Layer 3 Quantized Congestion Notification (L3QCN)

draft-yu-tsvwg-l3qcn-00

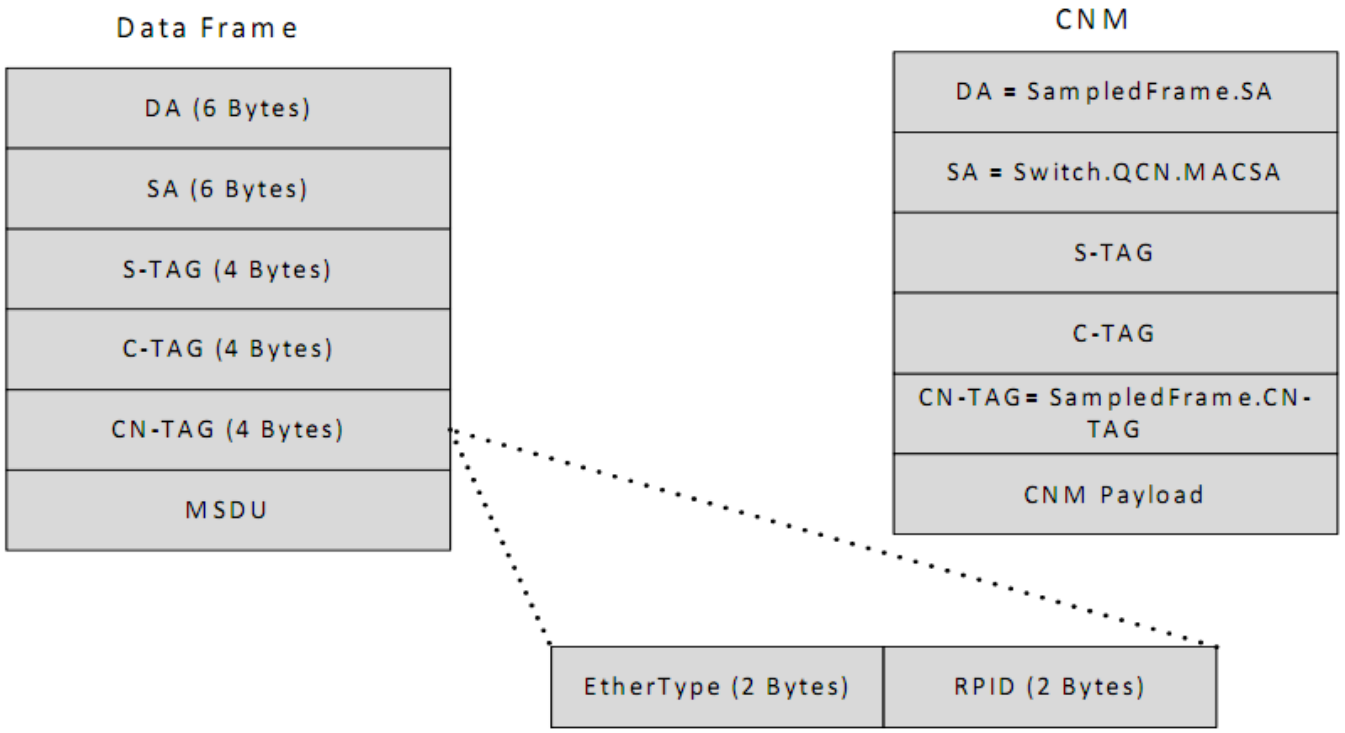
Yolanda Yu

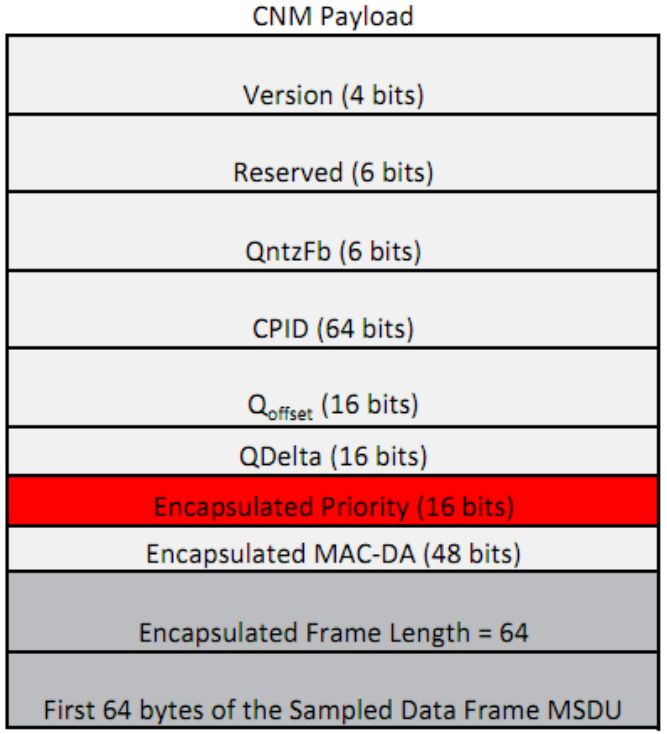
Yolanda.yu@huawei.com

IETF97-Seoul

Background

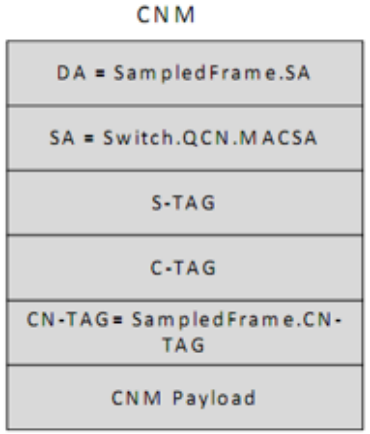
- QCN – Quantized Congestion Notification, defined in IEEE 802.1Qau
- Reduce the sending rate of the stream in congestion point
- Resolve the problem of re-transmission and flow control which may involve the latency
- IEEE 802.1Qau defines 2 new Ethernet packets
- Congestion algorithm – pls refer to IEEE P802.1Qau/D2.4 30.2.1

