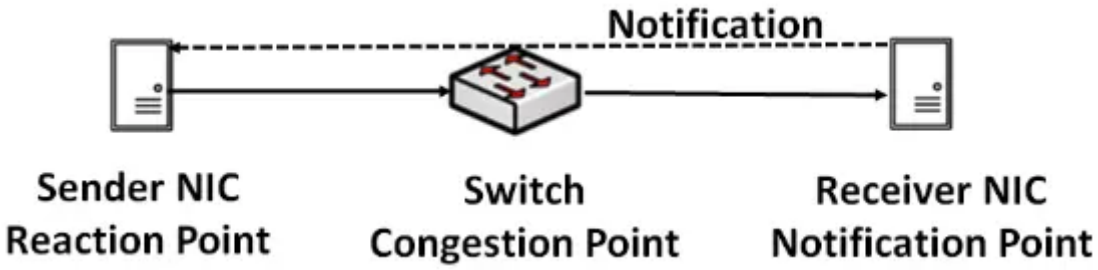
Fast Congestion Notification Packet (CNP) in RoCEv2 Networks

draft-xiao-rtgwg-rocev2-fast-cnp-05

Xiao Min	ZTE
Hesong Li	ZTE
Kan Zhang	China Mobile
Weiqiang Cheng	China Mobile
Jin Yang	China Mobile
Xiaoming He	China Telecom

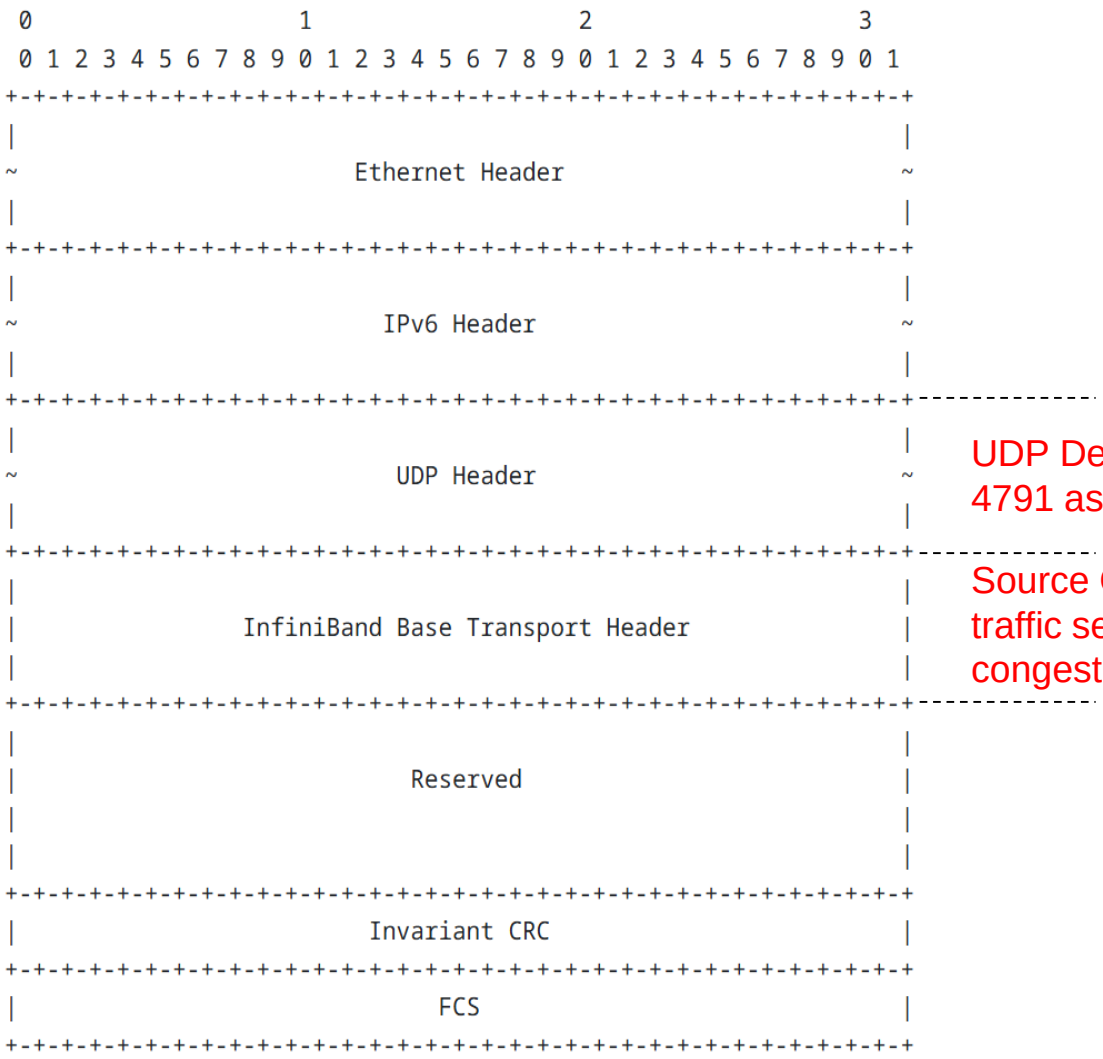
What's CNP in RoCEv2?



