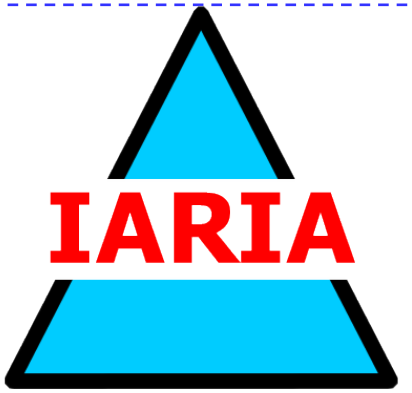
ECN Works Better for Selective Dropping in High Bandwidth-Delay Product Connections

Mahmoud Bahnasy, Ali Munir, Junjie Niu, Zhibo Yan, and Peng Dong, and Yashar Ganjali



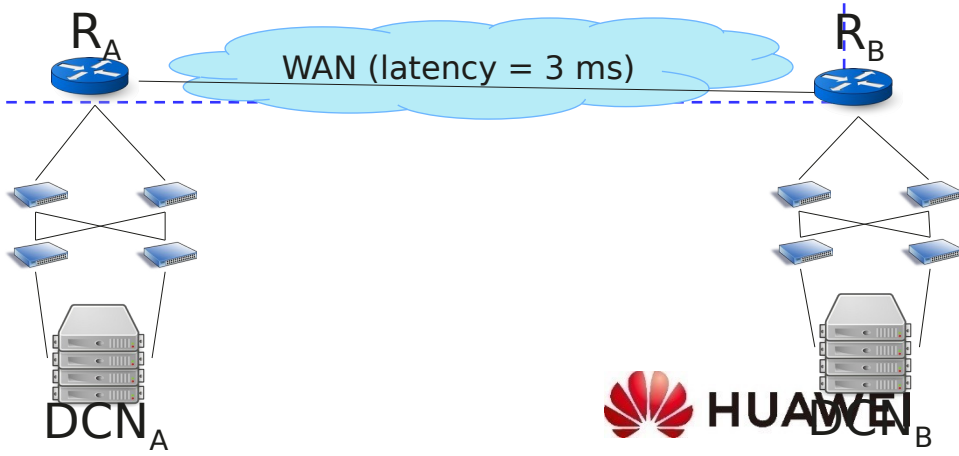
The Twenty-Fourth International Conference on Networks

ICN 2025

May 18, 2025 to May 22, 2025 - Nice, France

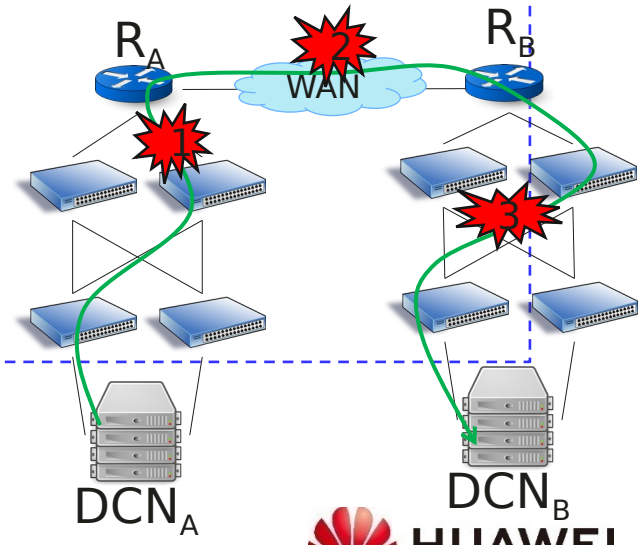
Background and Prior Art of the Idea

- Modern Internet services and applications distribute data and compute on different geographically distributed data centers.
  - Example, terrestrial observatories, remote telescopes.
- To handle congestion control in either DCN or WAN, existing solutions have leveraged either ECN (e.g., DCTCP and DCQCN) or delay (e.g., Vegas, TIMELY and BBR) as the congestion signal.
- However, using either method does not guarantee zero packet loss.
- Reason 1: Traditional TCP Reaction (Waiting for 3 duplicate acks or the RTO to expire) is too costly.
  - Hence faster packet-loss recovery or retransmission is mandatory.
- Reason 2: When congestion happens inside DCN, it takes the WAN traffic 1000x of DCN RTTs to react.
- Therefore, WAN traffic creates long queues due to its slow reaction time.
- And, DCI traffic greatly affects DCN as it reacts solely to congestion.



