P802.1Qau Reaction Point ID Tag

Adding a Reaction Point ID tag to congestion-aware frames

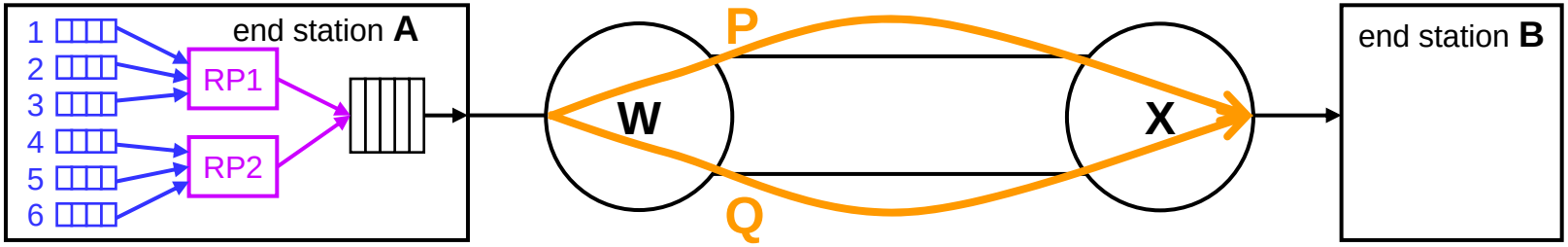
Version 2

Norman Finn

Cisco Systems

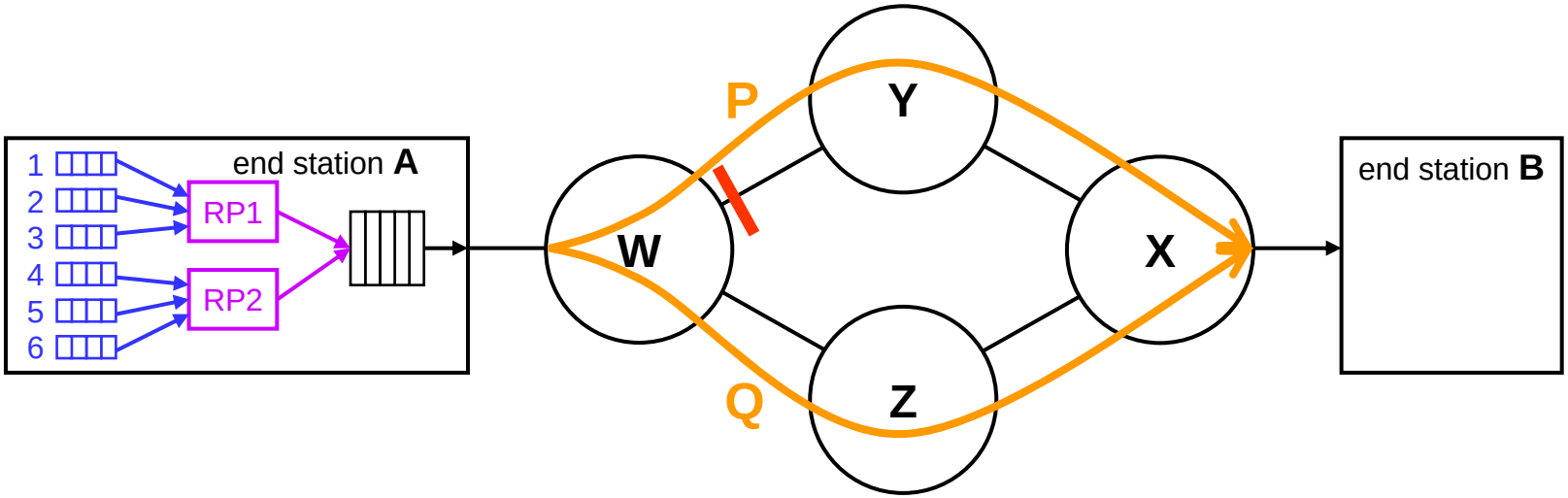
Four issues

Issue #1: Link Aggregation



- Two Bridges W, X. Bridges are doing Link Aggregation based on arbitrary criteria.
- End station A has six flows 1–6 on two RPs.