- Step 1, Congestion detected by Switch.
- Step 2, ECN bits in IP header marked by Switch.
- Step 3, Marked ECN bits detected by Receiver.
- Step 4, CNP sent to Sender by Receiver.
- Step 5, Transmission rate reduced by Sender.

Note: CNP is widely deployed in AIDC

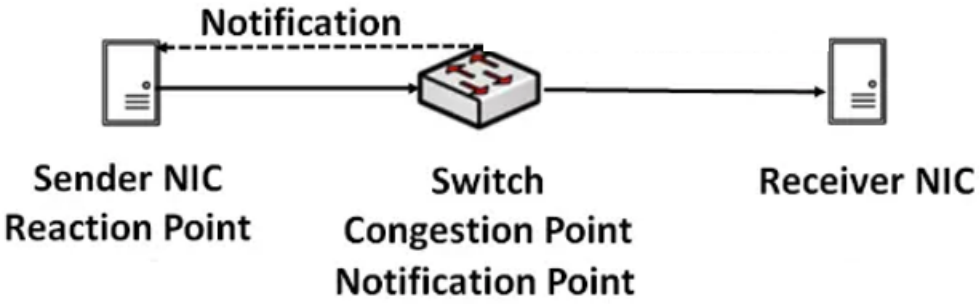
CNP Format defined by IBTA



UDP Destination Port number 4791 assigned by IANA

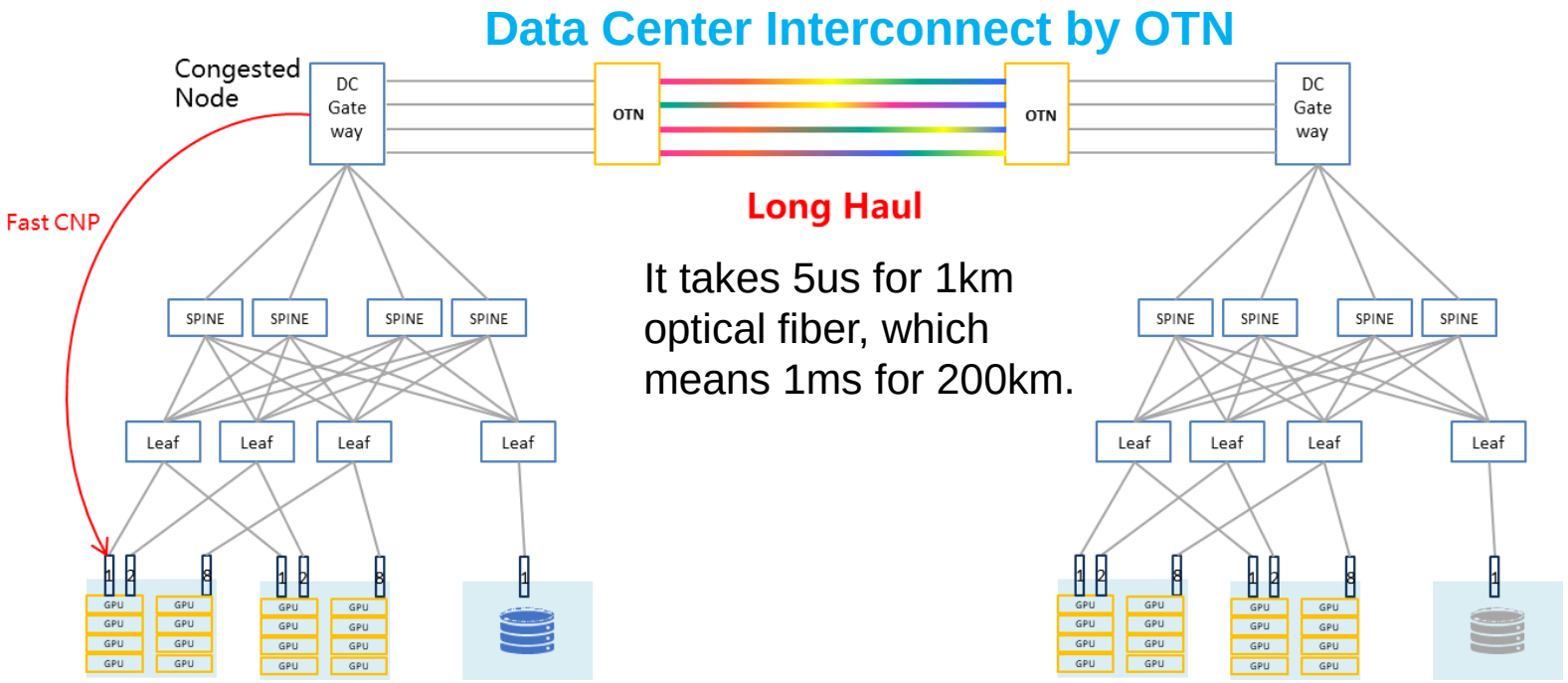
Source QP field used by the traffic sender to identify the congested traffic flow

