Requirements for Adaptive PFC Headroom

Fengwei Qin(China Mobile)

Ruixue Wang(China Mobile)

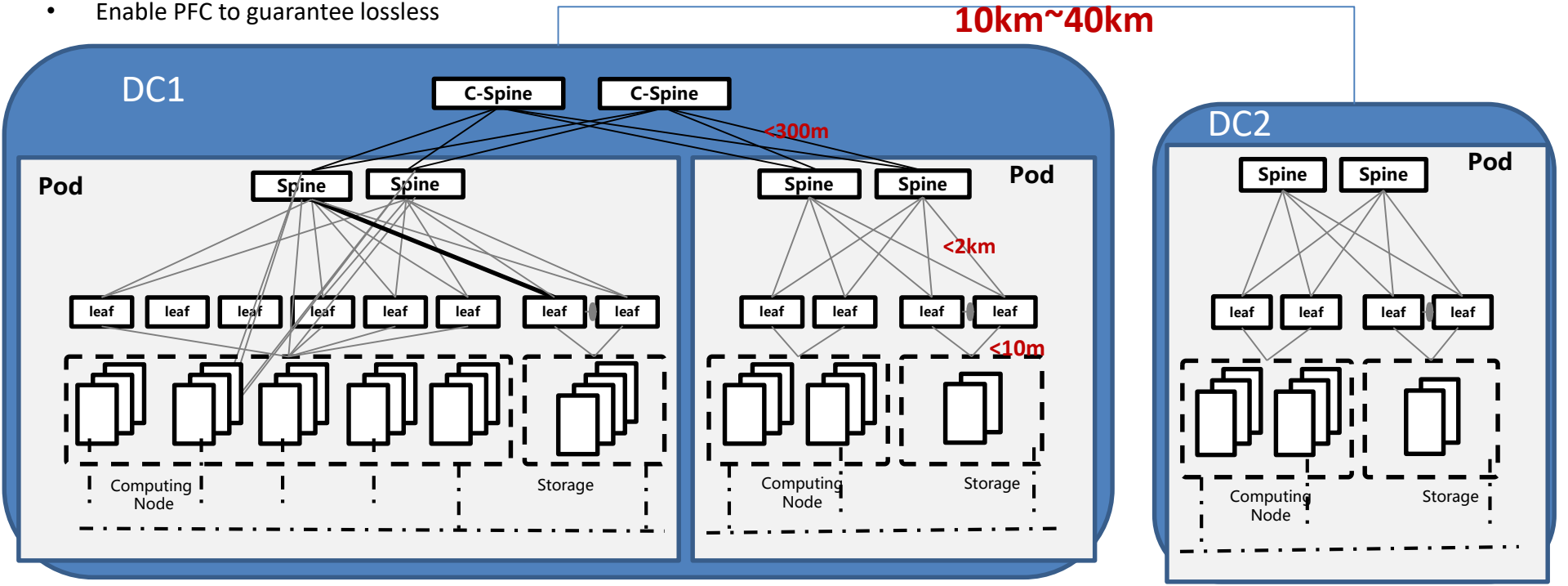
Introduction to CMCC DCN

- Data center service in CMCC
  - Including cloud DC, private DC, public DC
  - The number of DC exceeds 200, with more than 200000 servers

- Requirement of DCN in CMCC
  - Generally, a DC center contains multiple PODs, each pod has its own centralized storage, which requires the losses network
  - In future, one storage pod is deployed in each DC center, other PODs in the DC center will access to the storage POD
  - Meanwhile, replacing FC-SAN with RoCE-SAN step by step

Typical Deployment of CMCC DCN

- CLOS Architecture
  - The link distance is about 100m that accesses the local centralized storage
  - The link distance is from 2km to 5km
  - The link distance will be longer in the DCI scenario, from 10km to 40km , however the distance between two equipment is no more than 10km
- Enable PFC to guarantee lossless



Challenge of PFC Deployment

- No efficient way to configure PFC headroom
  - Link distance varies which cause vendor provided default value does not work.
    - 100m vs. 5km (in case of 100G port)
      - 100m needs ~6KB
      - 5km needs ~300KB
- Need more queues to support PFC
  - Normally only 2 queues are supported by commercial switches, that is not enough to support various services. One major reason is inaccurate PFC headroom estimation.

Requirement of PFC Deployment

- Require automatic configure PFC headroom when network is set up or is changed.
  - As PFC delay is the key to get PFC headroom size, an automatic measurement of delay is required.
- Prefer a simple mechanism to support PFC headroom measurement.
  - Previous discussions in 802.1 suggests several options to support the function.
  - One-stage measurement dedicated for pfc headroom (option1 or option 2) is preferred.
  - 2-stage measurement using PTP link measurement plus separate mechanism (option 3) may have issues.
    - PTP is not used and we do not think it will be used in our datacenter networks.
    - 2-stage procedures seem to add complexity in network maintenance.