Background and Prior Art of the Idea

- For Inter-Data traffic, congestion could happen in three main locations:
  1. Inside the source DCN.
  2. At the edge router of DCN_A towards the DCI link (uncontrolled link).
  3. Inside the destination DCN.
- In this paper, we present a solution that provides faster packet-loss recovery when congestion happens at point 2 and 3.
- Characteristic of operational connection.
  - Very high latency (100s of ms), hence, the reaction mechanism is very slow.
  - Heterogeneous traffic (DCI and DCN), hence, two different reactions must be adopted.
  - No explicit feedback signal is available.
  - SLA provides stochastic guarantees regarding packet loss and delays for DCI traffic.
  - Statistical multiplexing is used to increase network utilization and network providers' profit.



Technical Solution:

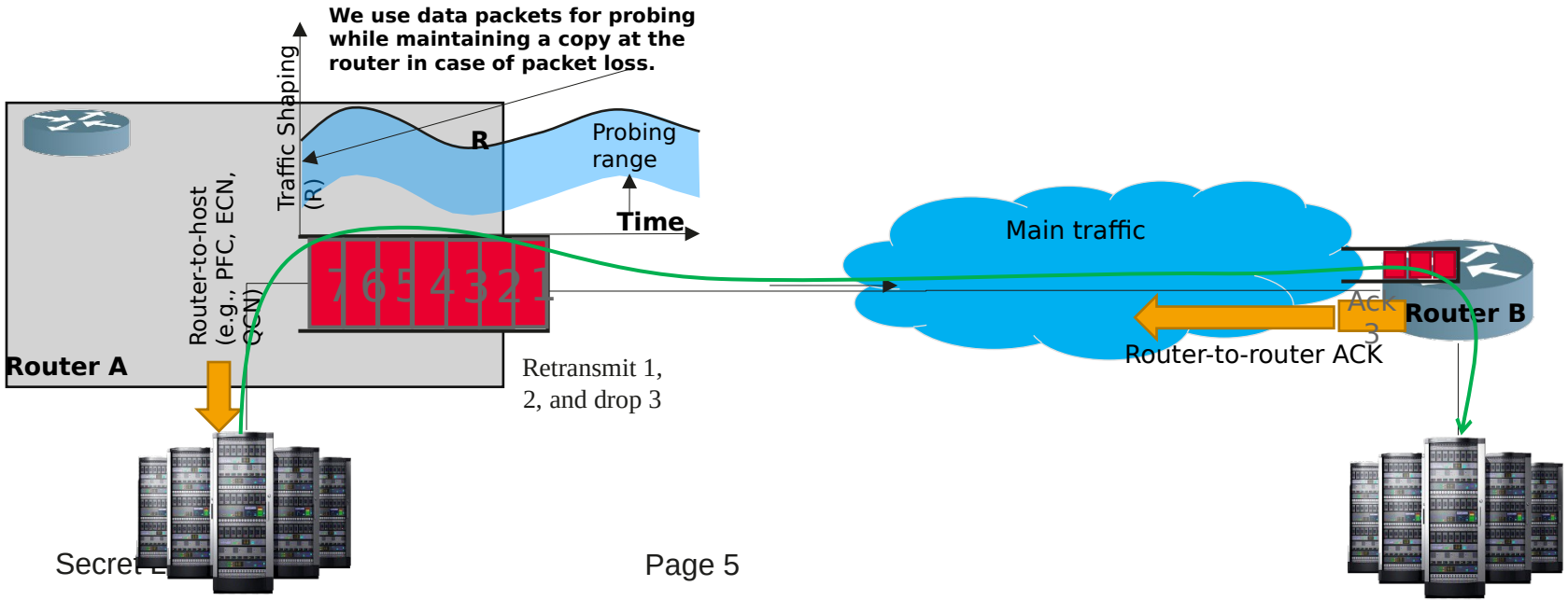
Design consideration:

- *The main idea is to introduce a new reaction to packet loss in the WAN without triggering host reaction.*
- *Packet loss recovery takes place at the edge router faster while preventing TCP conservative behavior.*
- *Our solution is be designed as a plugable to existing TCP protocols.*
- *In case of severe congestion, the source hosts falls back to TCP traditional behavior.*
- *Hence, we provide a faster recovery for packet loss for DCN and DCI connections, while keeping the control loop very short (between the edge routers)*
- *Such mechanism aims to resolve transient congestion.*

Inter-Data Center Bridging:

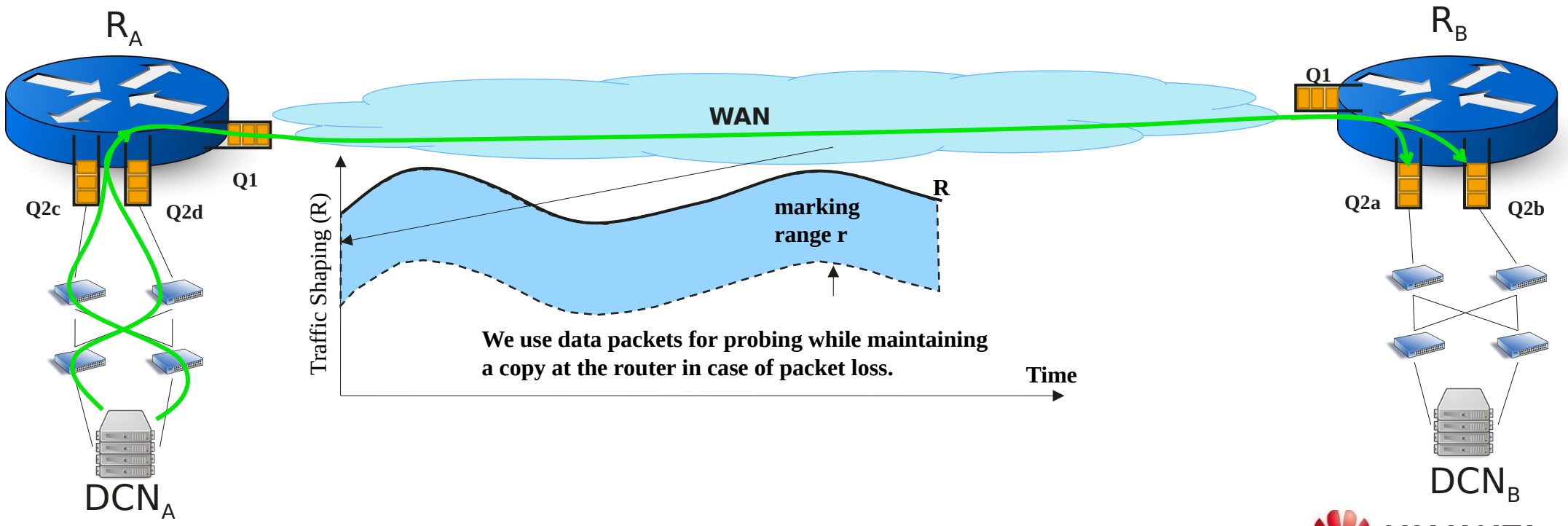
Details of the main idea:

1. Router A occasionally marks few packets for potential drop in case of congestion (ECN-disabled).
2. At the same time it clones these packets and saves them in a local buffer.
3. For each of these marked packets, router B must send an ACK.
4. Router A, deletes/drops packets appointed for received ACK, and retransmit packets with no ACK.
   - Hence, it allows the hosts to maintain high CWND and prevent CWND reduction triggered by packet loss or DupAcks.
5. In case of severe packet loss or congestion, the system fall back to the traditional TCP reaction.
6. Another option is to use a flow control signal (e.g., PFC or QCN) between Router A and local senders (Future work).