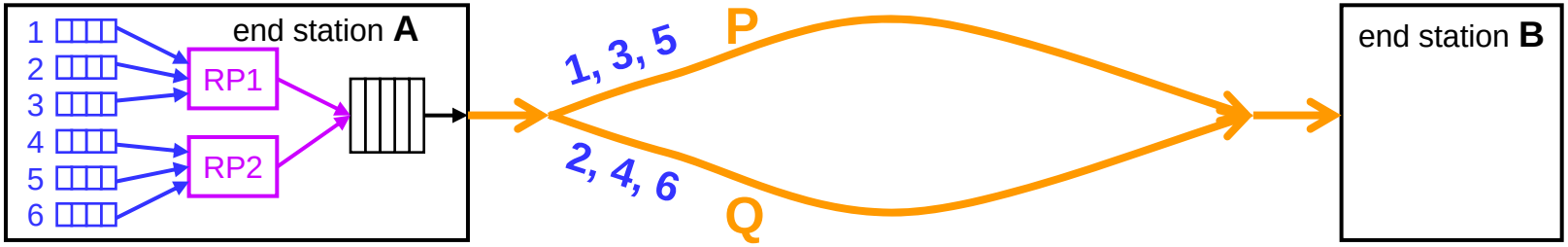
Issue #2: EoNECMP*



- Four Bridges W–Z. Bridges are doing EoNECMP* based on source address, VLAN ID, or other criteria. (Spanning tree and routing protocols can both do this.)
- End station A has six flows 1–6 on two RPs.

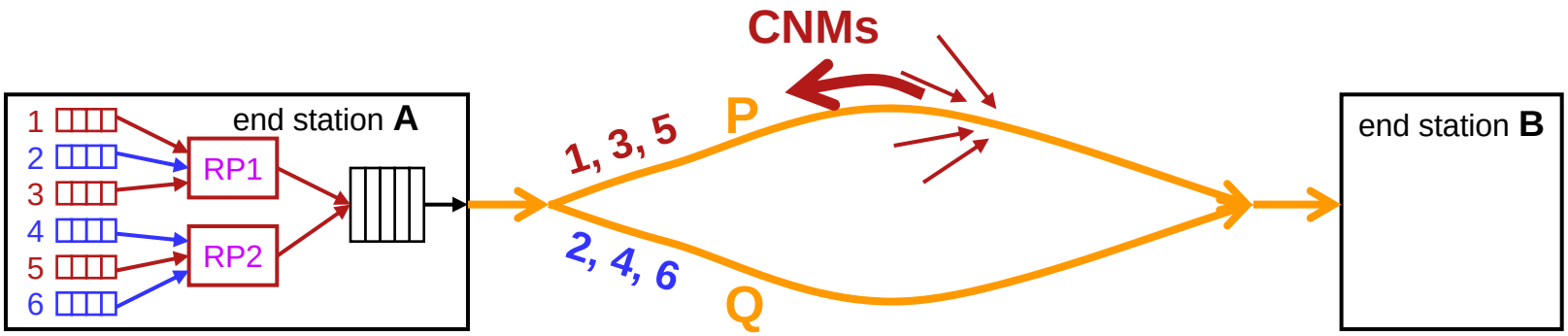
*** Equal or Nearly Equal Cost Multi-Pathing**

Independent selection criteria



- Flow-to-RP and Flow-to-route selection criteria are independent.
- Suppose RP1 has flows 1, 2, and 3. RP2 has 4, 5, 6.
- Flows 1, 3, 5 take route P, and 2, 4, 6 take route Q.

