PFC Deadlock Free in Data Center Network

LLDP use case: Self-learning of switch level and port type in CLOS network

Xiang Yu

yolanda.yu@Huawei.com

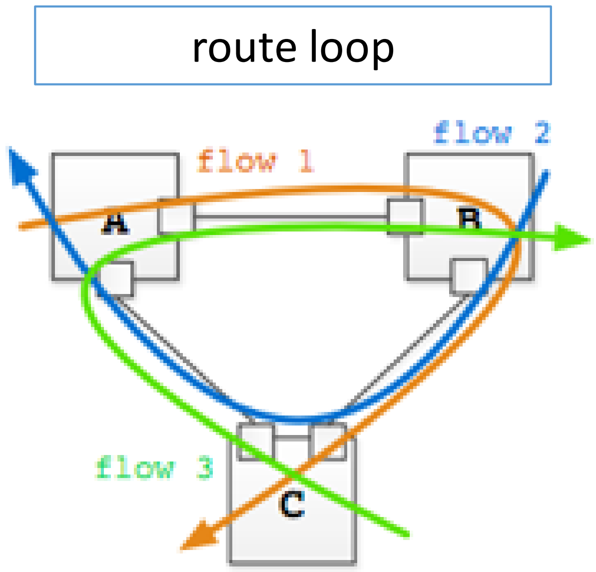
Vienna, Austria, July 2019

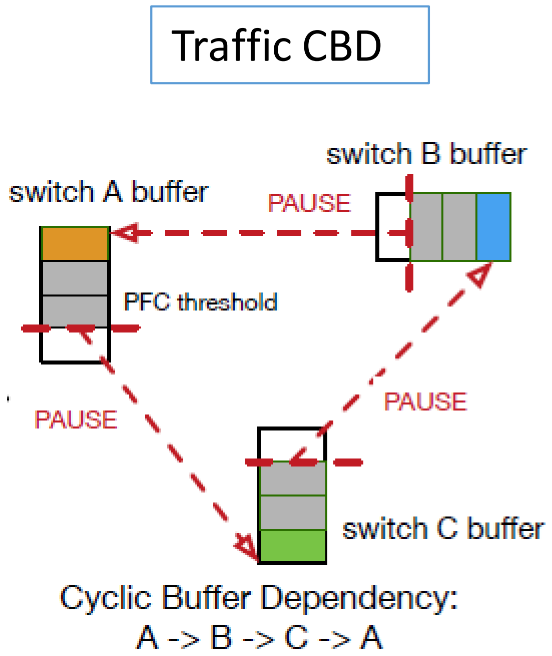
Purpose

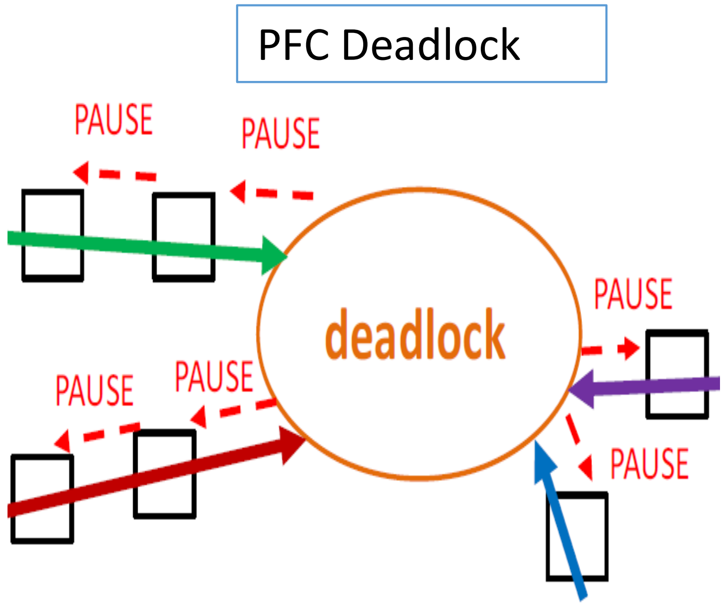
- Define a new TLV to support more switch and port information via LLDP protocol.
- Use case: Avoid routing loop to prevent PFC deadlock
  - Self-learning of switch level and port type in CLOS network

How does PFC deadlock form?

