IETF 118 – RTG WG

Coordinated Congestion Management

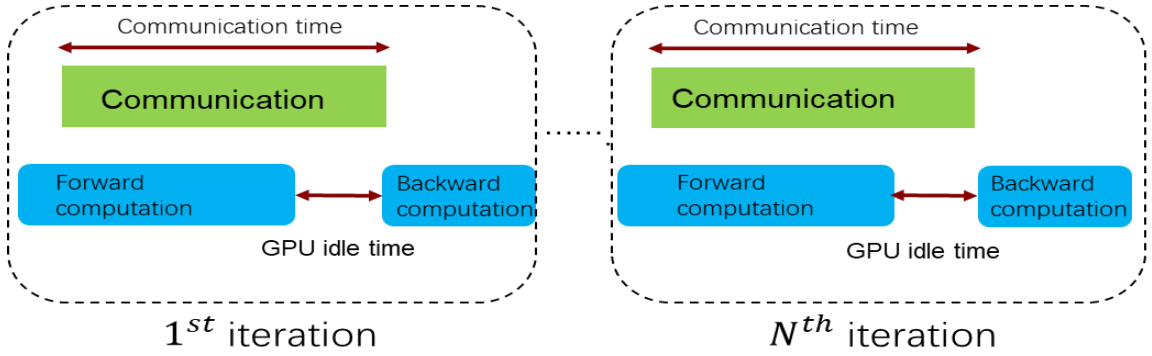
draft-lyu-rtgwg-coordinated-cm

Yunping(Lily) Lyu，Yuhan Zhang, Mengzhu Liu

Communication in AI Training

Large-scale AI models (e.g. LLM) have a huge number of parameters.

Distributed parallel training on thousands of GPUs(even more) is imperative.



	Characteristic	BW Requirement
TP	• Intra node communication (AllReduce) • 100s GB level traffic	★ ★ ★ ★ ★
PP	• Inter nodes communication (send/recv) • 100s MB ~ GB level traffic	★ ★ ★
DP	• Inter nodes communication (AllReduce) • GB level traffic	★ ★ ★ ★
MOE	• Inter nodes communication (AlltoAll/AllReduce) • GB level traffic	★ ★ ★ ★

AI fabric： GPU-GPU Interconnects for AI training

- **Bandwidth hungry**
- **Burst of low number, large flows**

AI fabric Needs Congestion Management

Shortening communication time is key to improve computation efficiency.

High EFFECTIVE network bandwidth is required.

- **Bandwidth hungry**
- **Burst of low number, large flows**

Congestion management is a MUST.

Load Balancing

ECMP is not suitable to handle AI traffic. Adaptive routing is a new choice.

- Adaptive routing: dynamically shift traffic path of a flow based on network status

Congestion Control

The congestion control mechanism identifies network congestion, then adjust the sending rate at the source host to alleviate congestion.