What's Fast CNP and why?



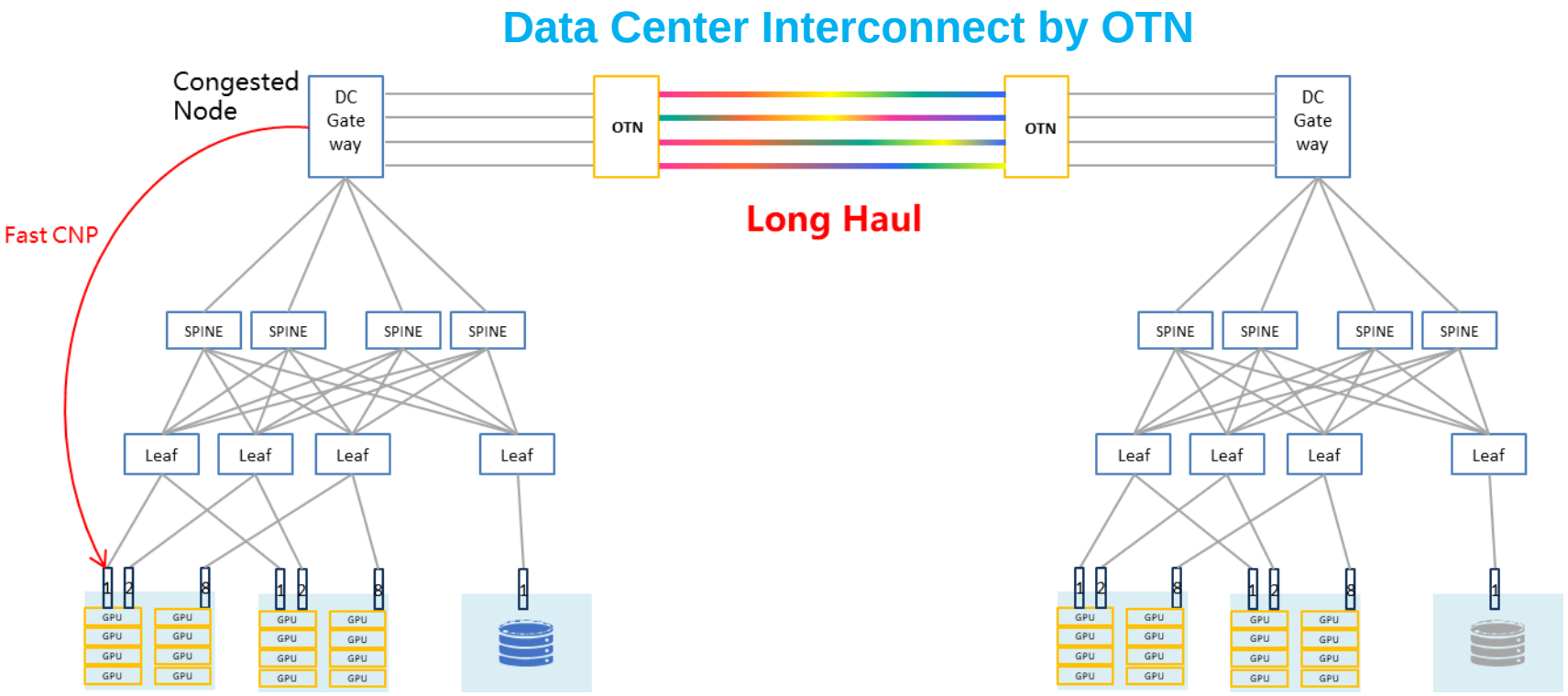
- Step 1, Congestion detected by Switch.
- Step 2, ECN bits marked by Switch. (Optional)
- Step 3, Fast CNP sent to Sender by Switch.
- Step 4, Transmission rate reduced by Sender.
- Fast CNP (as indicated by it's name) is faster than CNP
  - Notification Point is closer to Sender.
  - There is significant benefit when the distance between Switch and Receiver is non-negligible.