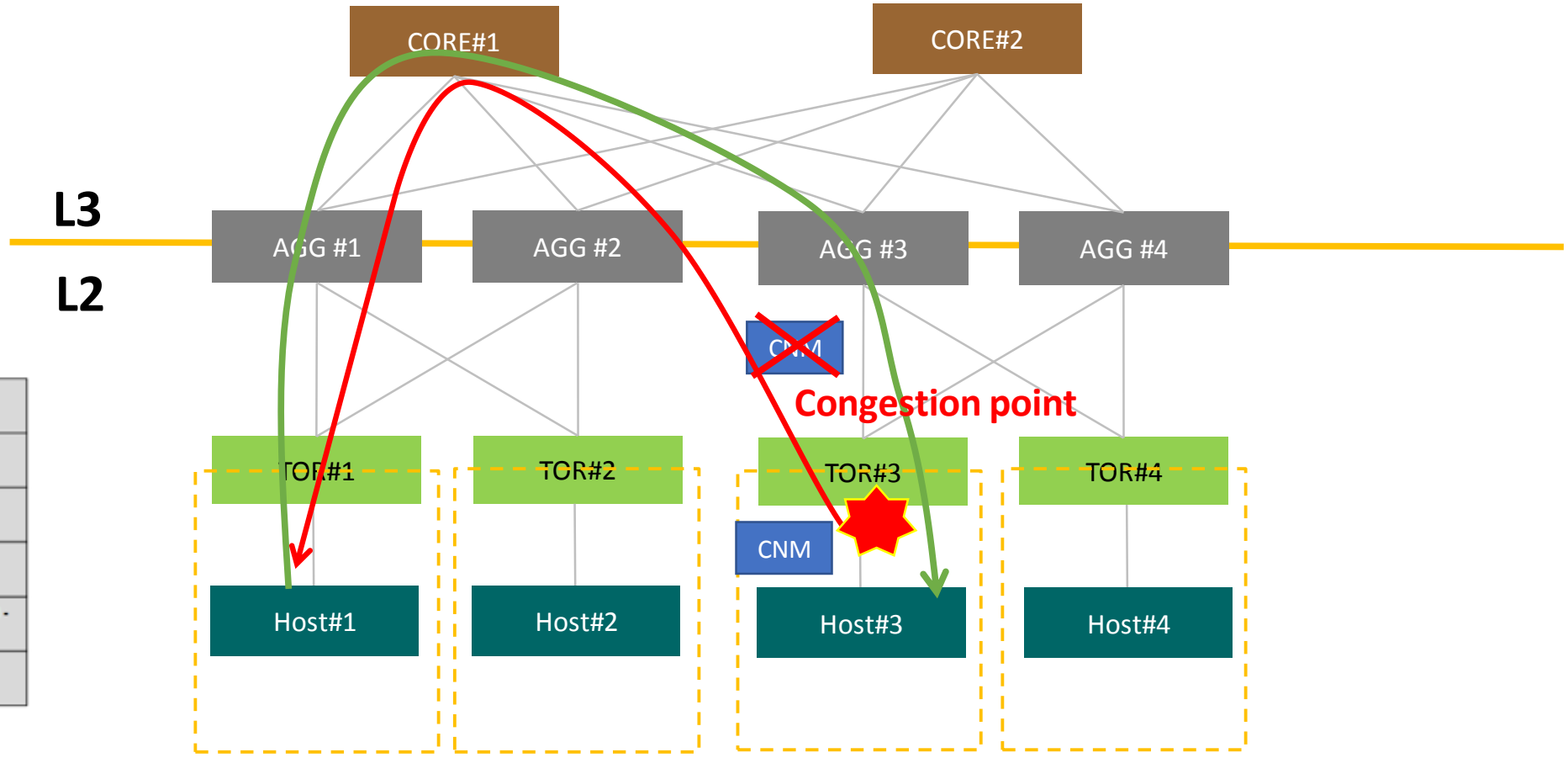




Limitation of CN

- CN could not be used in Layer3 interconnected DC network





The Trend of the Layer3 DC network hierarchy

- Due to the requirement of extremely high throughput, the multi-path L3 network is normally used due to its IP ECMP ability.

- We have visited TOP3 Internet companies and currently all of them are using Layer3 network in their own DC environment.