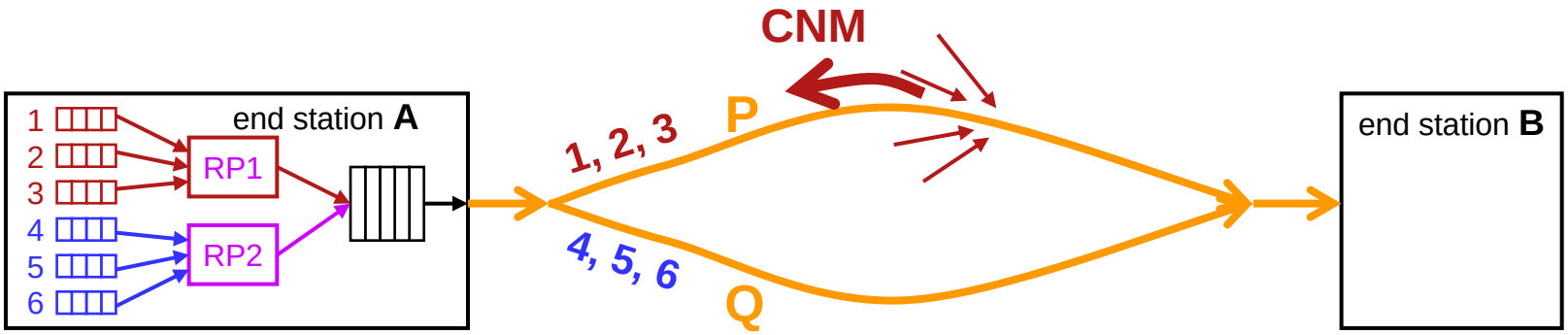
Independent selection criteria



BAD!

- Congestion triggers CNMs on path P.
- Flows 1, 3, and 5 are guilty, flows 2, 4, and 6 are innocent.
- Both RPs and all flows are slowed down.

Coordinated selection criteria



GOOD!

- Flow-to-RP and Flow-to-route selection criteria are coordinated.
- Flows 1, 2, 3 take RP1 and P. 4, 5, 6 take RP2 and Q.
- Congestion on path P affects only RP1's (guilty) flows.

Coordinated selection criteria

- If you have both:
  - Multiple path selection in the network; and
  - Multiple flows per RP;
- And if the flow-to-RP selection criteria are independent of the path selection criteria;
- Then, congestion on one path is likely to affect multiple RPs in a single end-station.
- This is fate sharing at its worst.
- Multiple paths will be common in Data Center networks, because the end stations' data rates will be close to the core's link speeds.

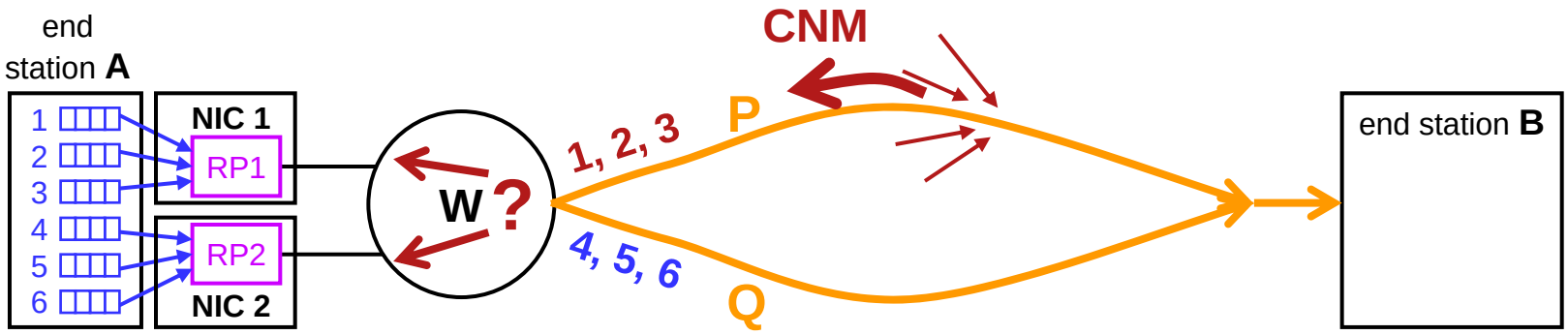
Coordinated selection criteria

- Assuming that an end station has more flows than RPs, then flows share fates when assigned to the same RP.
- The end station is in the best position to know what flows can best share the same fate.
  - This knowledge can be based on information supplied by the applications generating the flows.
  - The end station can also have knowledge of the network topology (more later).
- It is possible to configure higher-layer knowledge in the Bridges, so that they can make the same decisions as the RPs, but this is difficult when the applications are using, e.g., IPsec or secure HTTP.

