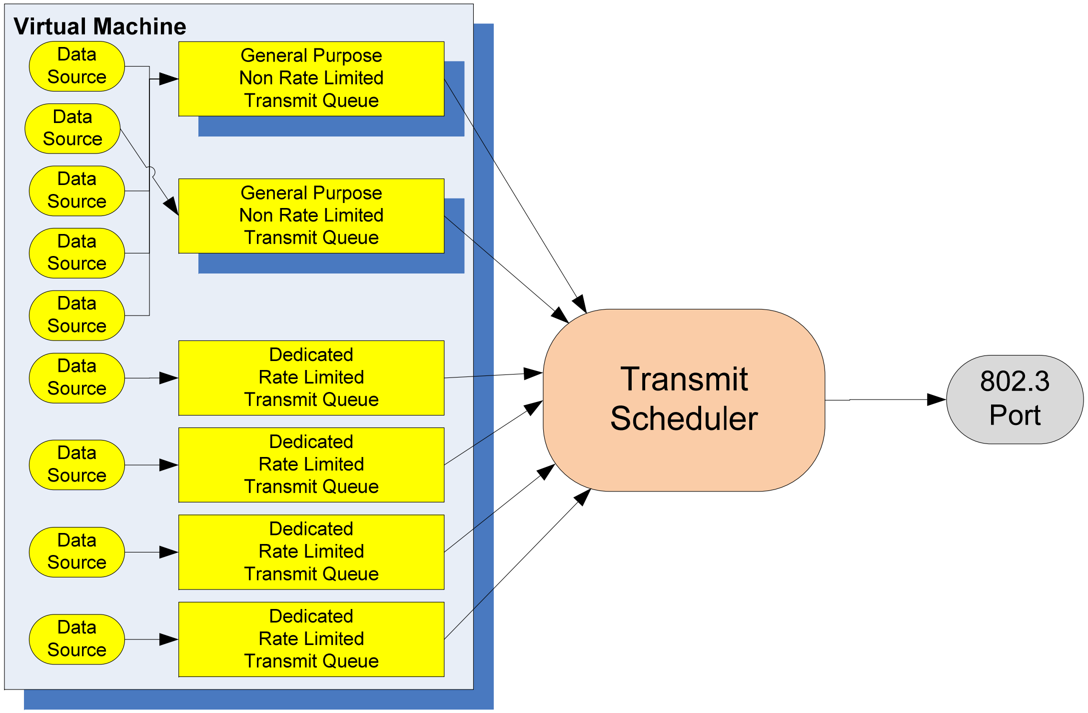
- **When a flow is rate limited the source must ultimately be slowed to match.**
- **With connection-specific RDMA style interfaces this is just a matter of not completing Send Work Requests.**
- **But existing IP stacks generally use a limited number of queues into a given L2 device.**
- **Possible results:**
  - Head of line blocking: a pause on one L2 flow will impact all traffic for the same Priority, whether to the same destination or not.
  - Buffer Drain: to avoid head-of-line blocking the driver will attempt to put rate limited frames in a side-queue.
    - Even if stack supports out-of-order completion, it will result in memory pressure.
    - Worst case: memory pressure causes swap out – to network storage that is reached via the problem Congestion Point.

Method of Participation may vary

- **QCN feedback to L4**
  - Any L4 socket that is impacted by a Rate Limiter is told of the rate limit in L3/L4 terms. It adjusts it's L4 congestion window accordingly.
- **Directed Queuing**
  - L2 driver informs its client that a specific flow should be placed in a distinct input queue.
- **Directed Pausing**
  - L2 driver informs its client that a specific submission cannot not be accepted at this time. The same frame should not be retried until a specified time (or callback). The source socket should block, but not any others.