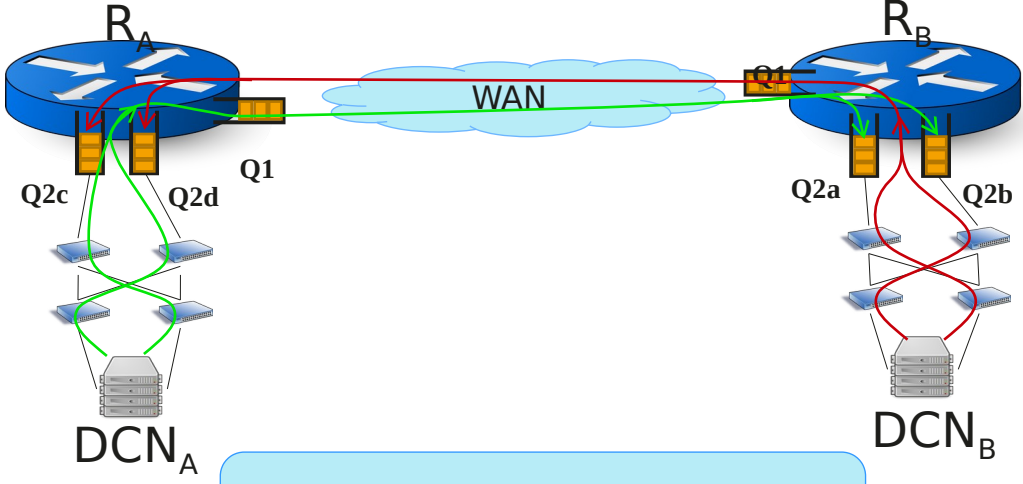


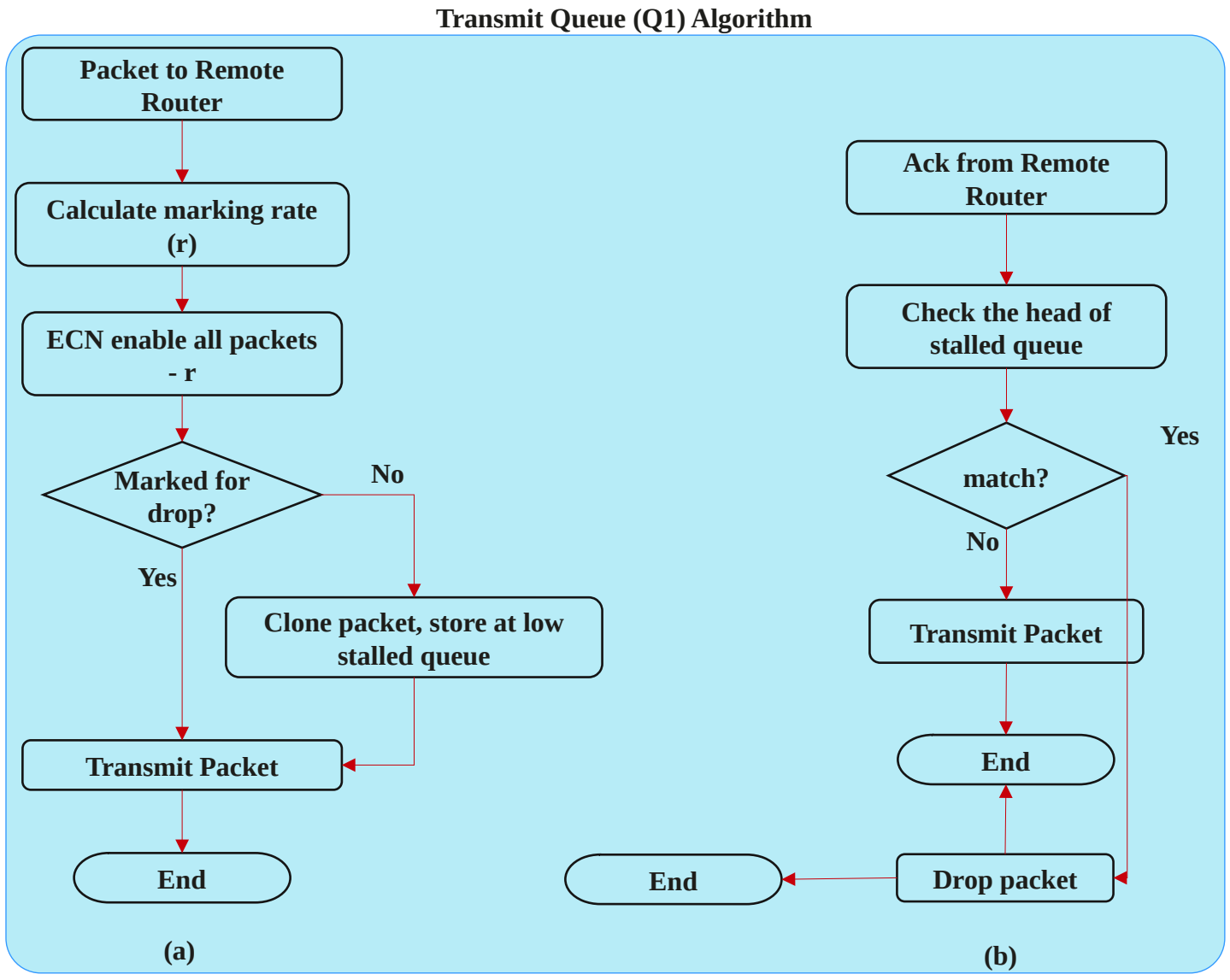
Inter-Data Center Bridging:

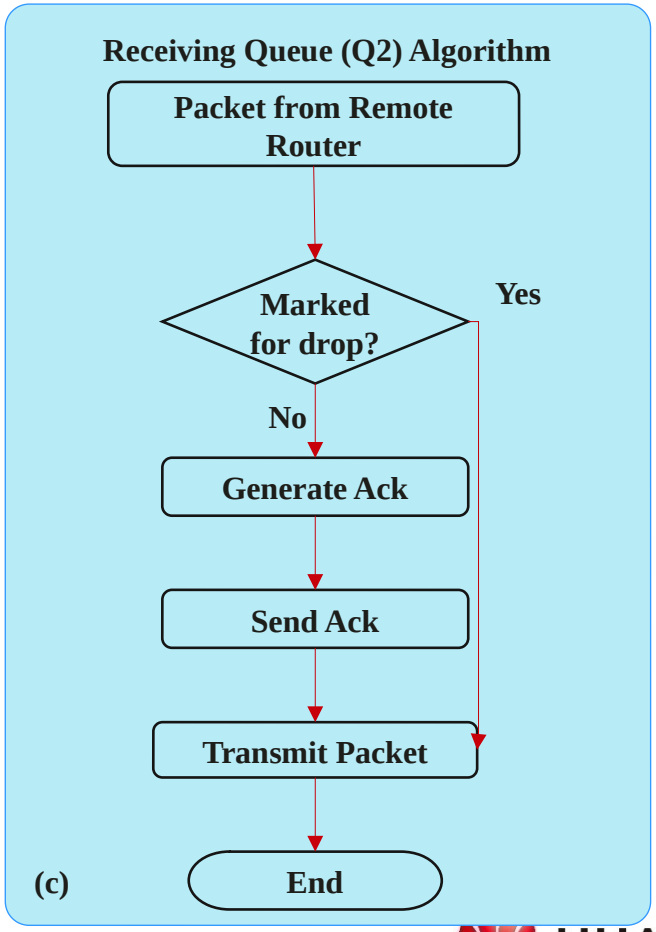
- The figure below shows the architecture of the Queue disciplines installed at the two edge routers for this invention to work correctly.
- We have two types of queue disciplines; namely transmitter Queue (Q1) and Receiving Queue (Q2).
- We assume that routers in WAN drop ECN-disabled packets in case of congestion (could be specified in SLA agreement).
- All packet marking happens below the rate R (i.e. [R, R-r] as shown in the figure below).