Issue #3: CNM encapsulated frame format

- The whole world is not necessarily 802.3.
- Other media use 802.2 LLC encapsulation, instead of the Length/Type encapsulation.
- CNMs can be generated on an LLC medium and a frame header returned to an RP on a Length/Type medium, and vice-versa. Therefore, either:
  - All RPs understand both encapsulations, the CNM carries a bit specifying which encapsulation is used for the returned frame header, and no new encapsulations can be invented; or
  - The CP always translates the frame header from the local encapsulation into a canonical encapsulation that RPs understand.

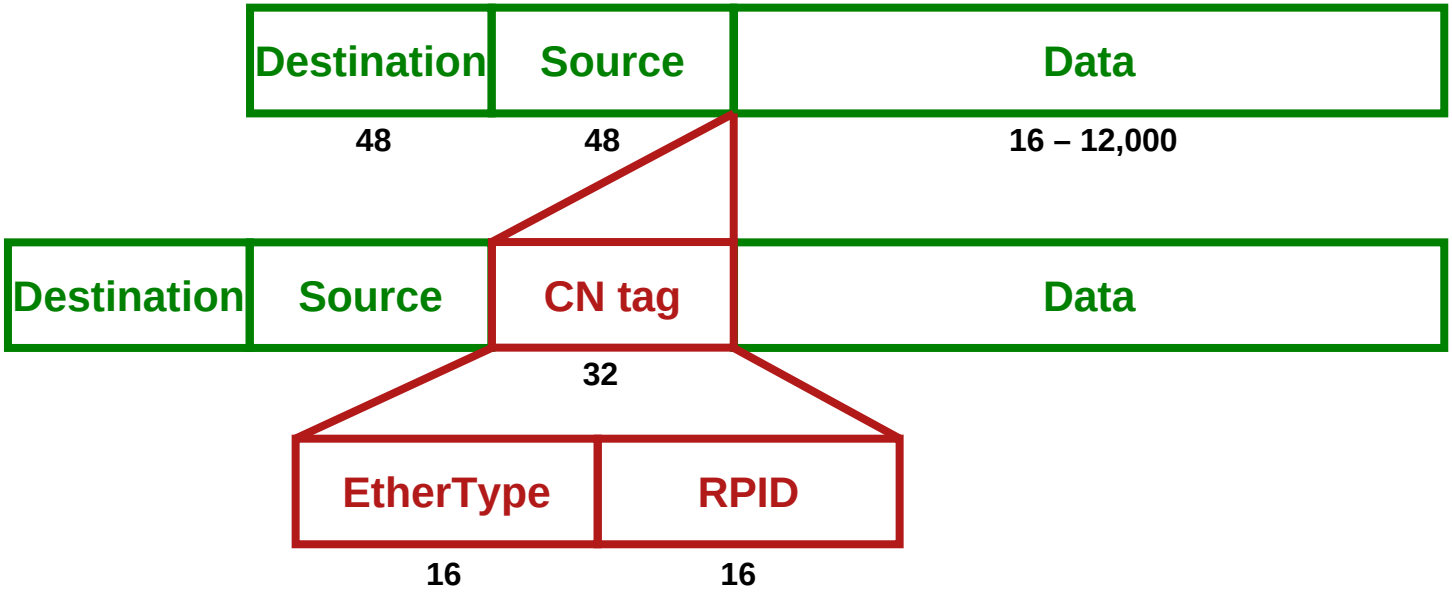
Issue #4: Link Aggregated NICs



- Two Network Interface Cards (NICs) on end station A connect to Bridge W via Link Aggregation.
- When a CNM is returned, to which NIC is the CNM delivered?
- If returned to the wrong one, it takes time for that NIC to notify the right NIC.

