Discussion on Switch-controlled Packet-level Load Balancing Solutions

Jieyu Li（China Mobile）

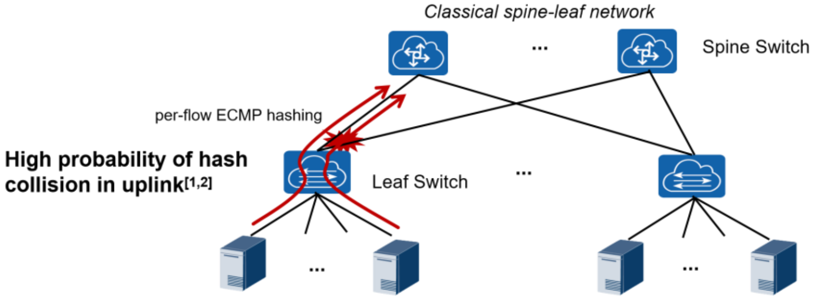
Weiqiang Cheng（China Mobile）

Ruixue Wang (China Mobile)

July Plenary 2025

Recap

- **Regarding Load balancing issues in AI computing network, several contributions have discussed the requirements and challenges in NENDICA AICN study item, here recap some key points:**
  - Unique AI traffic characteristics, large bandwidth, low entropy, cause load imbalance problem in scale-out network, which would greatly impact tail latency and AI training/inference performance.
  - Several efforts has been put forward to improve network balance in AI network. And the fine-grained packet spraying is almost the industrial consensus to completely solve imbalanced problem intra-network.