The Existing Implementation of Fast CNP



- Fast CNP has the same format with CNP defined in IBTA
- DC GW needs to populate Source QP into Fast CNP, but it's unable to know Source QP directly from the congested data packets, So
  - DC GW learned the mapping table between the Dest QP and the Source QP from the RoCEv2 data packets during the RDMA session establishment. The prerequisite for the learning is that data traffic in both forward and reverse directions must go through the exact same DC GW.
  - DC GW looks up the mapping table to obtain the Source QP before sending Fast CNP.

The Remaining Problems of Fast CNP



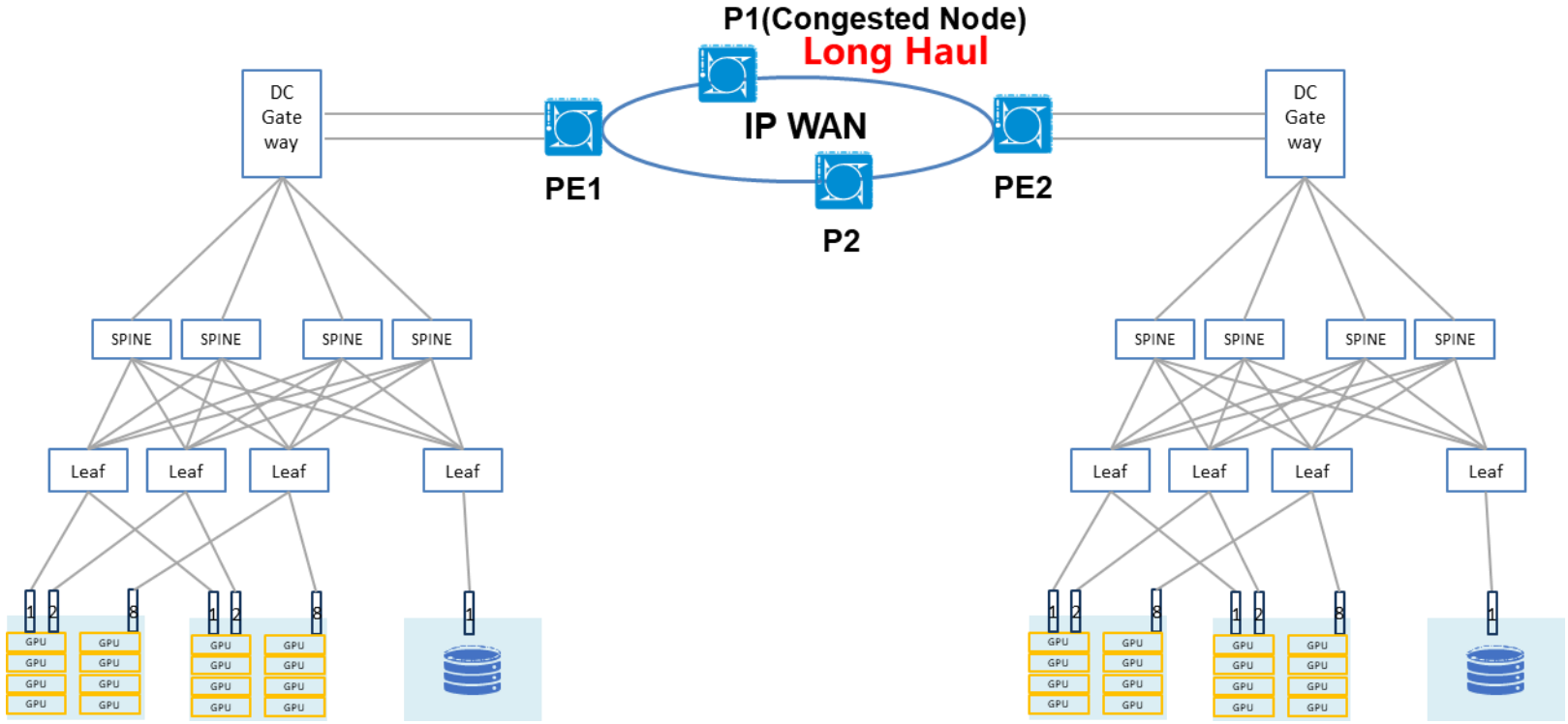
- When the SPINE switch is congested and needs to send Fast CNP, how can it learn the mapping table between the Dest QP and the Source QP during the RDMA session establishment?
- When there are two DC GWs for load balance, how can one DC GW learn the mapping table between the Dest QP and the Source QP?