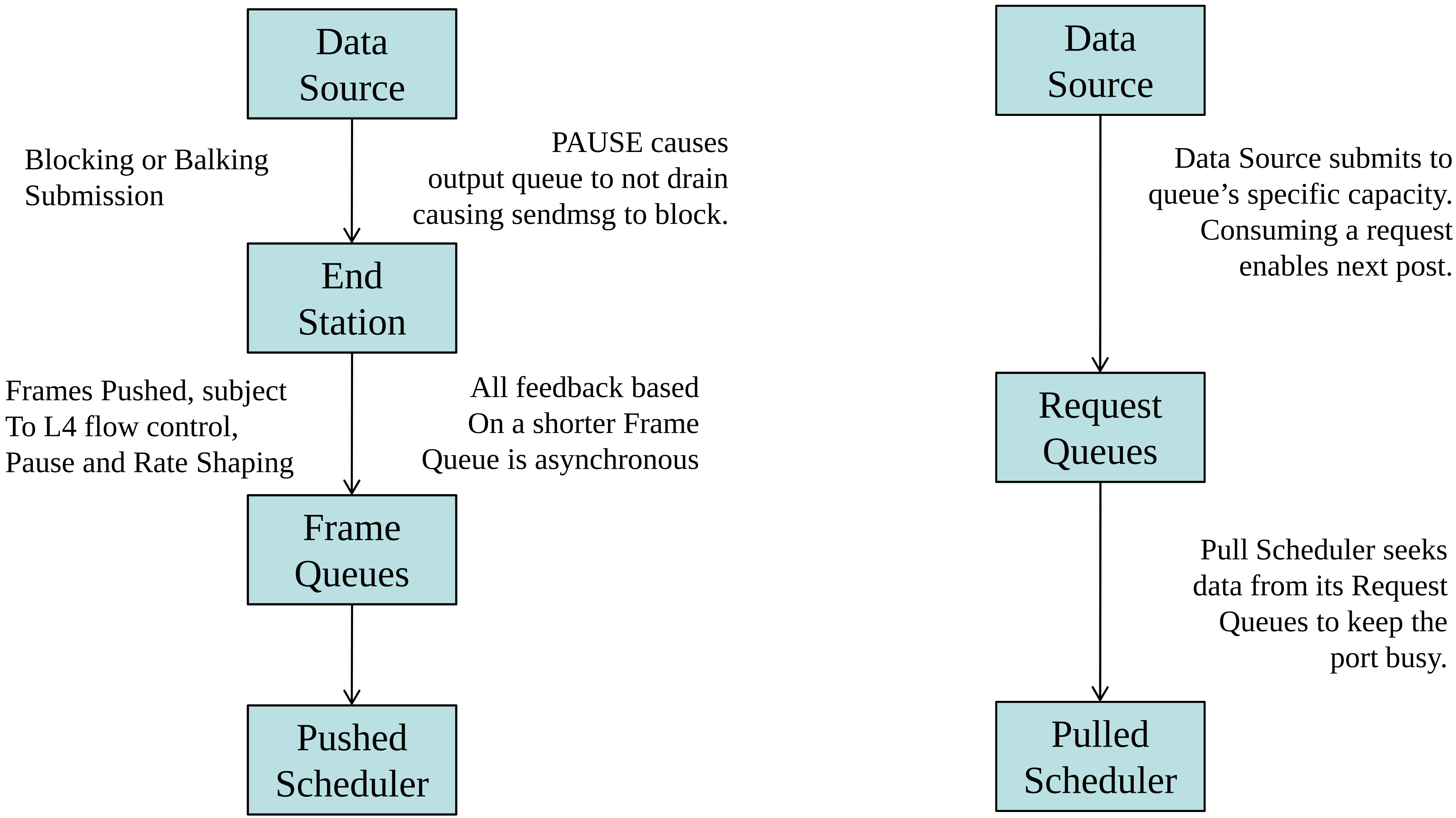
Some Background Information on End Station Design Issues