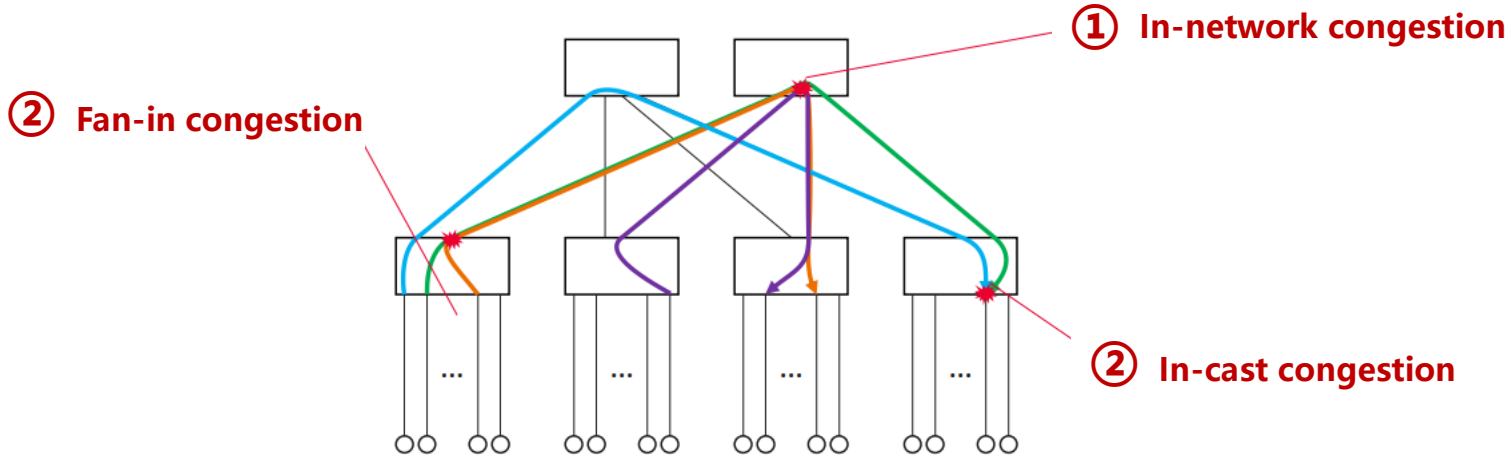
- ECN-based / RTT-based

Open Questions:

- how to decide what to do - shift flow or adjust congestion control?
- how adapt quickly and minimize reordering ?
- shifting flow too many times in a short period probably not going to help - add some sort of dampening?
- moving old vs new flows?
- Cross-group work

https://datatracker.ietf.org/meeting/117/materials/slides-117-rtgwg-sessa-14-dragonfly-routing

Issue of Uncoordinated CM



Type ① congestion:

If both AR and ECN-based CC are activated, the result is the flow is forwarded on the light-loaded path at a low rate. That reduces effective bandwidth.

Type ② congestion:

Performing CC is the only way. Any additional AR on the flow not only fails to mitigate congestion, but also may introduce unnecessary out-of-order traffic.

Proposal for Coordinated CM

Basic idea:

- Distinct *CC traffic* and *non-CC traffic*, and treat them differently when congestion occurs
  - CC traffic : packets in the flows which cause type 2 congestion
  - Non-CC traffic: packets in the flows which cause type 1 congestion
- For CC traffic,
  - do not perform AR
  - just perform CC
- For non-CC traffic,
  - perform AR first
  - when AR cannot find suitable path for switching, perform CC

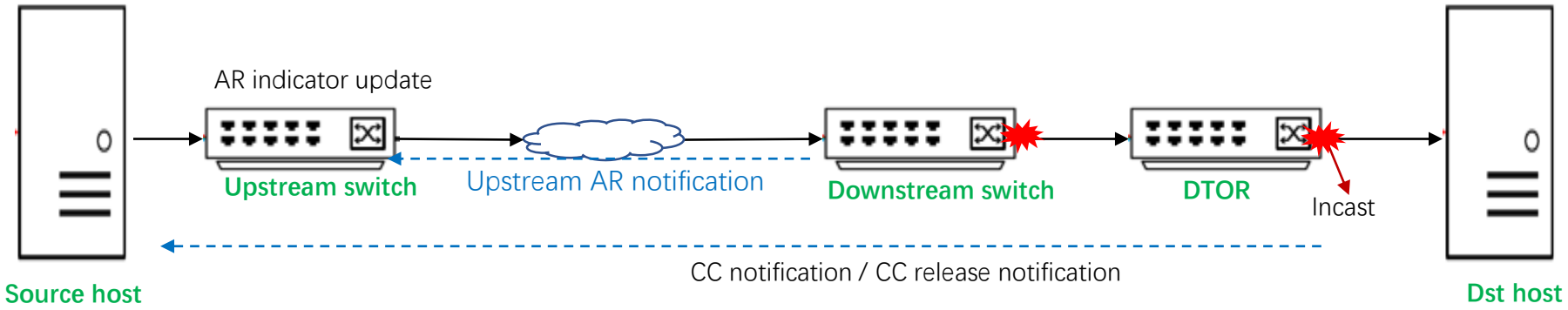
C-Tag

Coordination tag (C-tag) is in data packet.

It contains CC indicator and AR indicator.

Field	Description
CC indicator	indicate if the packet belongs to a flow which needs congestion control 0: non-CC traffic 1: CC traffic
AR indicator	indicate the location of upstream point where adaptive routing can be performed; can be an switch ID, or switch IP address

