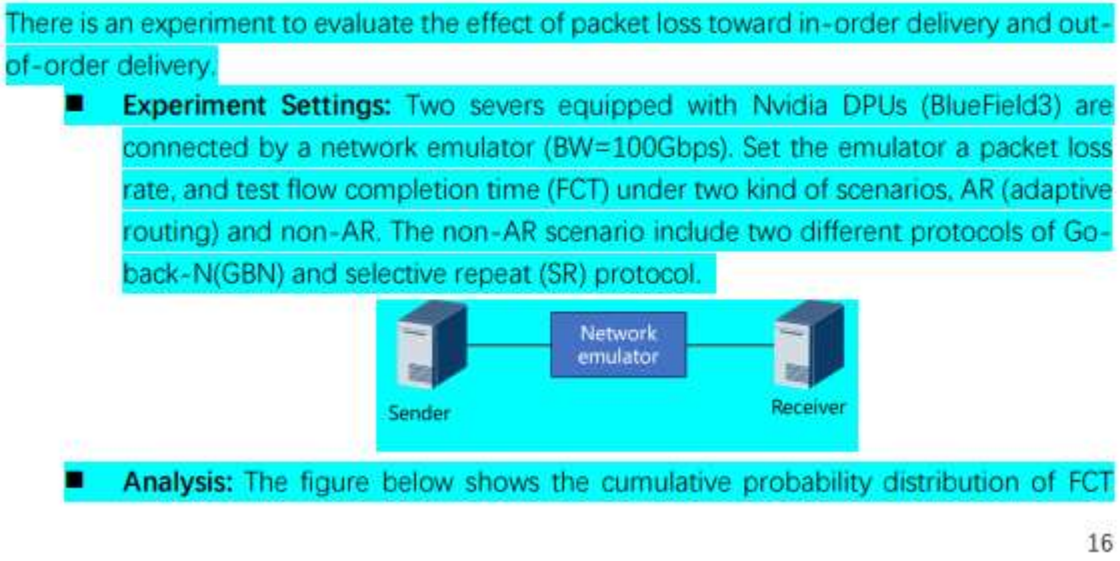
The challenges of per-packet Load Balancing in AICN

Jieyu Li (China Mobile)

2024.09

Purpose

- About the part of load balancing challenges in AICN study item draft report[1], one major comment is that it’s inappropriate to put a unpublished experiment data into the report.



There is an experiment to evaluate the effect of packet loss toward in-order delivery and out-of-order delivery.

- **Experiment Settings:** Two severs equipped with Nvidia DPUs (BlueField3) are connected by a network emulator (BW=100Gbps). Set the emulator a packet loss rate, and test flow completion time (FCT) under two kind of scenarios, AR (adaptive routing) and non-AR. The non-AR scenario include two different protocols of Go-back-N(GBN) and selective repeat (SR) protocol.

- **Analysis:** The figure below shows the cumulative probability distribution of FCT

16

- This contribution intent to give a discussion about the related problem and experiment.

[1] https://mentor.ieee.org/802.1/dcn/24/1-24-0028-04-ICne-aicn-report-draft.pdf

Background

- **Traditional ECMP-based per-flow load balancing solutions perform poorly in AICN**
  - Severe hash collision due to the low entropy and high bandwidth AI traffic.
- **Per-packet LB solution is widely considered as the technology trend to avoid per-flow LB's drawbacks for AI network**
- **Take further insights on the challenges of per-packet LB**
  - The main side-effect of per-packet LB is causing packets of a flow arriving at receiver out of order, and the change from network in-order to out-of-order delivery makes some troubles:
    - Re-ordering
    - Reliability problem: loss packet recovery

    ….
- **This contribution mainly discuss the loss packet recovery problem under network out-of-order delivery.**

Packet Loss Recovery

- **Packet loss is inevitable, even in lossless RDMA network**:
  - *Queue overflow, caused by congestion.*
  - Packet corruption, caused by bit error.
  - Silent packet loss, caused by some silent faults in switch/router.
- **How to recover loss packet?**
  - Link-level retransmission, not supported in DC ethernet yet.
  - End-to-end level retransmission, supported by RDMA NIC.
- **In commodity RDMA NIC, there are two general methods to trigger packet retransmission[1]**:

  a) Receive out-of-order packets at the receiver.
  - Network provide in-order delivery.
  - Go-back-N, and Selective Retransmission protocol.

  b) Wait for a timeout to expire at the sender[2].
  - Network don’t need provide in-order delivery.
  - Per-packet adaptive routing.