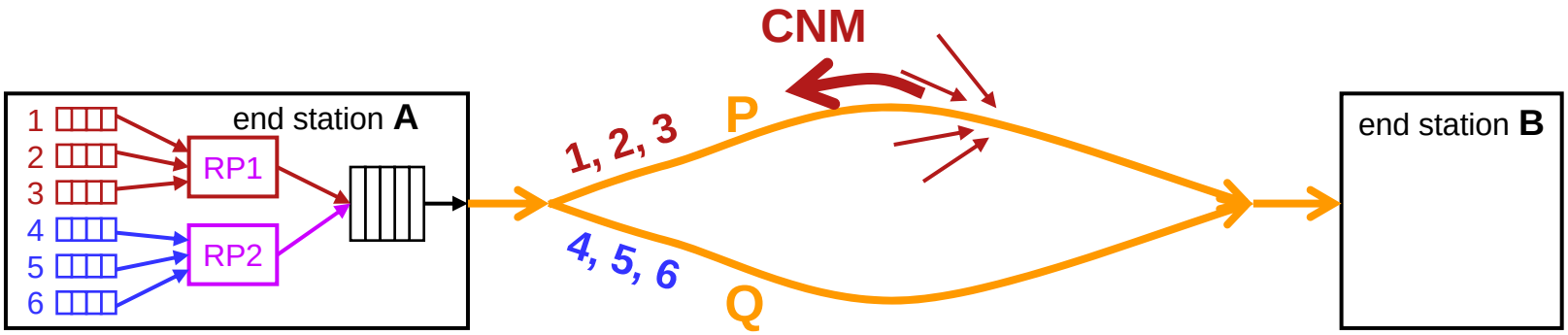
Reaction Point ID tag

Reaction Point ID tag



- 16 bits for the “Reaction Point ID” EtherType.
- 16 bits for a Reaction Point ID.

Issue #1: LinkAg and Issue #2: EoNECMP



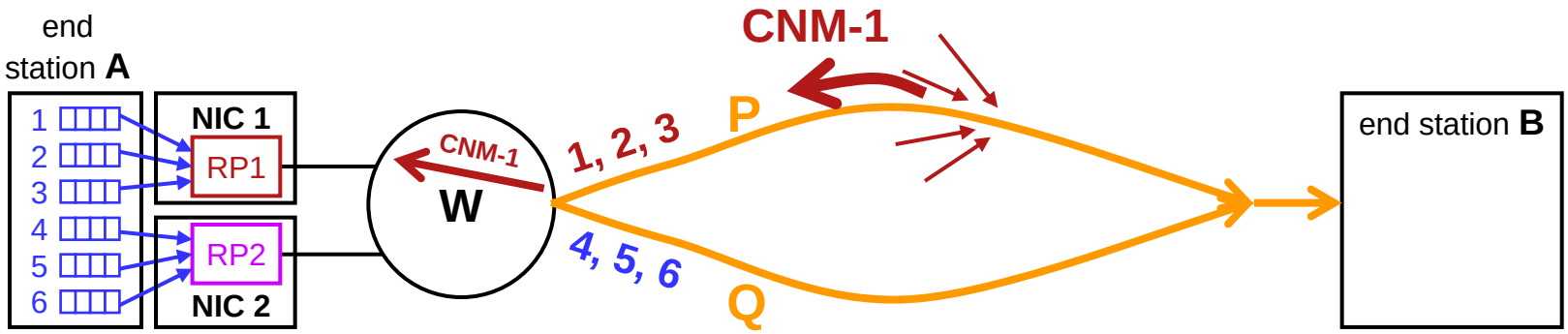
GOOD!

- LinkAg or EoNECMP use the Original Priority and/or RPID tag to select the path for a frame.
- RP selection matches path selection.
- CNM slows down the right RP. **This is good.**

Issue #3: CNM encapsulated frame format

- The CP only needs to return the RPID tag in the CNM; it does not need to encapsulate the offending frame's header.
- The RP needs only to decode the RPID tag in the CNM; it does not need to parse an encapsulated frame.
- Encapsulation translation by the CP and/or understanding "foreign" encapsulations in the RP are not required.
- CP is simpler. RP is simpler. **This is good**.