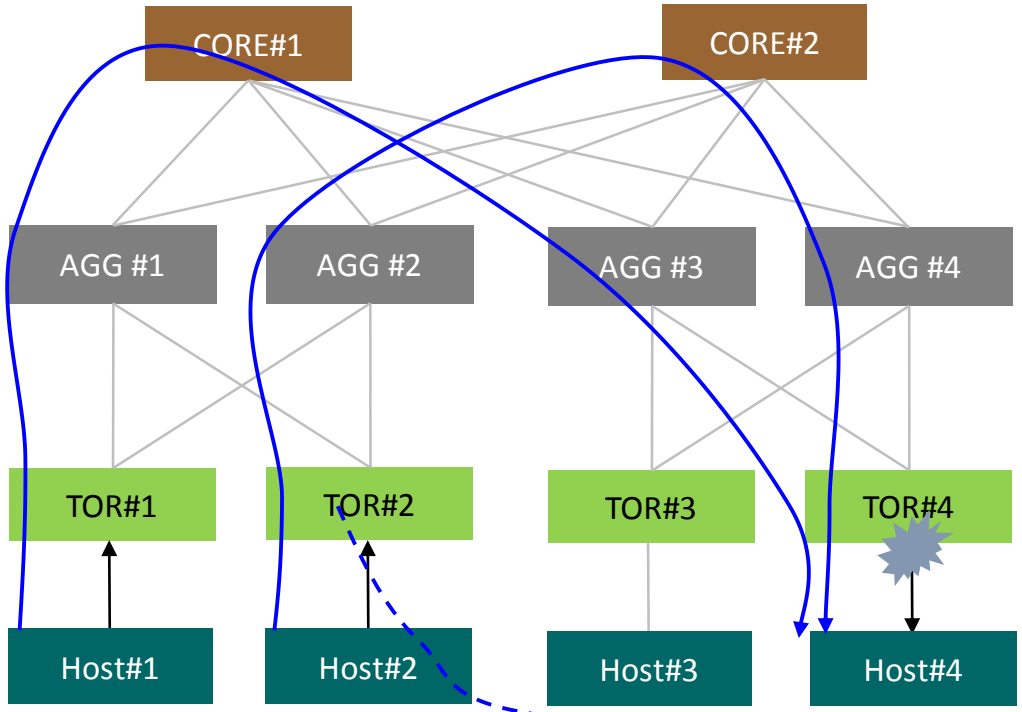
New Requirements of CN

- The essential of CN is to early alarm the potential Congestion
  - It directly reflect the congestion status of the network.
  - It’s an efficient mechanism to reduce the possibility of the network congestion.

- Could we extend QCN to L3 network?

L3QCN in a certain scenario

- General Mindset:
  - Use the constructed (private CNM) to forward the CNM (Congestion Notification Message) on the L3 network.
  - Use the standard CNM between TOR and HOST which are within L2 network.
  - The TOR transfer the private CNM to standard CNM or vice versa

