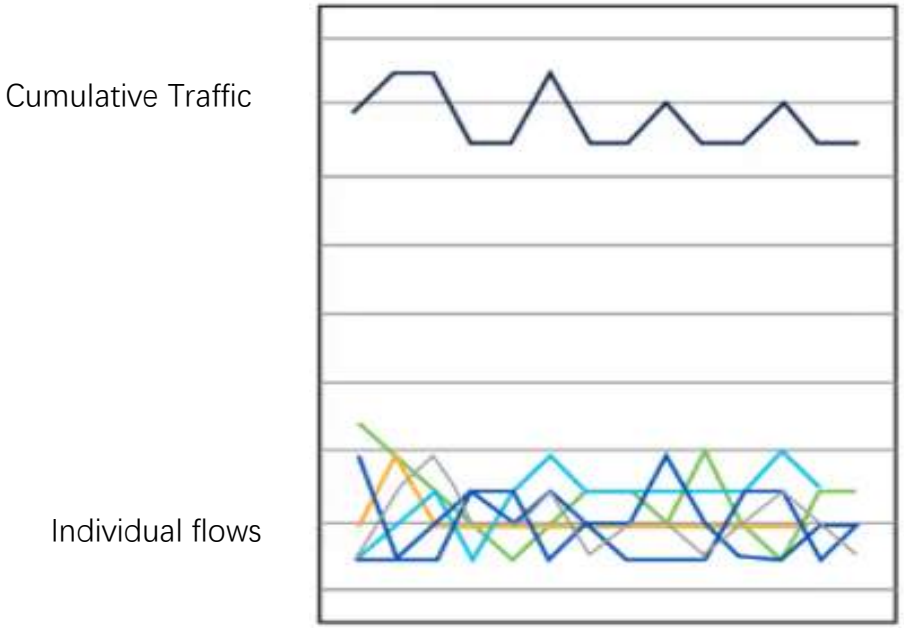
Load balancing challenges in AI fabric

Ruixue Wang（China Mobile）

Weiqiang Cheng（China Mobile）

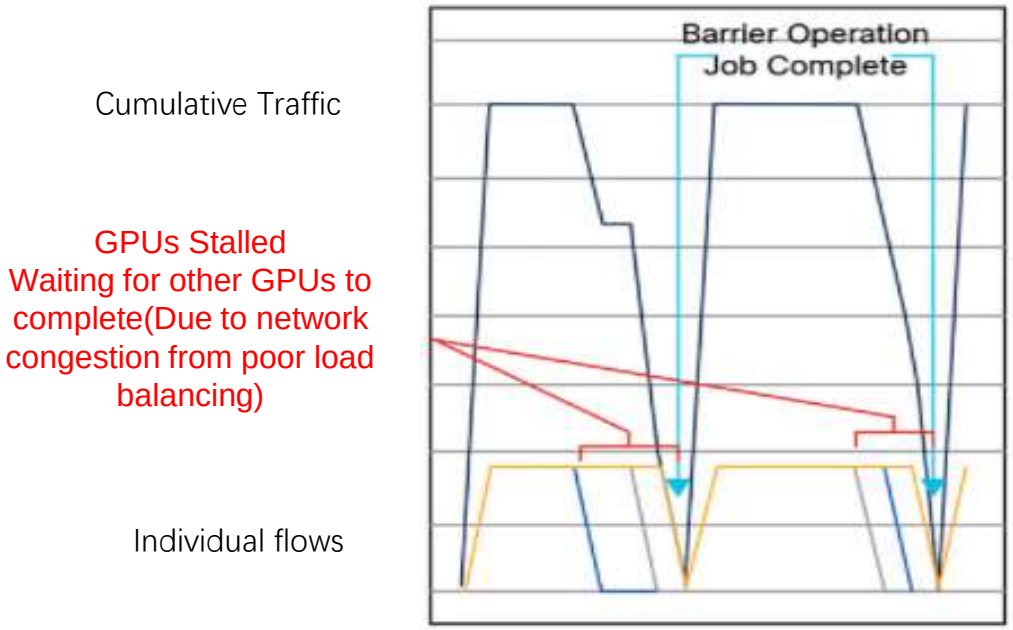
AI Traffic pattern challenge

Traditional DC Traffic pattern



- Many asynchronous small BW flows.
- Chaotic pattern averages out to consistent load.

AI (All-to-all Collective) Traffic Pattern



- Few synchronous high BW flows.
- Synchronization magnifies long tail latency and bad load balancing decisions.

Data from Cisco's public whitepaper.

