Reviewing Load balancing issues in AI Computing Network

Jieyu Li (China Mobile)

November Plenary 2024
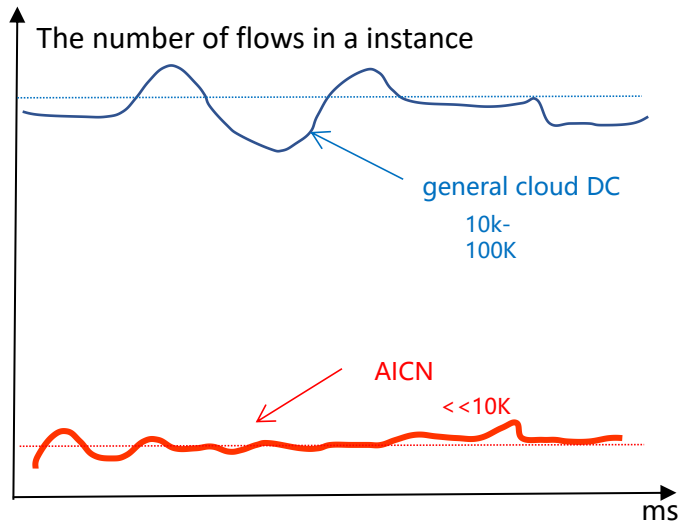
Purpose

- This contribution is related to the topic of load balacing in AICN study item.

- Reviewing the load balancing issues in AICN.
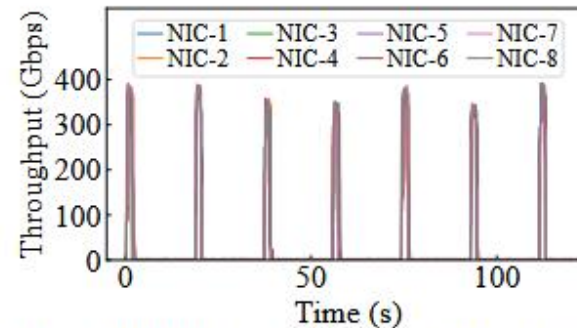
Traffic Pattern of AI Computing Network

- With the wide deploying of LLM, the traffic pattern in AI computing network is clear

Low entropy

The number of flows in a instance

general cloud DC
10k-100K

AICN

<<10K

ms

Elephant flows

Figure 2: NIC egress traffic pattern during production model training.

from Sigcomm2024: Alibaba HPN: A Data Center Network for Large Language Model Training

Collective communication

The bandwidth of single flow(Gbps)

start computing

one iteration of collective communication

Flow 1

Flow N

GPU waiting time

ms

- **General Cloud DC：** Great fluctuation，10K~200K[1]

- **AICN：** Relatively stable, a few dozen to hundreds of connections[1,2]

- **General Cloud DC：** Low bandwidth flows[3]

- **AICN：** High bandwidth flows[3], periodic burst
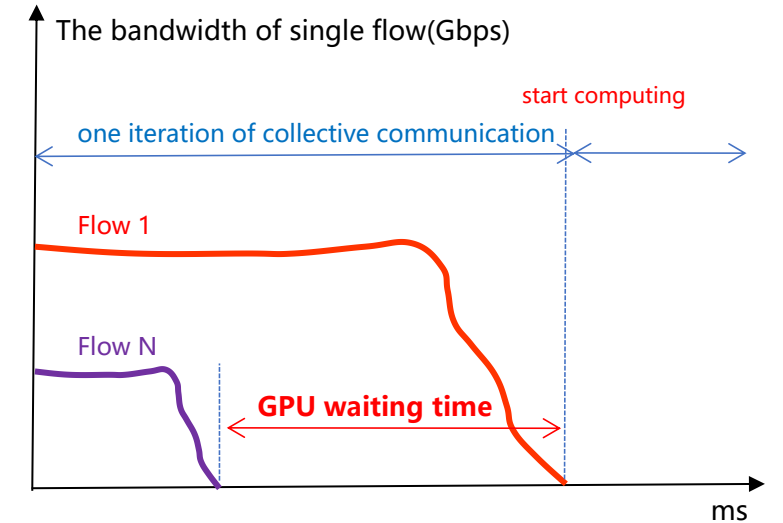
- **General Cloud DC：** Point-to-Point dominant

- **AICN：** Collective communication dominant[4]