Fast Congestion Notification Packet (CNP) with Proxy

draft-xiao-rtgwg-proxy-congestion-notification-03

Xiao Min	ZTE
Kan Zhang	China Mobile
Zehua Hu	China Telecom

Problems for Fast CNP when DCI by IP WAN Instead of OTN



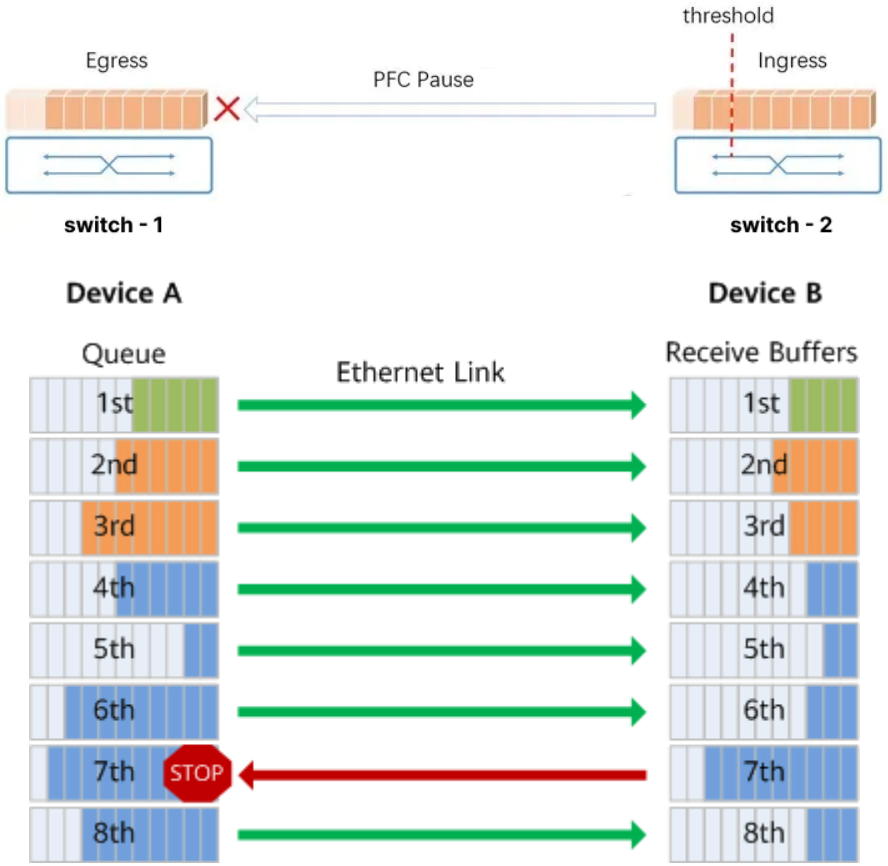
- When P1 is congested, how can P1 send Fast CNP to the traffic sender which is a server outside the IP WAN and within the Date Center?
- When P1 is congested, how can P1 learn the mapping table between the Dest QP and the Source QP during the RDMA session establishment?