Hypothetical Multi-function NIC with all Qau, Qaz and Qbb support

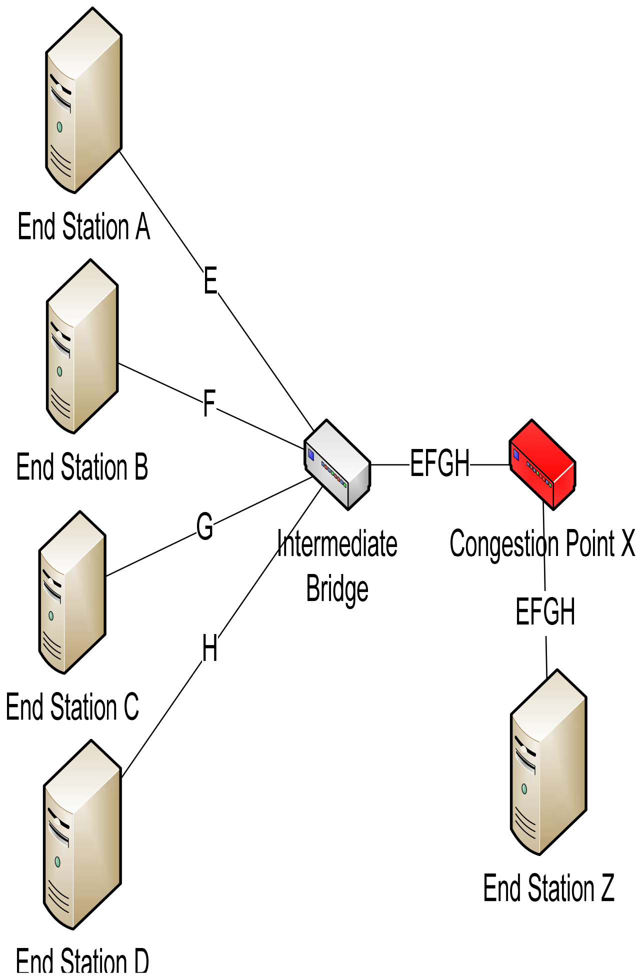


- **Most Data Sources feed general purpose transmit queues that are not rate limited.**
- **Data Sources may be diverted to dynamically allocated rate limited transmit queues**
- **Data Sources may have dedicated Transmit Queues which are optionally Rate Limited (RDMA/TOE/iSCSI).**
- **Each Transmit Queue is for**
  - Single Virtual NIC
  - Single Traffic Class
- **Each PCB priority applies to set of transmit queues.**
- **Each Transmit Queue is accounted for by one ETS priority.**
- **Additional weighted round robin likely applies to each VNIC.**

Push vs Pull

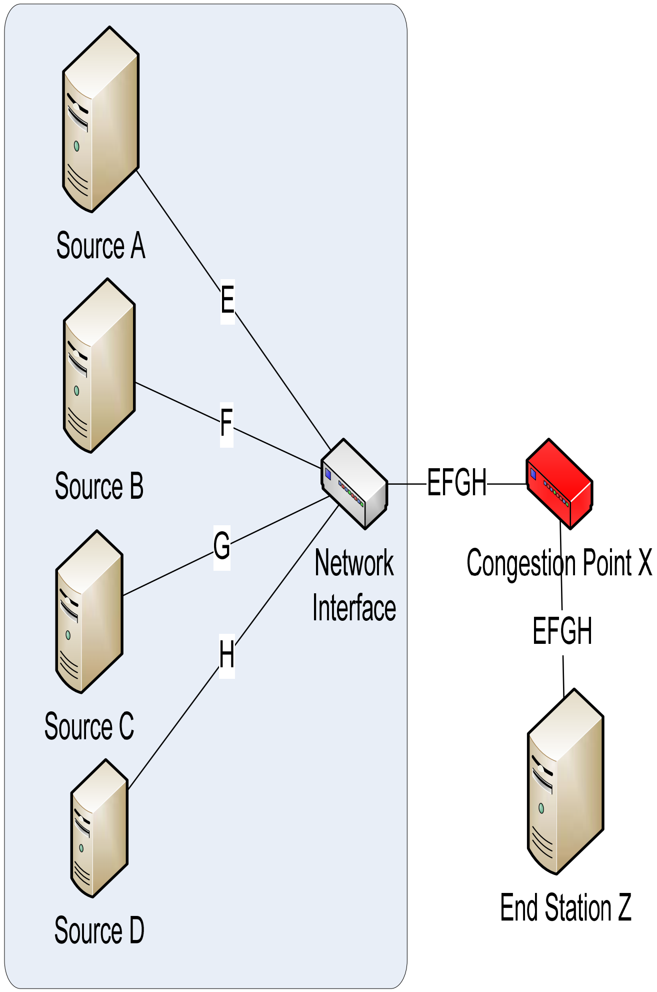

Data Source
Blocking or Balking Submission
PAUSE causes output queue to not drain causing sendmsg to block.
End Station
Frames Pushed, subject To L4 flow control, Pause and Rate Shaping
All feedback based On a shorter Frame Queue is asynchronous
Frame Queues
Pushed Scheduler
Data Source
Data Source submits to queue's specific capacity. Consuming a request enables next post.
Request Queues
Pull Scheduler seeks data from its Request Queues to keep the port busy.
Pulled Scheduler

Data Sources on Separate End Stations



- **Flows E,F,G and H to End Station Z all reach CP X in the Red Bridge.**
- **CP X has sent CNM for Flow E to End Station A.**
- **A reduce E's rate.**
- **Queues in the Intermediate Bridge are drained more rapidly because E's rate is reduced.**
  - Immediate reduction in aggregate flow to CP X is unlikely, but there is an immediate drop in the ingress rate (because E is reduced).
  - Draining of queues on the Intermediate Bridge will result in fewer PAUSEs to End Stations B, C and D.
  - Eventually this will cause F, G and H to speed up unless they get a CNM. Reducing the ingress rate reduction.
  - But it will not be immediate.

Data Sources on Single End Stations



- **Flows E,F,G and H to End Station Z all reach CP X in the Red Bridge.**
- **CP X has sent CNM for Flow E to Source Address used for flow E.**
- **Minimally scoped Rate Limiter:**
  - Only Source A reduces it's rate.
  - Network Interface, seeking to feed a hungry port, increases the rate at which it transmits from B, C and D.
  - There is no immediate reduction in the aggregate flow to CP X.
  - There is no reduction in the ingress to the network of frames destined for CP X.
  - There will be no reduction until all sources on the End Station have received a CNM.
- **End Station scoped Rate Limiter**
  - E,F,G and H are all reduced in response to the first Rate Limiter.