Notification Message

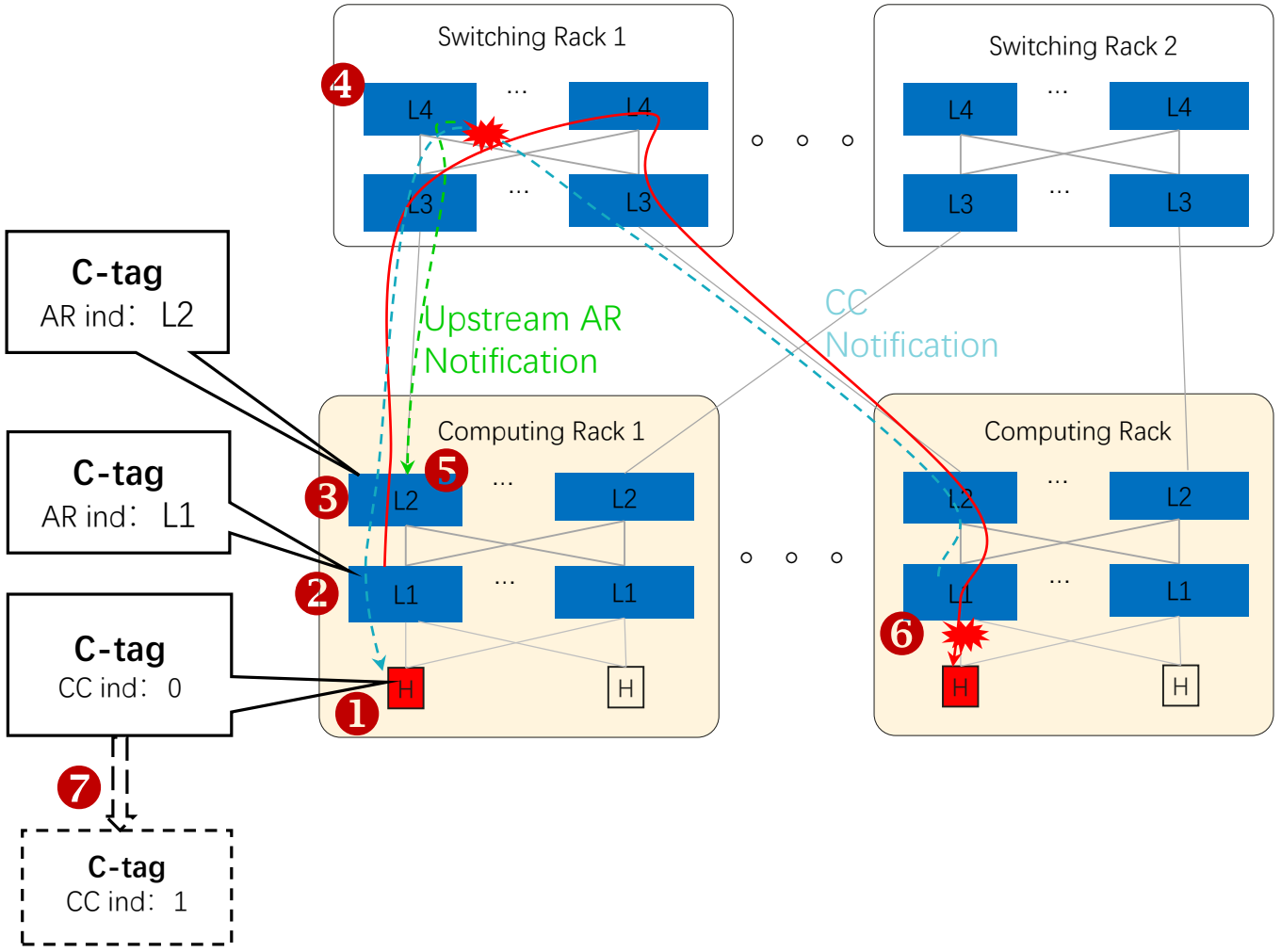


- CC notification / CC release notification
  - Example: the notification is sent from incast congetion switch to incast flow source host, notifying the source host to set/unset CC indicator of the packets in the incast flow.
- Upstream AR notification
  - Example: If the switch determines to perform AR upstream, upstream AR notification is sent to the upstream switch.

Field	Description
Source IP	IP address of the switch which sends the notification
Dst IP	IP address of the destination which will handle the notification
Type	Notification type
Flow key	Flow information, such as 5-tuple

Example of CCM

1. Default value of CC ind is 0
2. AR ind is updated to L1
3. AR ind is updated to L2
4. Upstream AR notification is sent to L2
5. L2 selects another path for the subsequent packets of the flow
6. Incast congestion triggers CC notification to host
7. CC ind is set to 1 for the subsequent packets of the flow.



Next Steps

- Encourage discussion on the mailing lists
- Update the draft based on feedbacks
- Welcome to contributions and co-authors

THANKS!