Congestion Notification for Pause

draft-xiao-rtgwg-congestion-notification-for-pause-01

Xiao Min ZTE
Kan Zhang China Mobile

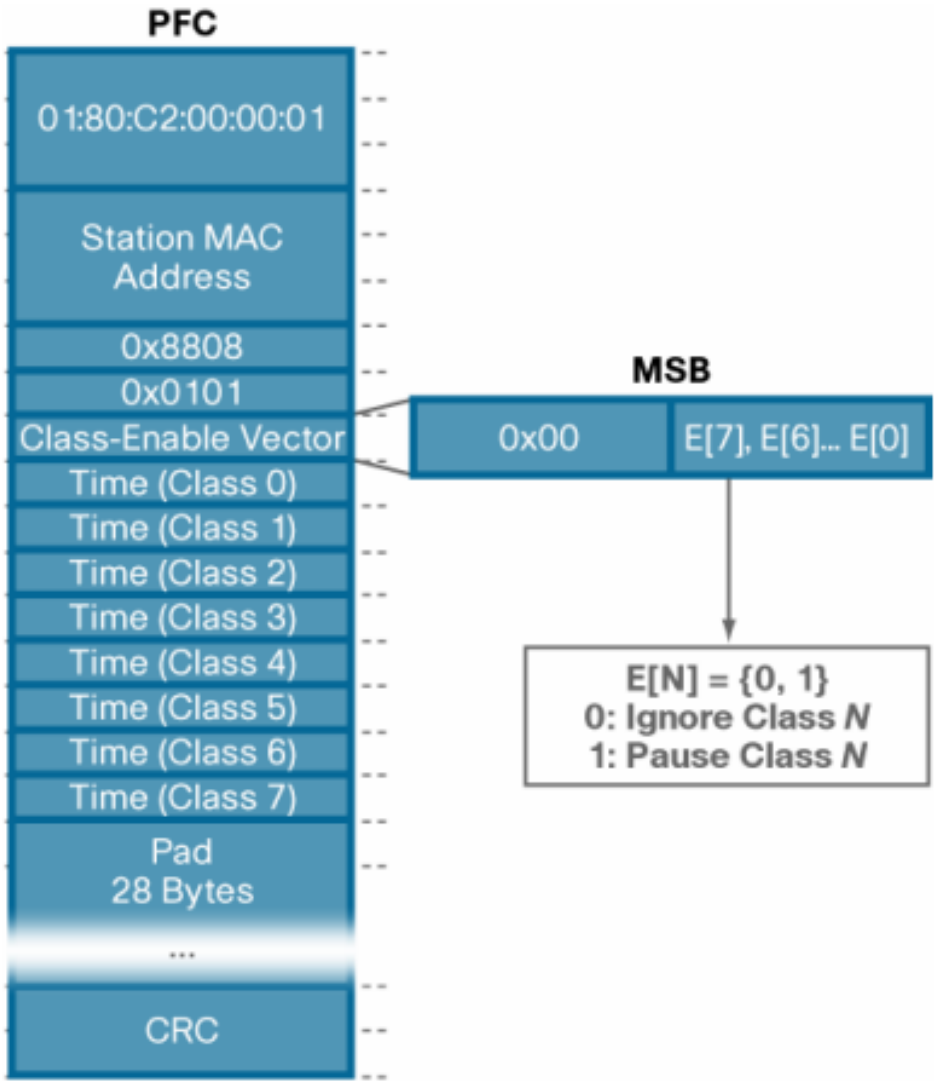
What's PFC?