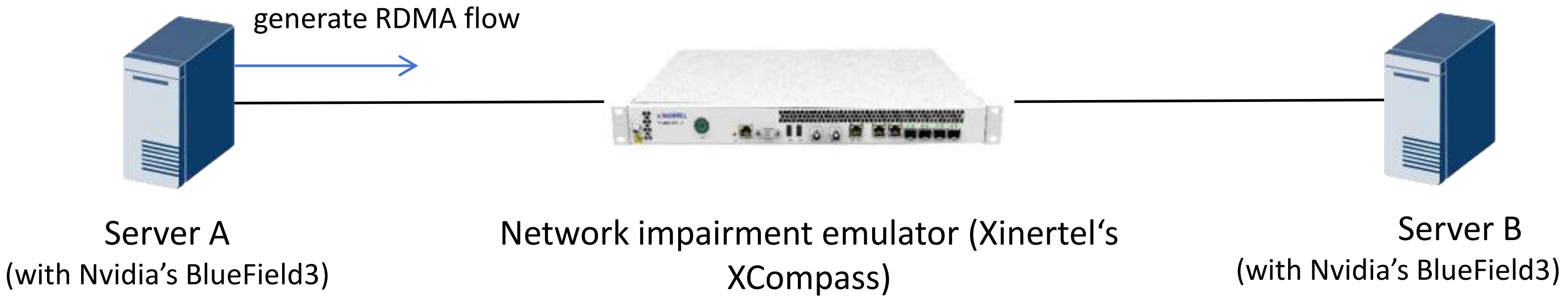
Higher recovery time

- **In per-packet Load balancing, if network no longer provide in-order delivery, RNIC can only rely on timeout mechanism to recover loss packet[2].**

[1] Gao Y X, Tian C, Chen W, et al. Analyzing and Optimizing Packet Corruption in RDMA Network[J]. Journal of Computer Science and Technology, 2022, 37(4): 743-762.
[2] Hoefler T, Roweth D, Underwood K, et al. Datacenter ethernet and rdma: Issues at hyperscale[J]. arXiv preprint arXiv:2302.03337, 2023.

Experiment settings

- **To verify the effect of packet loss under out-of-order delivery, compared with in-order delivery.**


generate RDMA flow
Server A
(with Nvidia's BlueField3)
Network impairment emulator (Xinertel's
XCompass)
Server B
(with Nvidia's BlueField3)

Topology

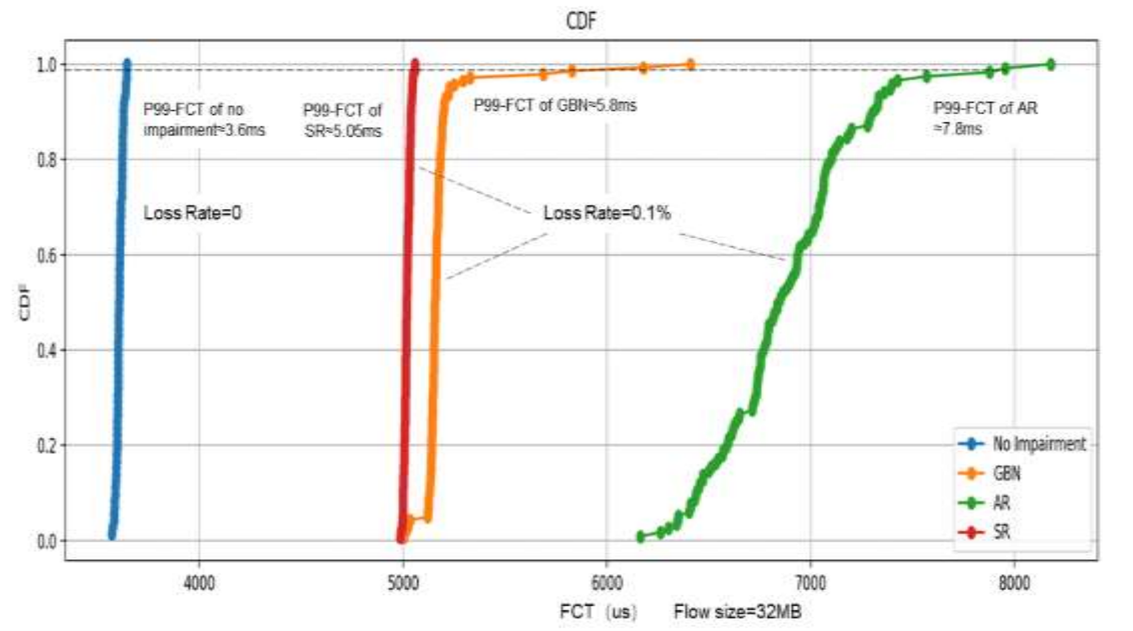
- There are two servers connected by an network impairment emulator, and each server is equipped with a Nvidia DPU (BlueField3).
- The network impairment emulator (BW=100Gbps) is used to cause packet loss in here.

Test case

- Generate RDMA flow in server A, set packet loss rate in network emulator, and record the flow completion time(FCT) under three condition:
  1. Enable RNIC Go-back-N protocol;
  2. Enable RNIC selective retransmission(SR) protocol;
  3. Enable RNIC adaptive routing(AR);

Results

- Flow size=32MB,loss rate=0.1%
- The right figure show the cumulative probability distribution of FCT under four conditions.
  - Blue line: the reference with no packet loss.
  - Orange line: enable Go-back-N
  - Red line: enable SR
  - Green line: enable AR
- The P99-FCT of AR is 34% higher than GBN, and 54% higher than SR.



- As show in the right table, lower loss rate into 0.05% and 0.02%, the P99-FCT of AR still obviously higher than non-AR conditions.

Loss rate	Go-Back-N	SR	AR
0.02%	4.88ms	4.86ms	5.44ms
0.05%	5.09ms	4.98ms	6.65ms
0.1%	5.8ms	5.05ms	7.8ms

- **Out-of-order delivery under packet spraying potentially has higher recovery time of loss packet than in-order delivery.**

Thank You !