- CBD(Cyclic Buffer Dependency) is a necessary condition for deadlock formation
- Routing loop is a necessary condition for CBD

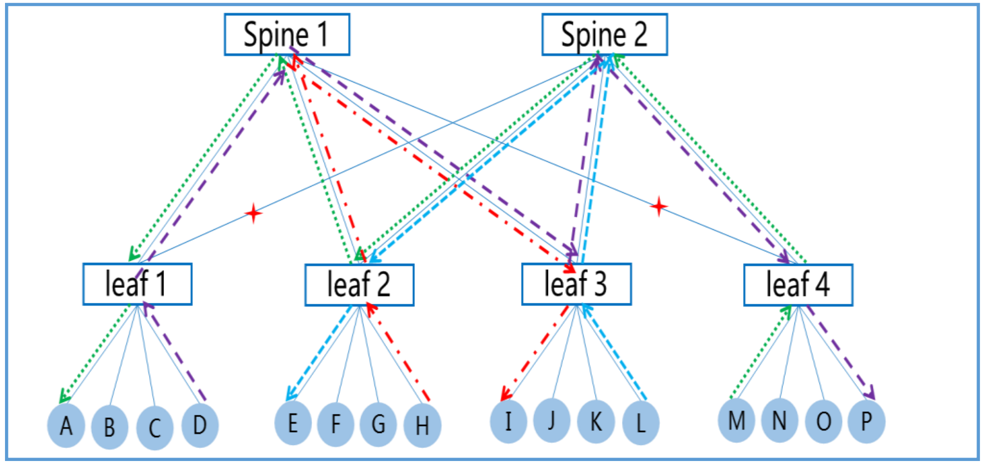




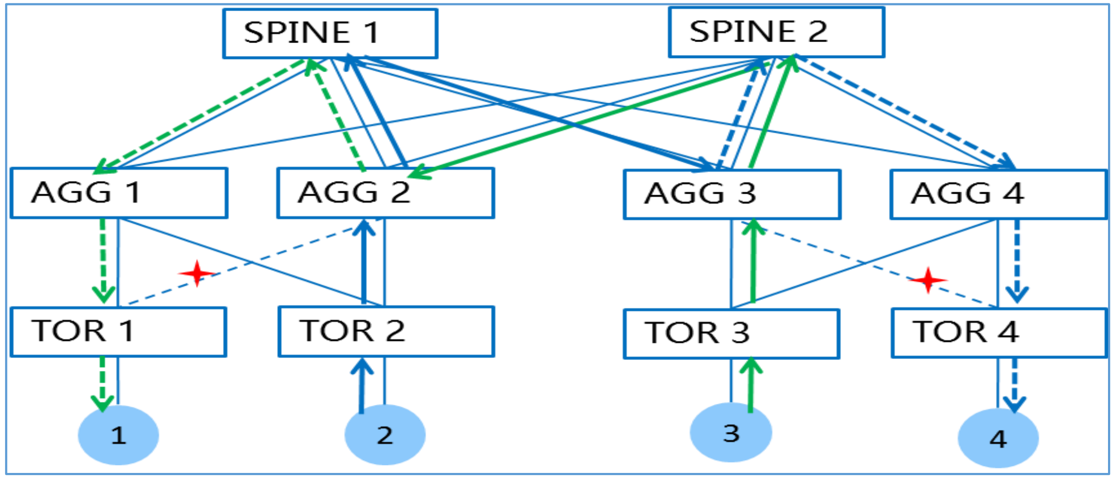


[Reference]: Hu, Shuihai, et al. "Tagger: Practical PFC Deadlock Prevention in Data Center Networks." *Proceedings of the 13th International Conference on emerging Networking EXperiments and Technologies*. ACM, 2017.