[1] Qian K, Xi Y, Cao J, et al. Alibaba hpn: A data center network for large language model training[C]//Proceedings of the ACM SIGCOMM 2024 Conference. 2024: 691-706.
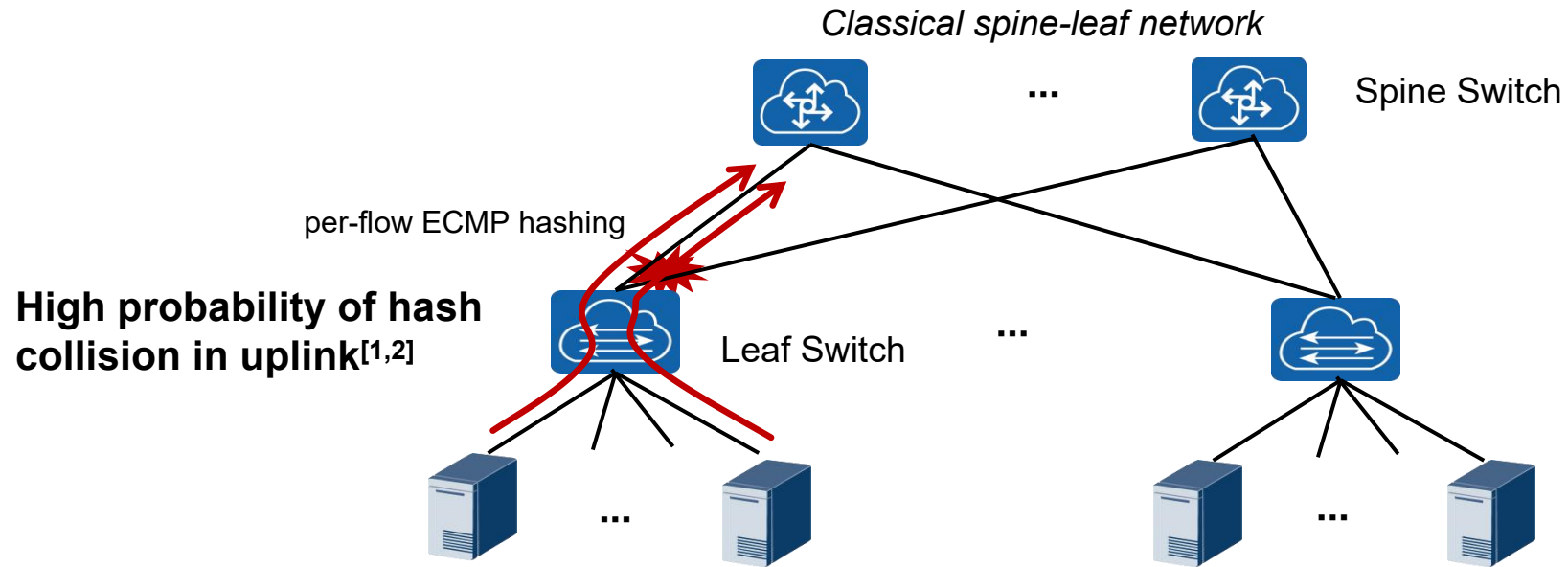[2] Gangidi A, Miao R, Zheng S, et al. Rdma over ethernet for distributed training at meta scale[C]//Proceedings of the ACM SIGCOMM 2024 Conference. 2024: 57-70.
[3] Nvidia: Next-Generation Networking for the Next Wave of AI.
[4] Cisco:Evolve your AI/ML Network with Cisco Silicon One.

Load Imbalance Problem in AICN

- **Hash collision** is the main problem when conventional per-flow Equal Cost Multi-Path (ECMP) hashing applying in AICN[1,2]

Classical spine-leaf network

- Due to the features of hash algorithm, **low entropy** will lead to the high probability of flows collision.
- Especially for **elephant flows**, the collision will cause more severer imbalance which magnifies long tail latency.
- As AI computing is mainly based on **collective communication**, long tail latency seriously affect the computing efficiency[3]

[1] Qian K, Xi Y, Cao J, et al. Alibaba hpn: A data center network for large language model training[C]//Proceedings of the ACM SIGCOMM 2024 Conference. 2024: 691-706.
[2] Gangidi A, Miao R, Zheng S, et al. Rdma over ethernet for distributed training at meta scale[C]//Proceedings of the ACM SIGCOMM 2024 Conference. 2024: 57-70.
[3] Cisco:Evolve your AI/ML Network with Cisco Silicon One.

Existing Improvement Options (1)

- There several solutions to release the load balancing problem in AICN, here sort out the existing solutions into two main categories: *per-flow-based* and *per-packet-based* solutions.

Per-Flow-based Optimization Solutions
- The granularity of distributing packets among multiple paths is per-flow.

1. Path planning: Set a definite path for each flow, avoiding uncertain hash collision.

- e.g., centralized traffic engineering.[1][2]

```
AI
scheduler  ──►  Network controller
           ──►        ▲
                      ▼
           ──►   Data plane
```
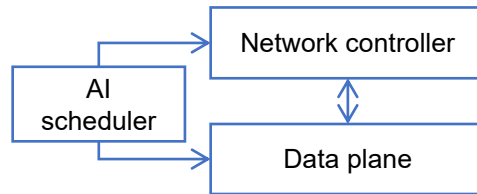
Limitations:
- "lower performance when multiple links failures happen"[1]
- "additional software complexity and manageability overhead."[1]
- hard to scale.
...