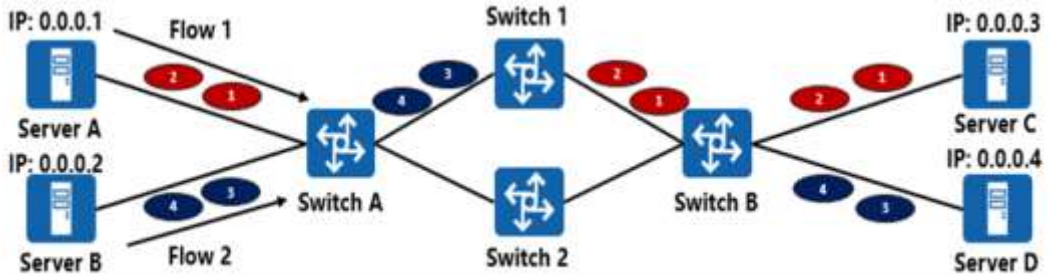
Traditional flow-based ECMP perform poorly

- **Flow-based load balancing means switches distribute packets to multiple paths in the flow granularity, and Packets within a flow take the same forwarding path.**

Limitations

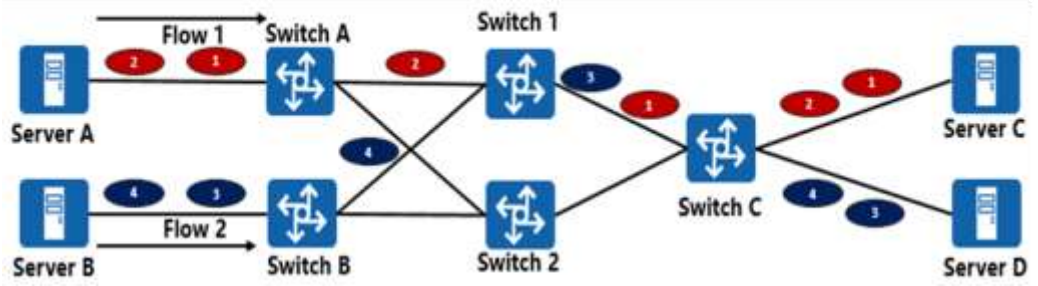
Flow-size collision:

It does not take into account the size of different flows. It is easy to forward multiple elephants flow to single path causing the congestion.



Local collision:

5 tuple based hash algorithm may output the same hash-key for different flows, resulting multiple flows to be forwarded to the same path causing local collision.

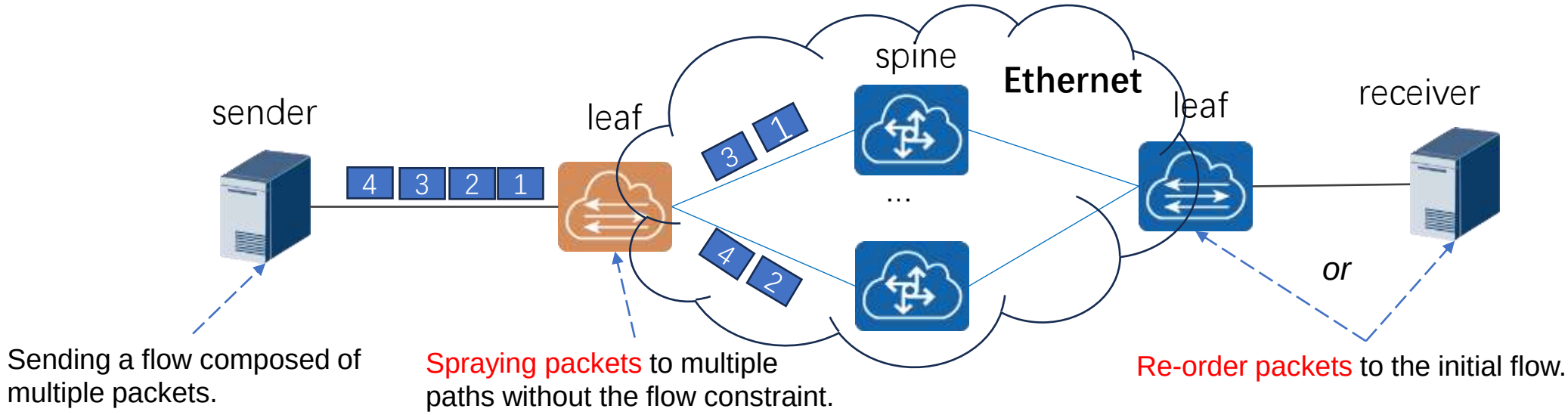


Downstream collision:

The local decision-making mechanism lacks of global view of the fabric (e.g. downstream nodes status) which may select multiple flows forwarded to the same downstream path, causing downstream collision.