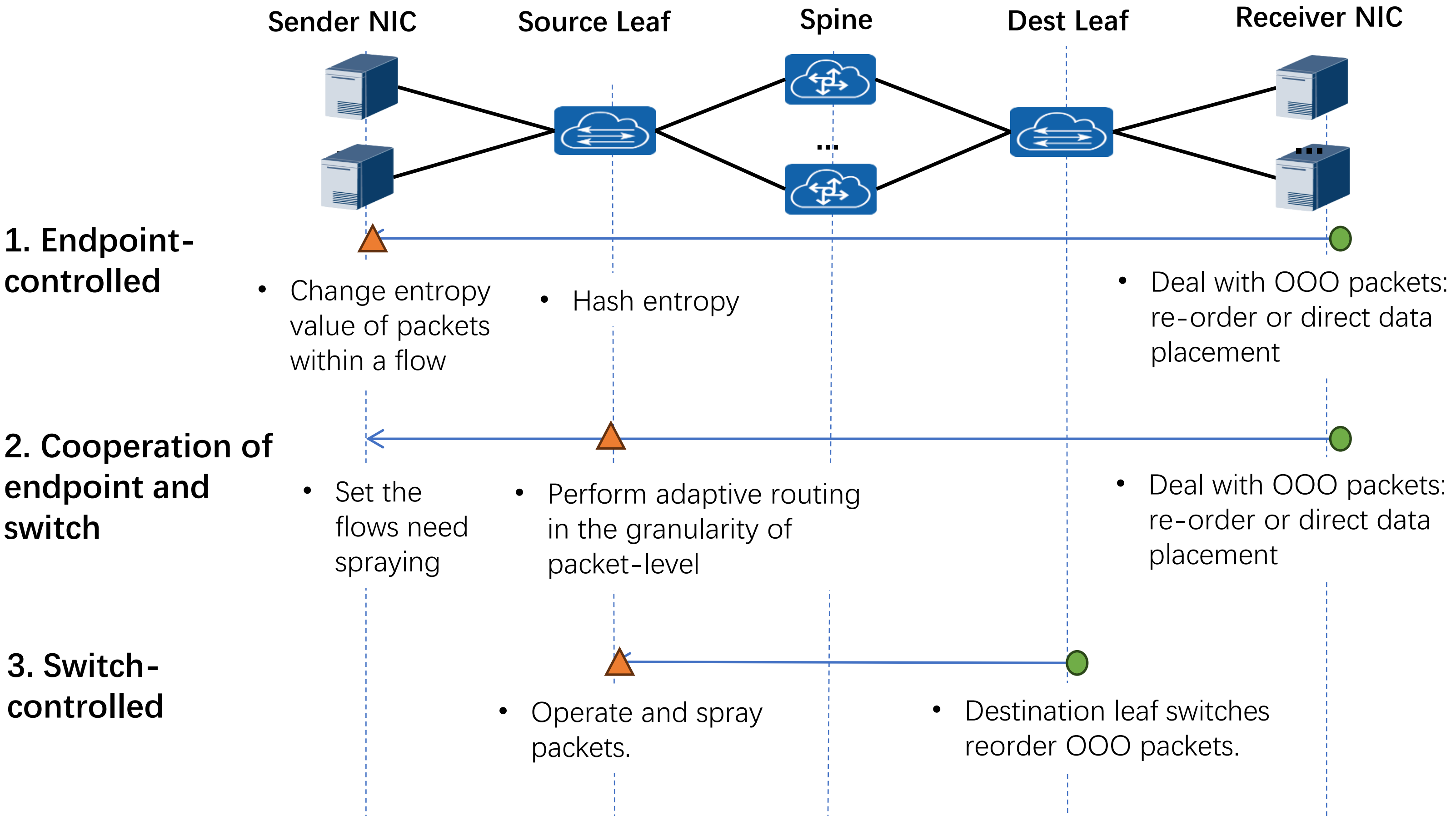

Classical spine-leaf network
Spine Switch
per-flow ECMP hashing
High probability of hash collision in uplink[1,2]
Leaf Switch
...

Ref: Contributions 802.1-24-0007, 1-24-0025, 1-24-0028, 1-24-0060

- **This contribution intend to further discuss switch-controlled packet-level LB solution and its standardization opportunities in 802.1.**

Different deployment patterns of packet-level LB

- Many mainstream venders and consortiums have put forward their packet-level LB solutions, those can be categorized into three types of deployment pattern based on different work division between network switch and endpoint.



Sender NIC

Source Leaf

Spine

Dest Leaf

Receiver NIC

1. Endpoint-controlled

- Change entropy value of packets within a flow
- Hash entropy
- Deal with OOO packets: re-order or direct data placement

- **Dependencies:** advanced NIC
- **Standardization activities**： UEC do this in its new transport layer[1].

2. Cooperation of endpoint and switch

- Set the flows need spraying
- Perform adaptive routing in the granularity of packet-level
- Deal with OOO packets: re-order or direct data placement

- **Dependencies:** advanced NIC, AR-supported switch.
- **Standardization activities**： several individual drafts about adaptive routing in IETF[2].

3. Switch-controlled

- Operate and spray packets.
- Destination leaf switches reorder OOO packets.

- **Dependencies:** advanced switch.
- **Standardization activities:** some efforts to standardize the control plane of scheduled ethernet fabric in OCP[3].

- **The switch-controlled pattern is fully decoupled with endpoint, and “provides flexibility & speed across multiple generations and types of accelerators and NICs”[4]**

[1]ultraethernet.org/wp-content/uploads/sites/20/2025/06/UE-Specification-6.11.25.pdf

[2]https://datatracker.ietf.org/doc/draft-dong-fantel-state-of-art/

[3]OCP 2024: Insights from Production: Scheduled Ethernet Fabric in Large AI Training Clusters

[4] Powering the AI Future Meta Vision for Open Systems for AI - presented by Meta

The existing switch-controlled packet-level LB solutions (1)

The existing switch-controlled LB solutions can be further divided into (1) cell-based and (2) packet-based according to the difference of basic forwarding unit.

(1). Cell-based

- **Basic forwarding unit:** fixed-length cell.
- **Source leaf switches** segment packets into cells**,** and spray them into all available ports.[1]
- **Dest. leaf switches** re-order and re-assemble cells, then regain original packets.[1]

- **Pros:** cell spraying can achieve optimal balanced load distributing multiple egress ports, regardless of the variable length of packets.
- **Cons:** complexities to assemble cells; not the standard ethernet packet structure intra-fabric thus needing two types of chips for leaf and spine switches respectively.

Ref: [1] Scheduled Ethernet Fabric for Large scale AI training cluster

The existing switch-controlled packet-level LB solutions (2)

(2). Packet-based

- **Basic forwarding unit:** ethernet packet.
- **Scheduling granularities**:
  - Packet[1];
  - Packet container, a logical group of packets to approximate a fixed-length unit.[2]
- **Source leaf switches** insert ordering information to each packet leveraging ethernet header extension, and spray them to egress port.
- **Dest. leaf switches** reorder packets based on information carried in the header.

