# Transport Mechanisms for the Data Center

## IEEE Data Center Standards

**Abdul Kabbani, Mohammad Alizadeh, Berk Atikoglu, Prof. Balaji Prabhakar**

Department of Electrical Engineering, Stanford University

## Data Center Transport: Opportunities and Challenges

- Data Centers see the true convergence of L3 and L2 transport
  - L3 or L2 transport mechanisms can be used almost interchangeably
- But there are "cracks" between the layers
  - Deficiencies exist in each layer which not even their union addresses

## L3 Transport: TCP/IP

- Design assumptions: Make ***no*** assumptions!
  - Arbitrary network sizes (1us---1s)
  - Wide range of link bandwidths (10s of Kbps---10s of Gbps);
  - Wide range of flow sizes (10s of Bytes to GBs)
  - Discover operating scenario gradually and optimize (reduce Min-RTO, etc)
- Consequences
  - ✓ No restrictions on deployability
  - ✓ Gradually optimal performance for large flows
  - – Per-flow state at sender and receiver: defeats hardwarization---think of TOEs
  - – Poor performance for short flows (e.g. inter-process comm., short memory/disk accesses)

## L2 Transport: IEEE 802.1

- IEEE 802.1 Data Center Bridging standards: Enhancements to Ethernet
  - Reliable delivery (802.1Qbb): Link-level flow control (PAUSE) prevents congestion drops
  - Congestion management (802.1Qau): End-to-end Ethernet congestion control, prevents congestion spreading due to PAUSE
- Consequences
  - ✓ Hardware-friendly algorithms: can operate on 10—100Gbps links
  - ✓ Partial offload of CPU: no packet retransmissions
  - – Corruption losses require abort/restart; 10G over copper uses short cables to keep low BER
  - – PAUSE absorption buffers: proportional to bandwidth x delay of links, high memory bandwidth

## L2 Challenges

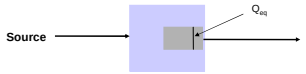
1. No per-packet acks in Ethernet, unlike in the Internet
   - Not possible to know round trip time!
   - So congestion must be signaled to the source by switches
   - Algorithm not automatically self-clocked (like TCP)
2. Links can be paused; i.e. packets may not be dropped
3. No sequence numbering of L2 packets
4. Sources do not start transmission gently (like TCP slow-start); they can potentially come on at the full line rate of 10Gbps
5. Ethernet switch buffers are much smaller than router buffers (100s of KBs vs 100s of MBs)
6. Most importantly, algorithm should be simple enough to be implemented completely in hardware

## Our Solution

- Decouple reliable delivery from congestion management:
  - QCN for Congestion Management (IEEE 802.1Qau)
  - L2.5 for Reliable Delivery

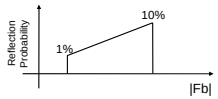
## Quantized Congestion Notification: Congestion Management

- Consider the single-source, single-switch loop below

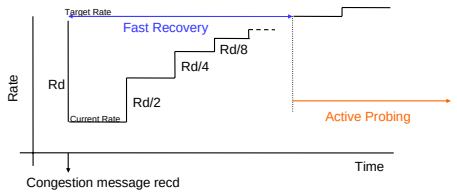



- **Congestion Point (Switch) Dynamics:** Sample packets, compute feedback (Fb), quantize Fb to 6 bits, and reflect only *negative* Fb values back to Reaction Point with a probability proportional to Fb.

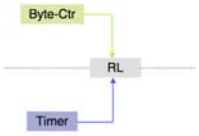
$$Fb = -(Q - Q_{eq} + w \cdot dQ/dt)$$
$$= -(\text{queue offset} + w.\text{rate offset})$$

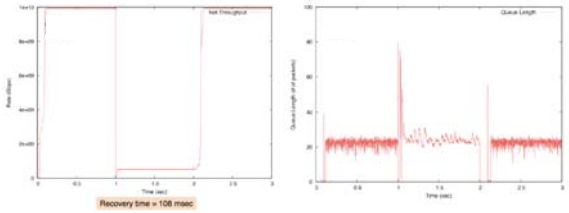



- **Source (reaction point):** Transmit regular Ethernet frames. When congestion message arrives:
  - Multiplicative Decrease: $R \leftarrow R(1 - G_d F_b)$
  - Fast Recovery similar to BIC-TCP: gives high performance in high bandwidth-delay product networks, while being very simple.
  - Active Probing




  - Rate increase (Fast Recovery/Active Probing) is triggered by two mechanisms:
    - Byte Counting
    - Timer



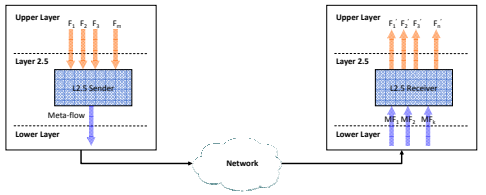




Single Link, 10 Sources, RTT = 200 µs

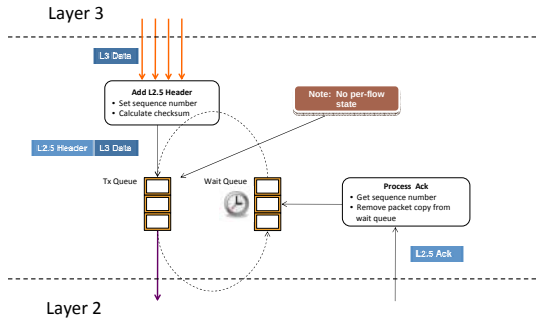
## L2.5: Reliable Delivery

- A shim layer between L3 and L2; uses the main observation:
  - The RTT in Data Centers is small, typically less than 500 us
  - The link speeds are high and pretty uniform, so there isn't much congestion
  - Therefore, amount of data from a 10G source "in flight" is less than 750 Kbytes

➡ Collapse all source flows into a L2.5 "Meta-flow", store in flight packets in a single queue, retransmit dropped/corrupted packets




## L2.5 Sender Block Diagram




## Consequences & Further Work

- Consequences
  - ✓ Reliable packet delivery at high-speeds (for all flow sizes) across Data Center
  - ✓ Reduced latency for flow transfers
  - ✓ Gets rid of short cabling requirement, PAUSE absorption buffering
  - ✓ Combines well with QCN to give good packet transport solution
  - ✓ As shim layer, completely transparent to L3 and L2: incrementally deployable
  - – Note: Uses the assumption that Data Center is a homogenous environment
- Lots of further work
  - Storage (FCoE, RAMCloud, …)
  - TCP Offload Engines; crossing IP boundaries
  - Short latency makes it interesting for real-time traffic: Video