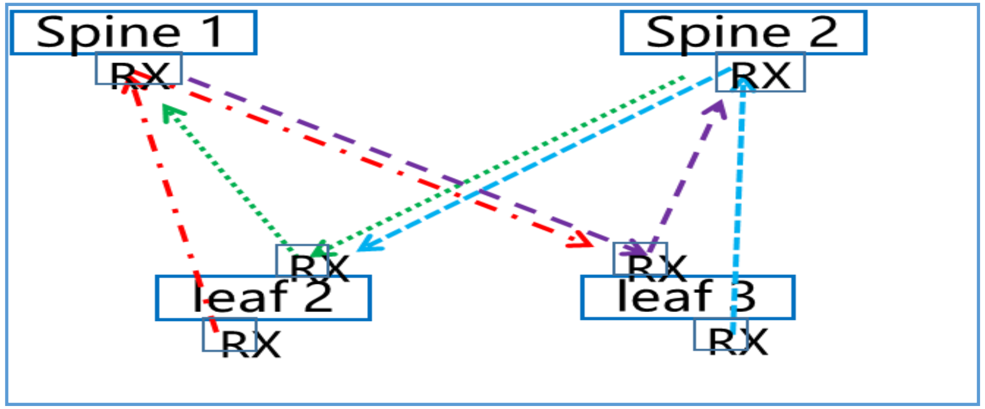
Reproduce the PFC deadlock problem



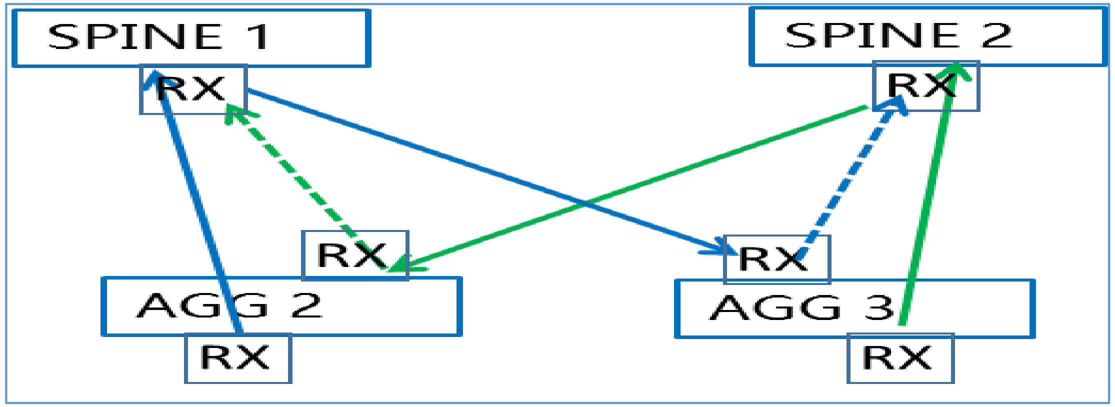
CASE 1 traffic flow: H→I L→E M→A D→P



CASE 2 traffic flow: 3→1 2→4



CASE 1 CBD: Spine 1→Leaf 3→Spine 2→Leaf 2→Spine 1



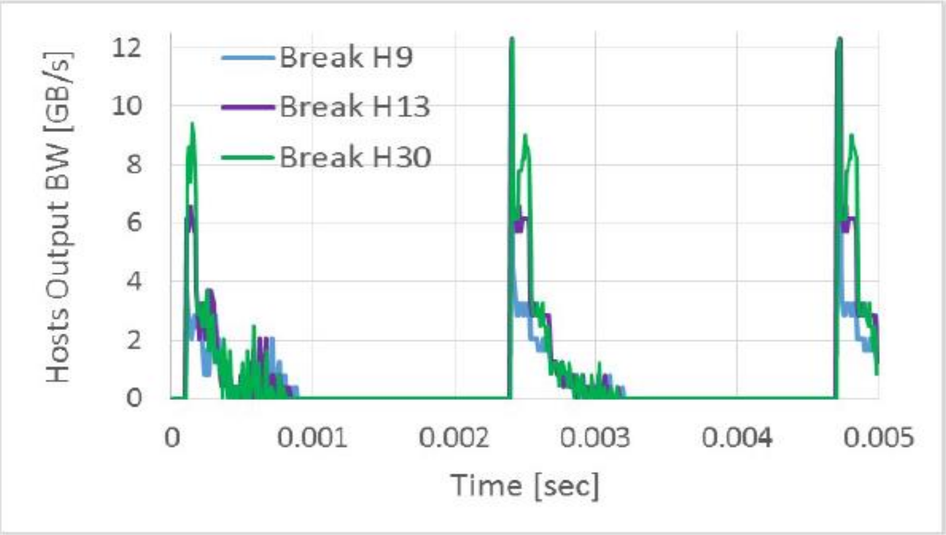
CASE 2 CBD: AGG 2→Spine 1→AGG 3→Spine 2 → AGG 2

- Reproduce the PFC deadlock in both level 2 CLOS and level 3 CLOS network.
- Although CLOS network does not have loops, when link fails, route loop happens and CBD appears. PFC deadlocks may happen.