Packet-based LB become the trend for AI fabric (1)

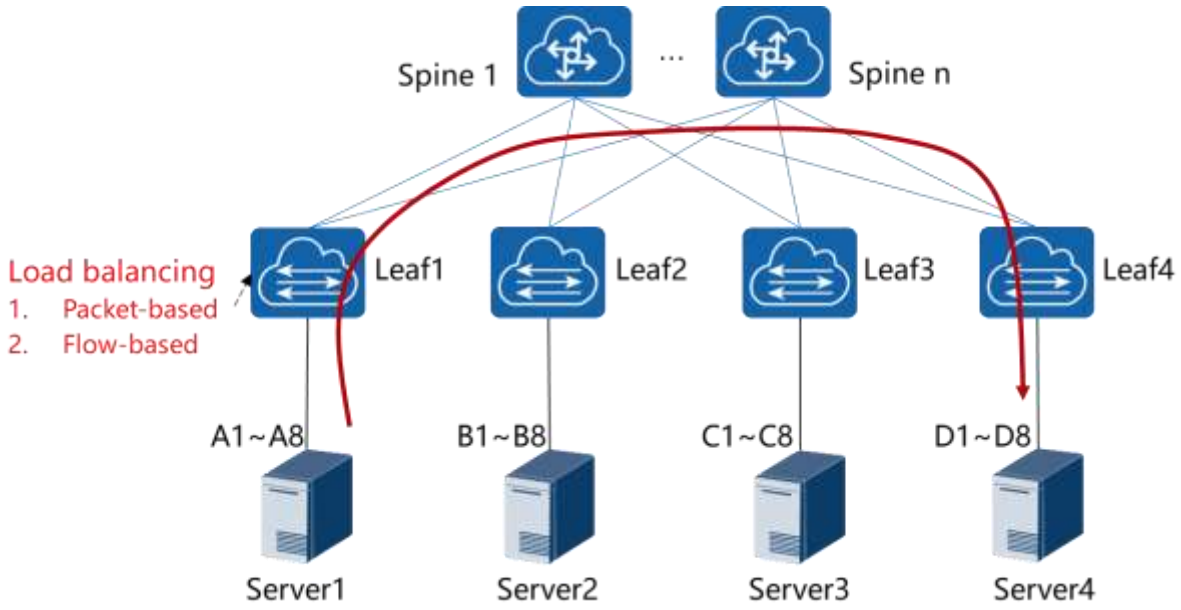
- **Packet-based load balancing means switches distribute each packet to multiple paths independently, making the load on the network more balanced than flow-based.**
- **There are several routes supporting packet-based LB:**
  - Cell-based in dedicated network or ethernet-based: Standardization → ✓ **Ethernet-based.**
  - NIC-driven or Network-driven: Applicable to different scenarios. → Focus on **network-driven** solution in this document.
- **Basic Architecture of network-driven packet-based LB in ethernet:**



Packet-based LB become the trend for AI fabric (2)

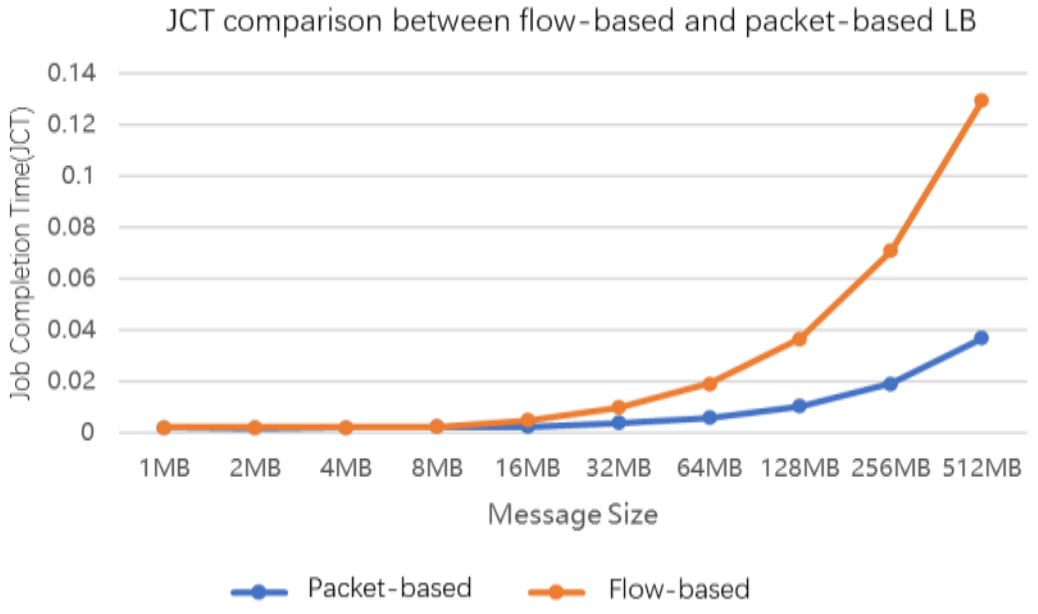
- **We conduct an experiment to evaluate the performance of flow-based and packet based LB.**