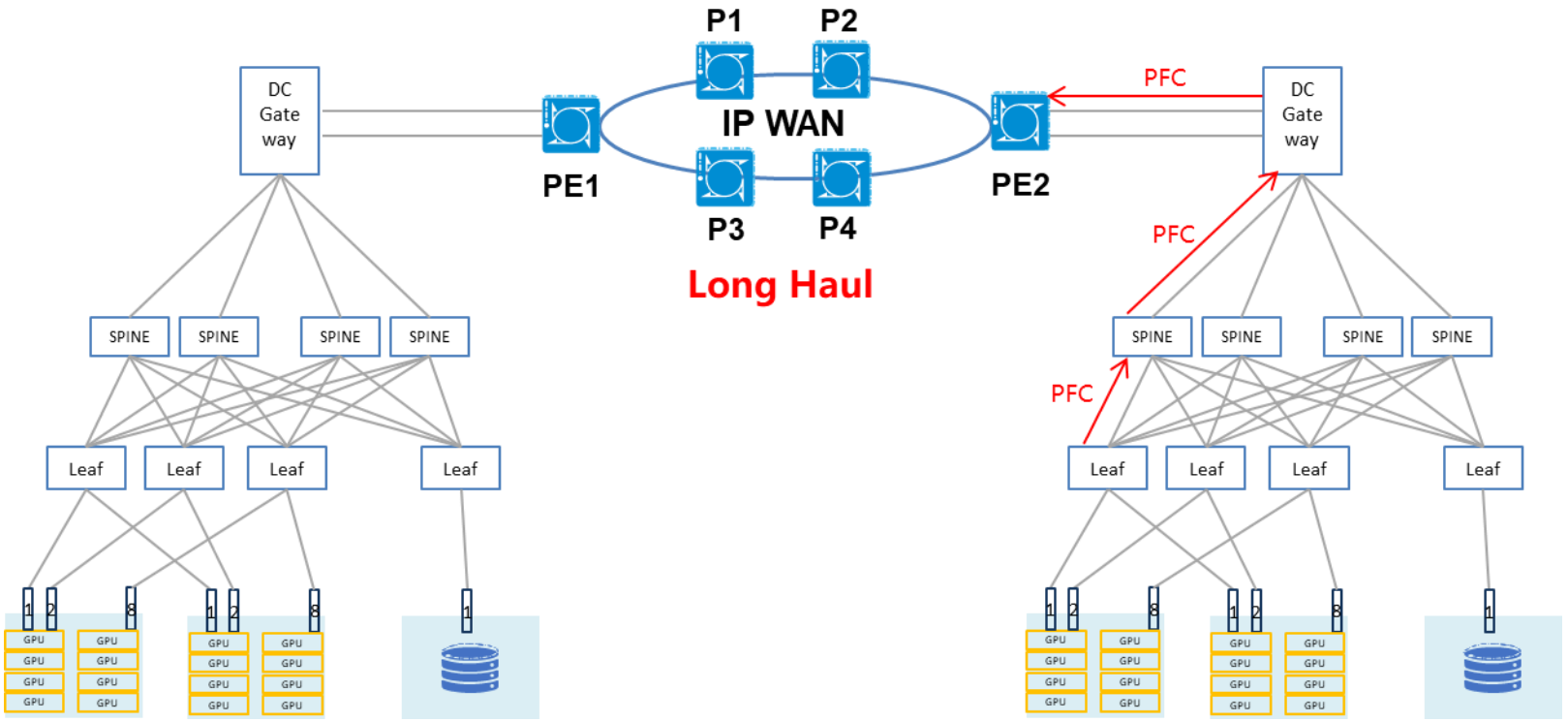
- Priority-based Flow Control (PFC) provides a link-level flow control mechanism that can be controlled independently for each Class of Service (CoS)

Note: PFC is widely deployed in AIDC

PFC Frame Format defined by IEEE



Problems for PFC when DCI by IP WAN



- When PE2 receives PFC frames from the DC GW, the first thing it can do is to pause traffic forwarding with the corresponding CoS indicated by PFC
- When PE2 gets also congested due to traffic forwarding pause, what else it can do to mitigate the traffic congestion?

Next step

- Discuss the solutions if the problems are agreed by FANN WG