Inter-Data Center Bridging

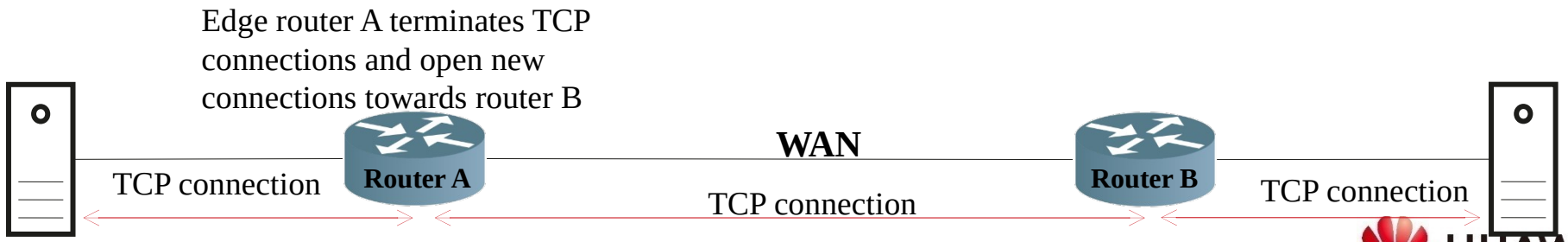






Inter-Data Center Bridging VS. TCP termination at Edge routers

- Edge router A terminates TCP connections and open new connections towards router B.
- In such an approach, hosts do not need to wait for the end-to-end ACK to control transmission.
- However, Edge routers need, not only to store all inflight packets but also need to keep the state for each connection both to the hosts and to the other edge routers.
- In addition, managing these connections encounters more delay in processing the whole TCP stack.



Simulation Results:

Experiment I:

- 15 long-lived flows from all hosts of DCN A towards DCN B.
- We artificially drop 5% of data packets.
- Cooperative WAN and drop 50% of the 10% ECN-disabled traffic only (i.e., 5%).
  - DCIB enhances the performance 6x and 4.5x for TCP Reno and TCP Cubic respectively.

- Non-Cooperative WAN (drop packets regardless of the marking).
  - DCIB enhances the performance 4x and 3.24x for TCP Reno and TCP Cubic respectively.