Experiment settings



- The topology is the classic two-layer clos network, 4 servers, 8GPU with 8 NICs in a server.
- There are 8 jobs running: A1~D1、A2~D2….A8~D8.

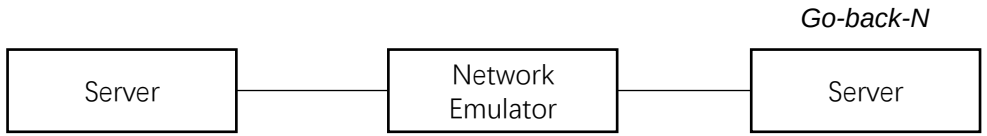
Results



- Testing the task completion time (JCT) of flow-based and packet-based load balancing under different message size.
- In a 512MB scenario, JCT of packet-based LB is reduced to about one-third compared to flow-based.

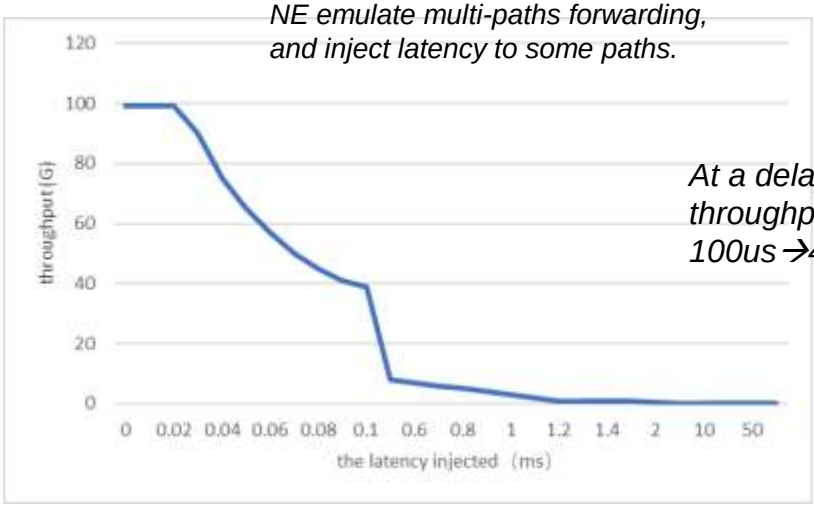
Challenges in Packet-based LB

- **The main side-effect of packet-based LB is causing packets of a flow arriving at receiver out of order:**
  - Re-order problem.
  - **Reliability problem: Loss-detection and retransmission;**



NE emulate multi-paths forwarding, and inject latency to some paths.

- **Out-of-order cause performance degradation significantly under Go-back-N mechanism.**
  - The mainstream RNIC adopt Go-back-N mechanism to provide reliability.
  - A lot of out-of-order packets may trigger frequently Go-back-N, resulting in a precipitous decline in throughput, as shown in the right emulation.



At a delay of about 20us, throughput begins to decline. 100us→40%, 500us→10%

- **RNIC can adopt Selective ACK to improve GO-back-N, but still existing problems hindering performance.**
  - The receiver can not directly determine whether the packet is lost or just out of order through the PSN,
  - relying on the timeout mechanism to detect packet loss reduces the sending rate.
  - Accurate fast-retransmit is necessary, but only by receiver is often not possible.