- **Pros:** compatible with ethernet forwarding, relatively low extra overhead on restoring packets compared with cell-based one.
- **Cons:** not extremely balancing due to slight packet-length differences.

In the perspective of constructing a fully unified and low overhead ethernet-based solution, the packet-based one is the better choice and more appropriate to consider standardization.

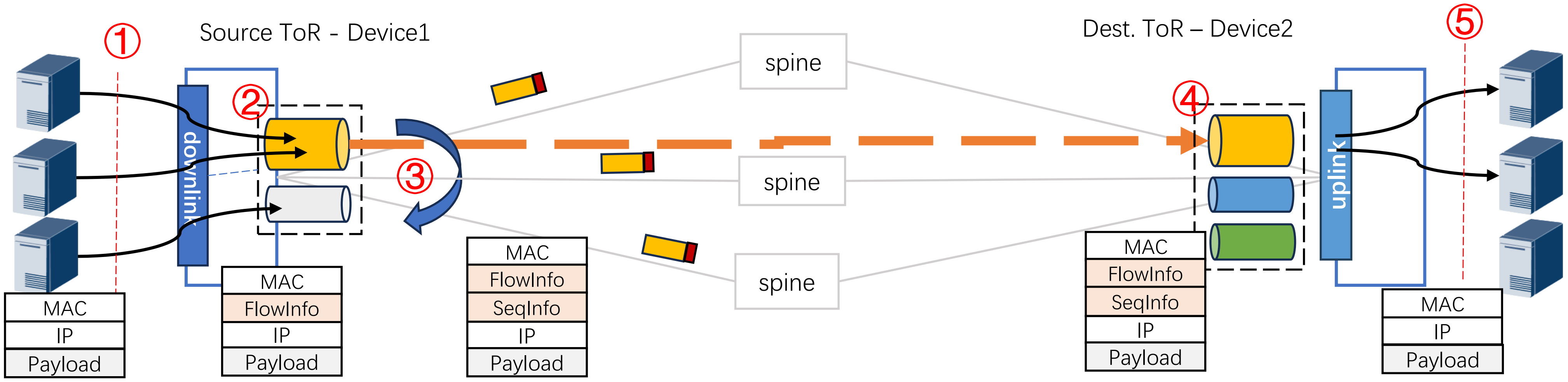
Ref:

[1] Evolved Networking: the AI/ML Challenge

[2] https://regmedia.co.uk/2024/11/26/china_mobile_gse_whitepaper.pdf

A simple example of possible end-to-end processing in fabric

If we do this based on ethernet packet:



① Send original flows.

② Aggregate into a new spray-then-reorder "flow"

- Aggregated Source: from the certain source ToR
- Aggregated Destination: toward the certain dest. ToR or dest. ToR's output port or priority.

→ New Flow Information (e.g., src. deviceID) should be carried.

③ Spray the new "flow"

- Granularities:
  - One packet
  - A Group of packets
- Strategies：
  - Round-robin
  - Congestion-aware
  - …

→ Sequence information within the new 'flow' should be carried.

④ Re-order the "flow"

- Identify the flow based on FlowInfo
- Reorder based on SeqInfo
- Remove the extra tag.

⑤ Receive original flows.

Position consideration: Layer 2 could be the optimal option

1) Faster switch processing.

2) Oblivious to upper protocol: can be used to the cases without IP layer, pursuing the low latency.

802.1 Standardization Considerations

Benefits

- The in-network packet-level spraying can eliminate the congestion intra-network with any traffic characteristics.
- Regarding the incast congestion in the last hop, there are some valuable mechanisms can solve, like SFC (P802.1qdw) or some VoQ-based credit mechanism.
- Packet spraying intra-network + incast congestion control at edge-side≈ no congestion in network.
- “Provides flexibility & speed across multiple generations and types of accelerates and NICs”.

Considerations on IEEE 802.1 standardization opportunities

- The existing switch-controlled packet-level LB solutions are almost proprietary, and lack of international standard.
- It’s appropriate for IEEE 802.1 to consider standardizing the switch-controlled per-packet LB since the emphasis on the network switch.
  - Define the extra information needed and its encapsulation way.

  …

Thank You !