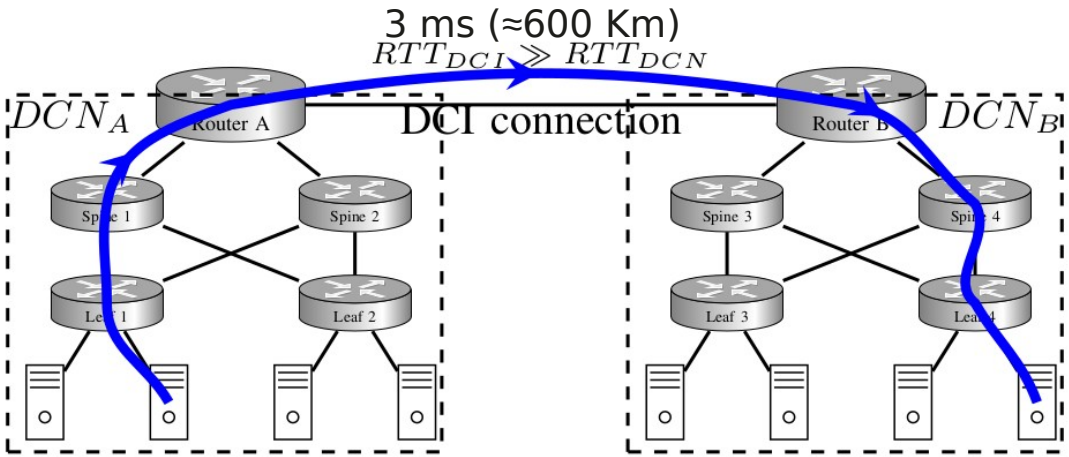


Figure 1. A typical DCI topology

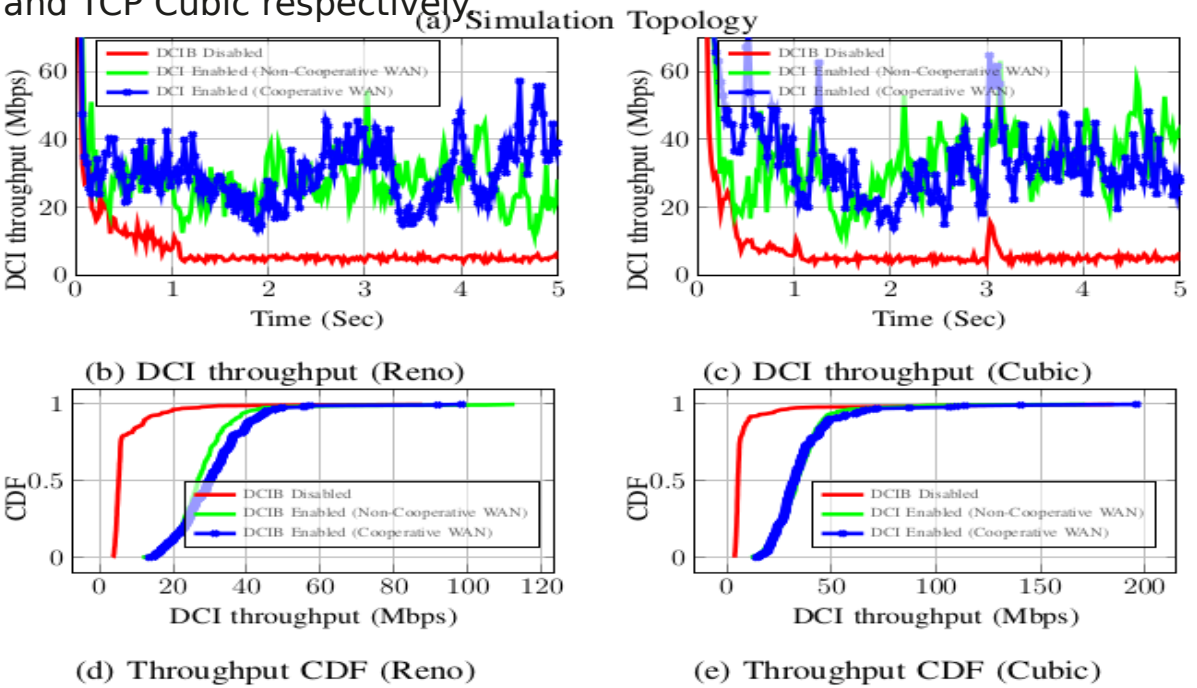


Figure 10. DCIB increases the average throughput by 4x and 6x when packet drop occurs in all packets or in selected packets only.

Simulation Results:

Experiment II:

- Experiments with different RTT starting from 10 ms (2000 Km) up to 30 ms (6000 Km).
- DCIB enhances the average throughput of TCP Cubic and TCP BBR by 5x and 7x, respectively.
- DCIB g

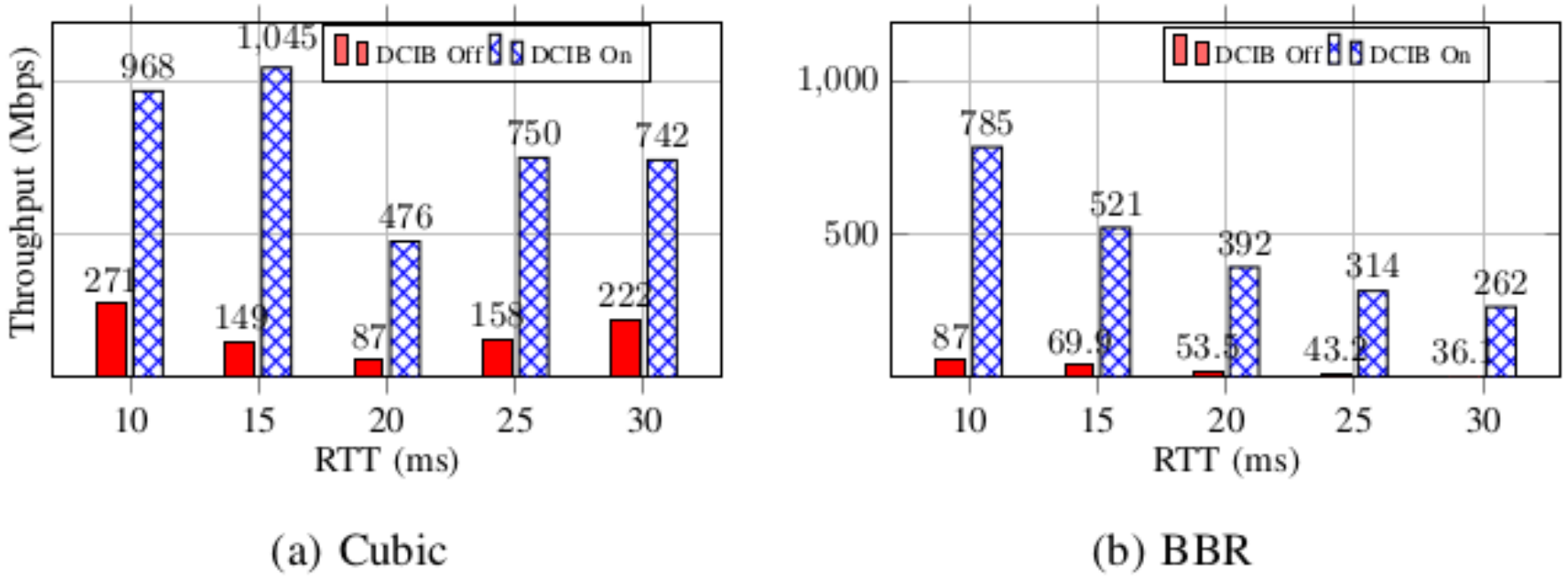


Figure 12. 5x and 7x higher throughput compared to TCP Cubic and BBR.