Current mechanism

Two broad categories:

- **Reactive:** mechanisms/systems detect that a deadlock has formed, and then try to break it by resetting links/ports/hosts etc. May cause the network performance seriously.



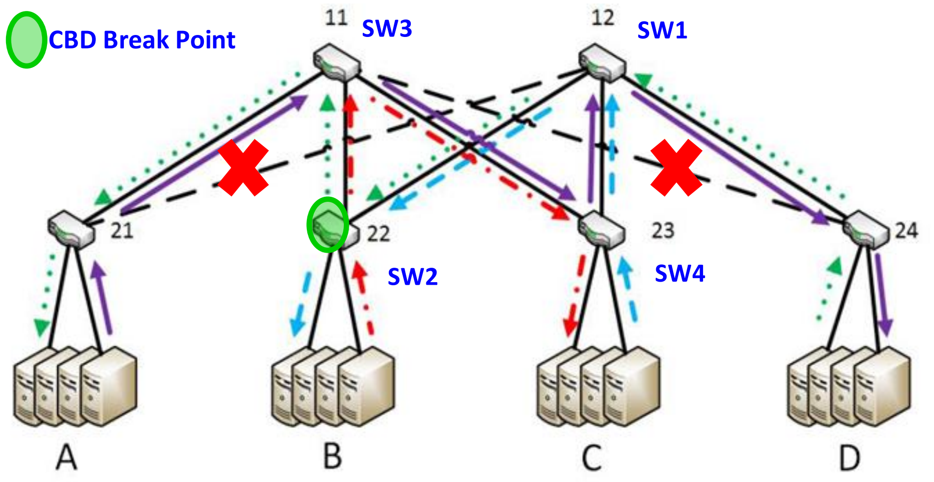
Shpiner, Alex, et al. "Unlocking credit loop deadlocks." *Proceedings of the 15th ACM Workshop on Hot Topics in Networks*. ACM, 2016.

- When a queue continues to be in the PFC-XOFF state for a period of time, it is considered that a deadlock has occurred. Software will trigger to interrupt notification to perform deadlock recovery.
- The software allows the scheduler to ignore PFC-XOFF state of the deadlock queue for some time (configurable) and continue scheduling (send packet to the peer/direct drop the packet)
- If the CBD persists, then it will invoke deadlock immediately after recovery, and the throughput will be greatly affected.

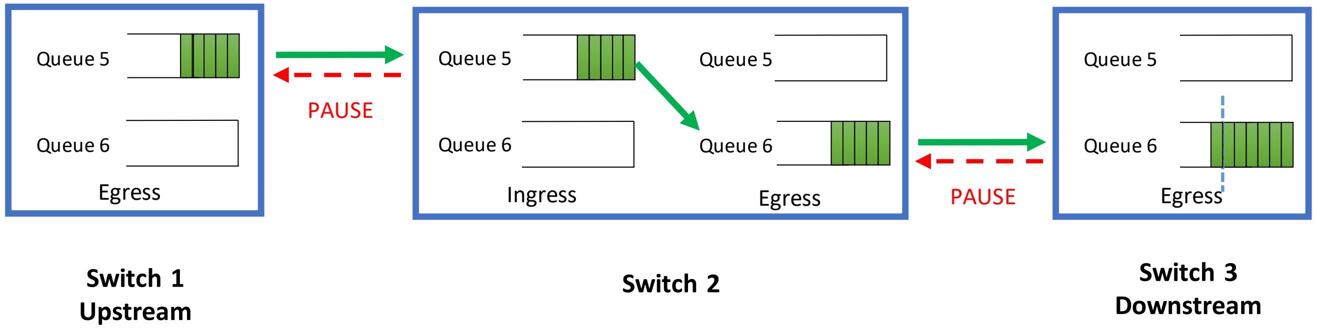
- **Proactive:** deadlock prevention is a more principled approach to this problem.

Deadlock free mechanism (Proactive)

- Identify CBD break point and prevent the PFC Deadlock
- Mindset:
  - Although the traffic in CLOS network itself has no loop, when the link fails or jitters, it may cause the reroute which may form CBD.
  - Use some attributes of the switch or server (Device type, Device Level, Port type) to design a method to judge if packet reroute happens.
  - Use some certain mechanisms to prevent CBD, then PFC deadlock. For example switch the priority queue.