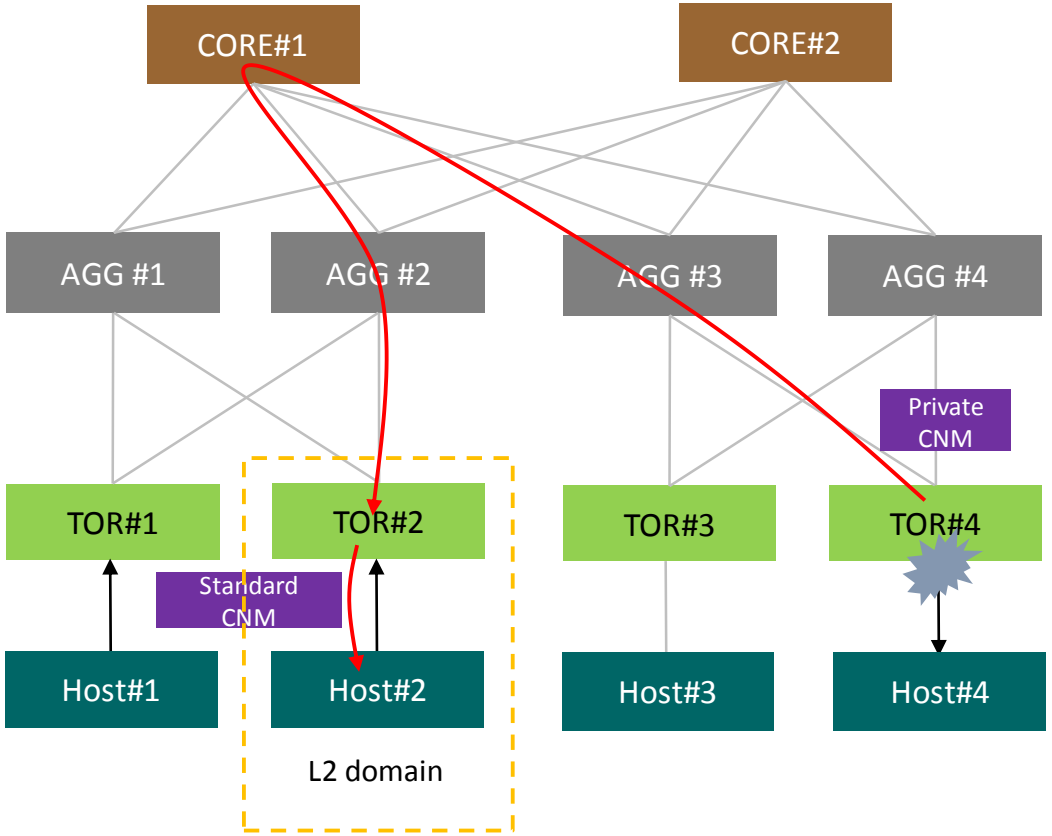


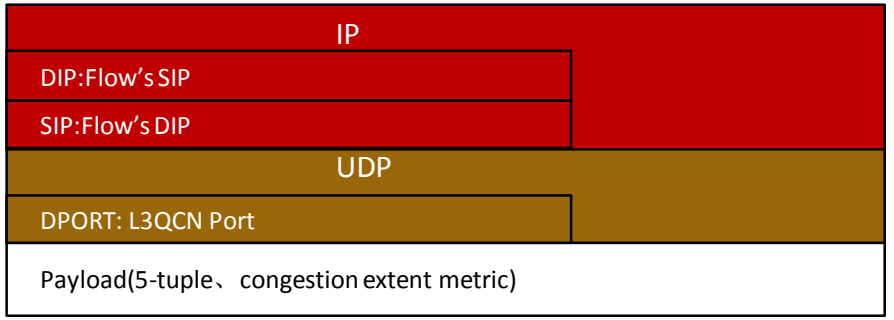
In the TOR, save the Src MAC, FLOWID(RPID), 5-tuple in the local map table. Then use the normal routing function

FlowID Map Table

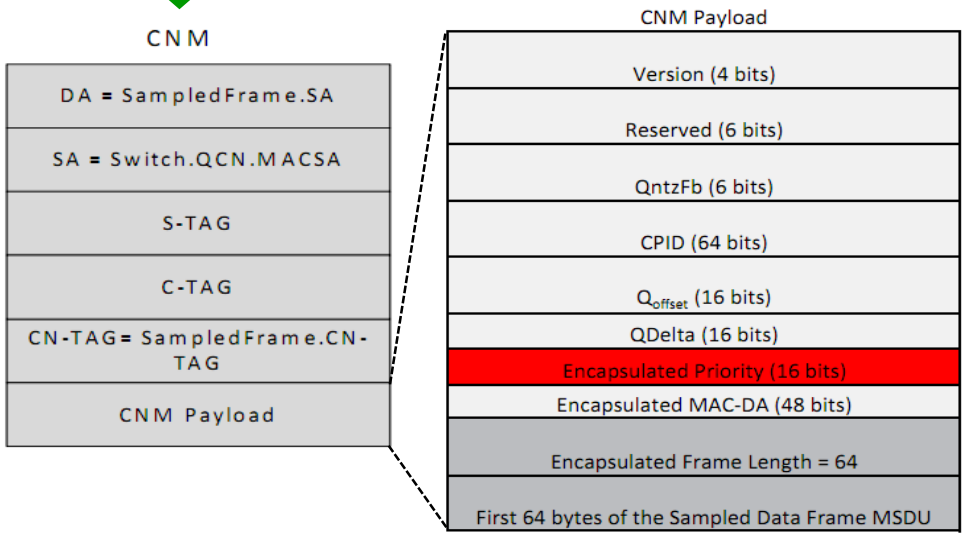
MAC SA	Flow Identifier	SIP	DIP	PROTOCOL	SPORT	DPORT

L3QCN in a certain scenario





Private CNM → Standard CNM



- T4 detected the congestion on the port of T4->H4， judge the congested stream. Constructed the private CNM (5-tuple、Src MAC, RPID). Encapsulate in IP+UDP. Use the specific UDP port. Set the Des IP as the Src IP of the stream to make sure the CNM could be routed to the origin TOR.

More Generic L3QCN

Think about a More Generic L3QCN machenism

1. Tunneling is common used in DC network which may make this scenario more complex, such as VxLAN, NVGRE, GPE MAC-in-MAC
2. Nested tunneling may even increase the complexity.
3. Different network topologies, such as Fat tree, CLOS, …