Simulation Results:

Experiment III:

- We simulate using realistic workload (1000 flows using web search workload).
- The average load is 80%.
- The base TCP variant used in this simulation is TCP Cubic.
- The Figure below illustrates that DCIB reduces Flow Completion Time (FCT) by up to 88%.

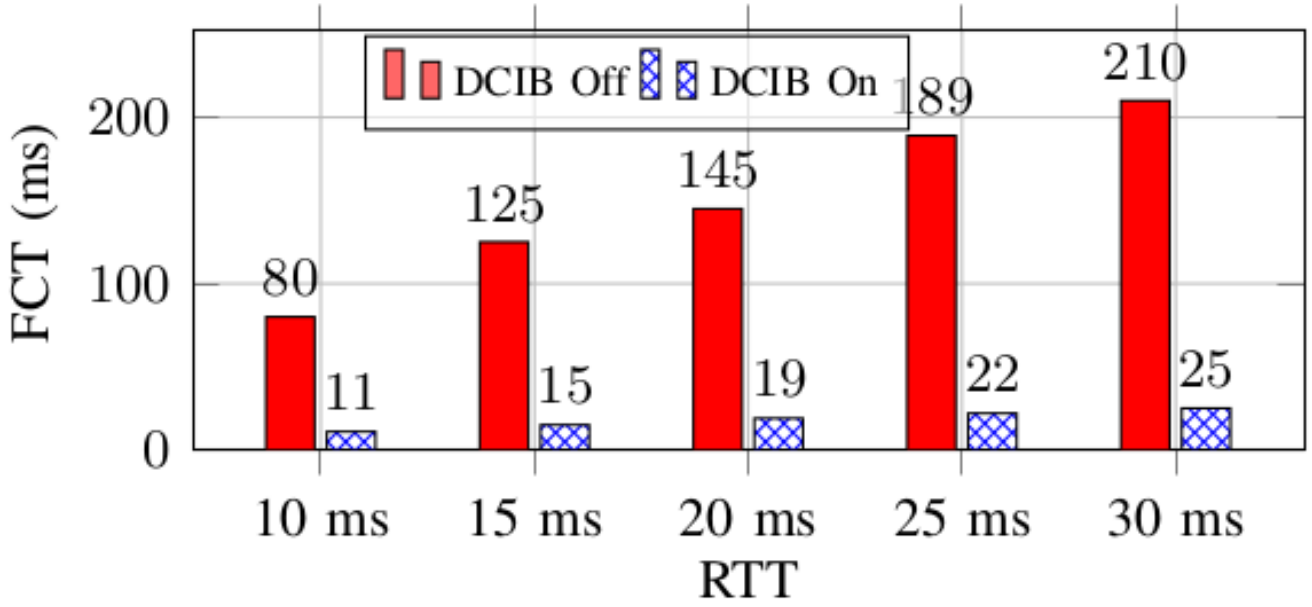


Figure 13. DCIB with stochastic retrans (FCT reduced by 88%)

Conclusion and Future Work:

Conclusion:

- We present DCIB as a plugable for TCP to enhance the performance of inter data center communication.
- DCIB uses ECN marking to selectively drop certain packets instead of sending congestion notification.
- This allows intermediate routers to drop recommended packets that can easily be recovered.
- Hence, we avoid causing any TCP severe reaction at the host.
- Our experimental results show that DCIB can increase the DCI throughput by a factor of 6x , 4x, and 7x for both TCP Reno, Cubic, and BBR, respectively.

Future Work:

- Defining an interaction between DCIB and the TCP at the host should allow better performance.

Thank you.

Bring digital to every person, home and organization for a fully connected, intelligent world.

Copyright © 2022 Huawei Technologies Co., Ltd.
All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.

HUAWEI