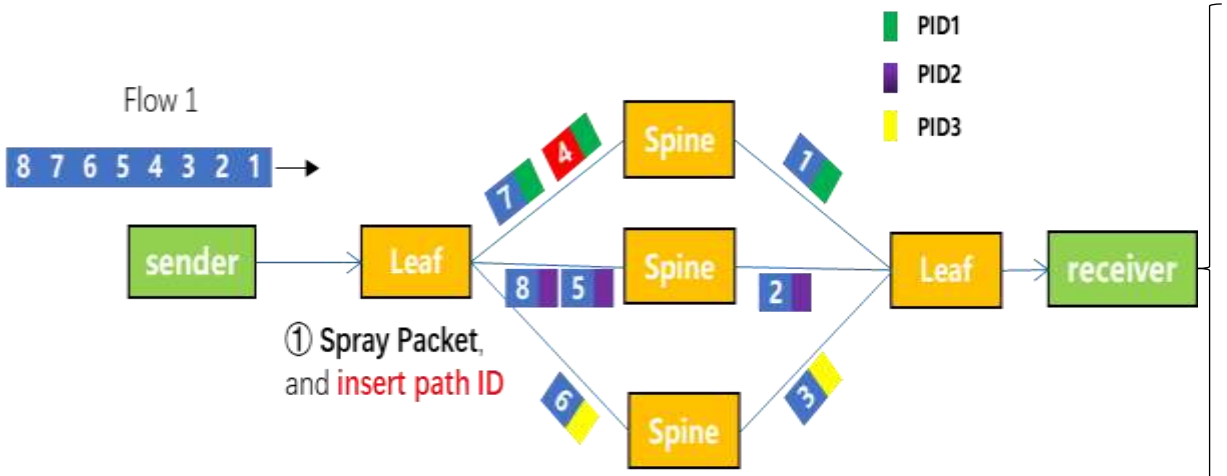
- **A preliminary conclusion is that processing out-of-order packets exclusively on the receiver NIC can hardly achieve optimal performance.**

Network can do more…

- **In packet spraying, the root difficulty of receiver dealing with out of order packets is that it does not know the forward path and state of each packet.**
- **An intuitive solution is that network provide receiver the path information of packet forwarding to help loss detection and fast retransmission.**

Key idea: network device insert the path information(e.g. Path ID) into packet header, so that the receiver can detect the loss more quickly and execute fast retransmission.

Example



② Update the receiving window of flow 1, *assume the 'hole' is packet 4*:

PSN	1	2	3	4	5	6	7	8
state	1	1	1	0	1	1	1	0

③ Update the max receiving PSN of each path of flow 1:

- Path 1: maxRcvPSN[1]:7
- Path 2: maxRcvPSN[2]:5
- Path 3: maxRcvPSN[3]:6

④ Compare the hole number with maxRcvPSN of each path:

- If hole number < maxRcvPSN of all paths → Packet 4 loss

Current industrial support for packet-based LB

① Cisco：Silicon one

Figure 1: Cognitive routing features

Global load balancing

Prior generations of Tomahawk and Trident switches support Adaptive Routing via the Dynamic Load Balancing (DLB) feature. DLB is a quality-aware load distribution scheme that selects the next hop for a packet based on the local switch's port quality. It supports both per-packet spray and flowlet modes of operation and can be enabled selectively for different traffic types with ineligible flows falling back to hash-based ECMP. DLB is successfully deployed in multiple networks today.

② Broadcom：Tomahawk 5

Table 3. Ethernet ECMP vs. scheduled fabric

Characteristic	Unscheduled Ethernet fabric	Fully scheduled fabric
Distribution method	ECMP hash	Spray and re-order
Link utilization	Low	High

③ Nvidia spectrum x

Spectrum-X Technology Innovations

Spectrum-4 switches and BlueField-3 SuperNICs work in tight coordination to form a **NCCL-optimized network fabric** built to optimize AI cluster performance using a suite of end-to-end innovations:

- **RoCE adaptive routing** avoids congestion by dynamically routing large AI flows away from congestion points. This approach improves network resource utilization, leaf/spine efficiency, and performance. The Spectrum-4 switch employs fine-grained load balancing, re-routing active flows to eliminate congestion. Additionally, the BlueField-3 SuperNICs work in tandem to handle out-of-order packets, placing packets in the correct order in the destination memory. RoCE adaptive routing supports profiles for efficient provisioning and automation.

- **The mainstream chip venders have supported the packet-based load balancing, but their solutions are different. → standardization of packet-based load balancing on ethernet is needed.**

Summary

- **Introduce the drawbacks of traditional flow-based ECMP for AI fabric, and packet-based load balancing become the trend.**
- **Analyze the challenges bring to receiver in packet-based load balancing.**
- **Network can assist receiver to solve the challenges.**
- **Potential Standard Requirements：** Need to standardize packet information in L2 for network-assisted fast retransmission, such as path ID.

Thank You !