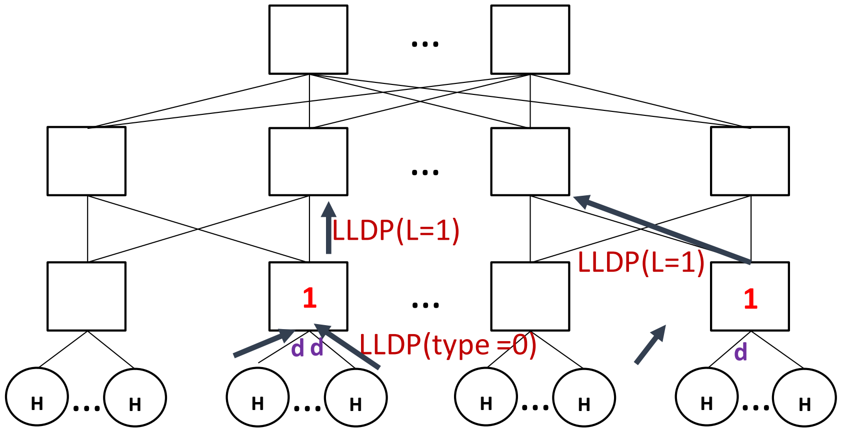


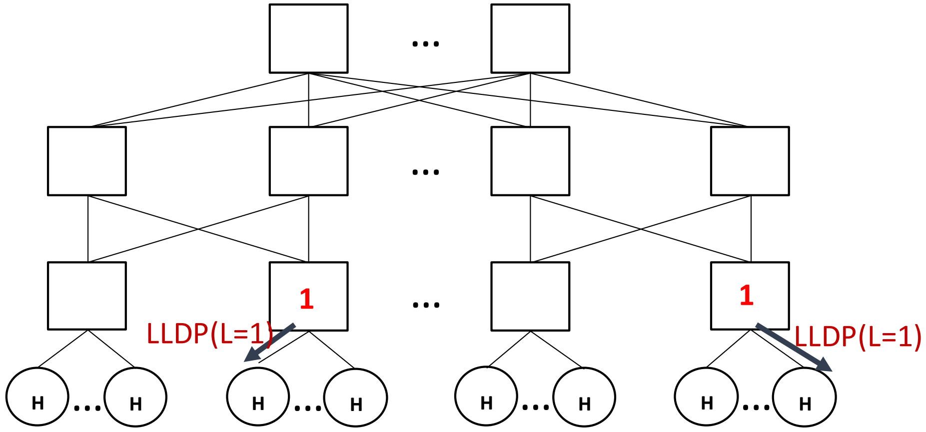
Recognize down-up change, identify the CBD break point



- Both Queue5&6 are lossless queues(Enable PFC)
- Switch 2 judge the packet and enqueue to Queue6, modify the DSCP
- When downstream triggered the PFC on Queue 6 in switch 3, PFC will map to Queue 5 in switch 2.

LLDP carry necessary information -1





Step 1:

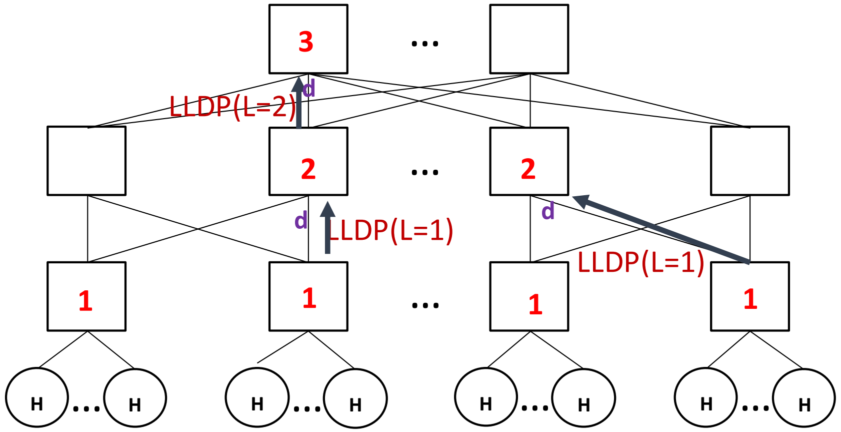
- When the switch receives a LLDP packet, if the Device Type = 0 (host) in the packet, the switch knows that it is the switch closest to the server. If the switch does not have level information or has level ≠1, set its own Device Level to 1, the corresponding port is set to downlink, and the other ports are unmarked.
- If you already have level=1, just set the corresponding Port Type to downlink.