2. Splitting into subflows: Make an elephent flow into multiple subflows to enlarge entropy, lowering the possibility of hash collision.

- e.g, QP scaling[1]

```
QP  ►  QP1
       QP2
       QP3
```
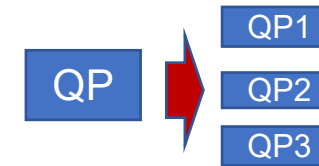
Limitations:
"the underlying probabilistic nature of hashing was a persistent downside"[1]
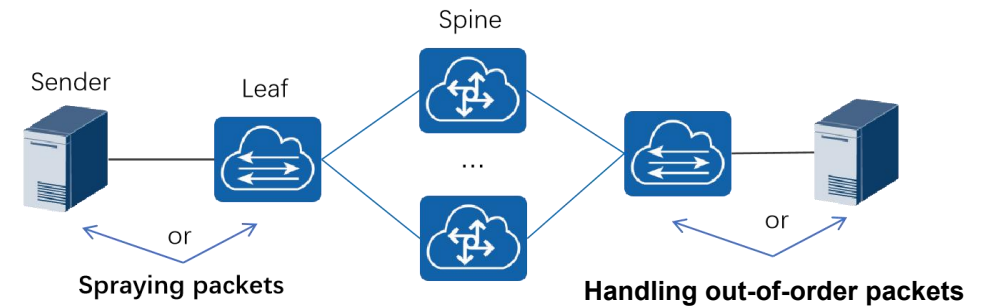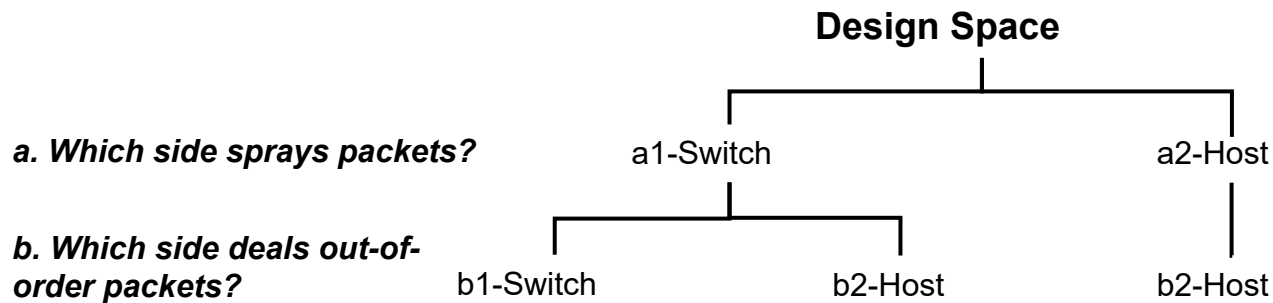...

[1] Gangidi A, Miao R, Zheng S, et al. Rdma over ethernet for distributed training at meta scale[C]//Proceedings of the ACM SIGCOMM 2024 Conference. 2024: 57-70.
[2] Huawei:https://info.support.huawei.com/info-finder/encyclopedia/en/NSLB.html

Existing Improvement Options (2)

Per-Packet-based Optimization Solutions
- Distributing packets among multiple paths in a finer granularity.

Design Space

a. Which side sprays packets?

a1-Switch a2-Host

b. Which side deals out-of-order packets?

b1-Switch b2-Host b2-Host

Spine

Sender Leaf

...

or

Spraying packets or Handling out-of-order packets

---

a1-b1: Switched-based solution
- switches spray packets and re-order out-of-order packets.
- fully decoupled with hosts, more requirments to switches.
- e.g.,Cisco's DSF[1]...

a1-b2: Coperation of host and switch
- source switch sprays packets, and receiver handles out-of-order packets.
- switches and hosts share the responsiblities.
- e.g.,Nvidia's Spectrum-X+BlueField3[2]....

a2-b2: Host-based solution
- hosts spray packets and handle out-of-order packets.
- low requirements for switches, more advanced NICs are needed.
- e.g., some new transport layer protocols supporting out-of-order service...

- Different technical route has different requirements on infrastructure, and match for different demand senarios.

[1] Cisco:https://www.ciscolive.com/c/dam/r/ciscolive/global-event/docs/2024/pdf/AIHUB-1004.pdf
[2] Nvidia: Next-Generation Networking for the Next Wave of AI.

Conclusion

- Load balancing is still one of the central issues in AICN.

- There are lots of load balancing solutions for AICN that can be mainly classified into flow-based and packet-based, and per-packet load balancing is the direction that attracts more attention.

- It's necessary to give a deep analysis on per-packet-based solutions' pros and cons, application scenarios in AICN study item report.

Thank You !