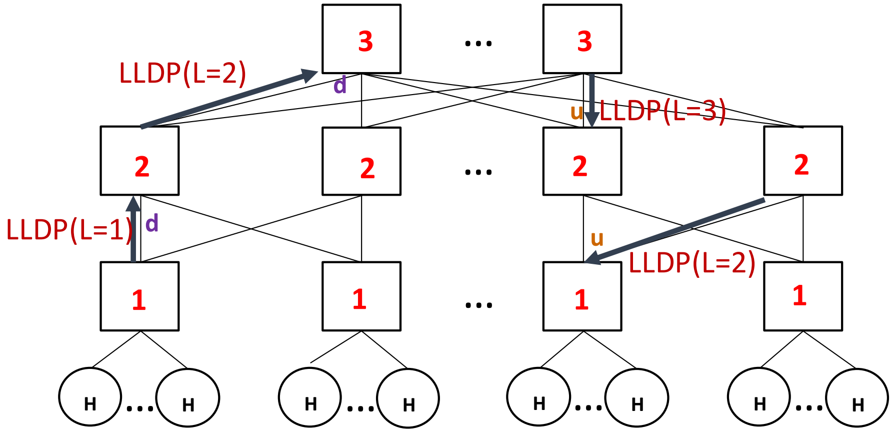
Step 2:

- When the Server (Device Type = 0) itself receives an LLDP packet containing the level information , the level information is ignored.

▹ The LLDP function are enabled both on the servers and switches

LLDP carry necessary information -2





Step 3:

- When the switch (Device Type = 1) received a LLDP packet containing the level=L information
- At this time, If the switch does not have level information and the minimum of the level from LLDP messages received from all the ports is Lm. Set the level of the switch to Lm+1.
- The corresponding port is set to downlink and the other ports are set to unmarked.

Step 4：

- When the switch (Device Type = 1) has its own Device level = N value and received a LLDP packet containing the Device Level = M information.
- If N = M + 1, the corresponding Port Type is downlink. If N = M - 1, the corresponding Device port is uplink.

Support Level/Port type/Device type in Organizationally Specific TLVs

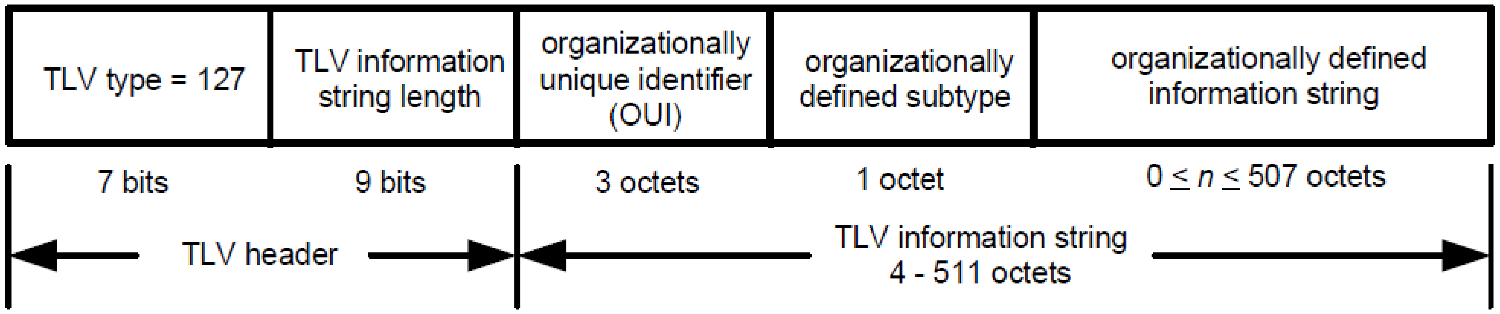


Figure 9-12—Basic format for Organizationally Specific TLVs

LLDP TLV:

TLV Type(7 bits)	TLV Length (9 bits)	OUI (3 octets)	Subtype (1 octet)	Device Type (1 octet)	Device Level (1 octet)	Port uplink/downlink(1 octet)
127		LLDP OUI	0x1	0~0xFF 0： host 1： switch …	0~0xFF 0： Server 1： Level 1 2: Level 2 …	0~0xFF 0： unmark 1： uplink 2: downlink …reserved

Next step

- Define the new Organizationally Specific TLVs.
- Support Device type/Device level/Port type in Organizationally defined information string.

Thank you