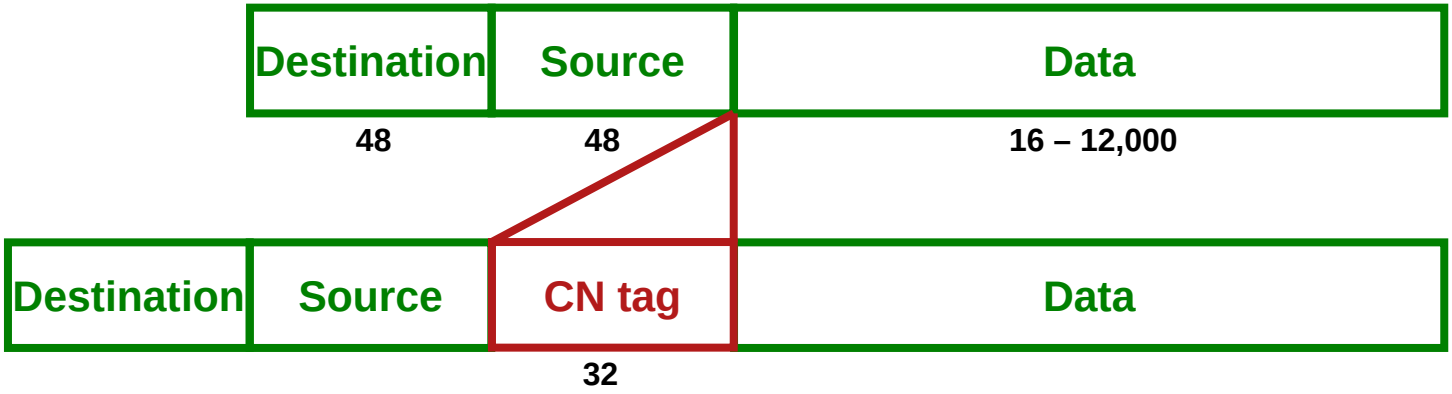
Issue #4: Link Aggregated NICs



GOOD!

- CNM has the same RPID tag as the guilty data frame. If Bridge W uses the RPID in the same way as the NICs label their RPs, then the CNM is returned to the right NIC. **This is good.**

Reaction Point ID tag: downside



- On the other hand:
  - The CN tag adds 32 bits to every data frame.
  - The CN tag must be removed before a frame is delivered to a non-CN-aware end station.
- This is not good. How bad is it?

Reaction Point ID tag: downside

- On the other hand, a minimum length frame is 84 bytes, including the preamble, CRC, inter-frame gap, etc.
- So, even for minimum-length frames, the CN tag adds only (84 + 4) / 84 **< 5%**. (0.4% for 1500-byte payloads)
- Most CN traffic will be among CN-aware stations, so the need to remove CN-tags should be an unusual case.
- The network knows where the boundaries of a CN Domain lie, so knows when it must remove a CN tag.
- **This is not so bad, after all.**

Further notes: Linktrace

- As mentioned above, there is a way for a station to determine the path of a frame through the network.
- A station can issue an 802.1ag CFM Linktrace message to determine the path of a frame. This allows the station to tell whether two different flows will take the same path or not.
- If the Linktrace includes an RPID, either in a CN-tag or as payload, then network path determination will be accurate, and could be used by the end station when assigning flows to RPs.
- Clearly, excess Linktrace activity could impact network performance. But, it is a possibility.

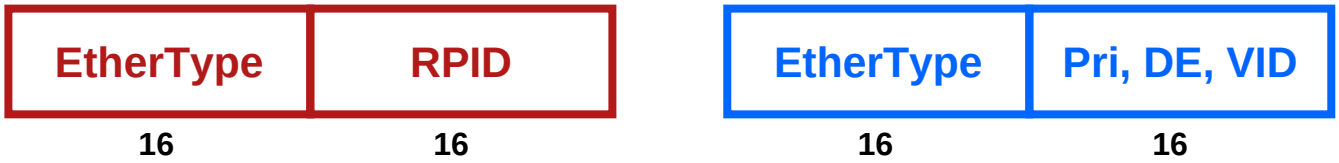
Further notes: RP/LinkAg packing

- If an end station has 8 RPs, and Link Aggregation is splitting flows on only two links, then some fate sharing is inevitable.
- But, if each RP's traffic takes the same physical link, e.g., RPIDs 1, 2, 5, and 6 take one link and 3, 4, 7, and 8 take the other, then the necessary fate sharing is minimized – four RPs' flows will be unaffected by congestion on one aggregated link.

Further notes: RP/LinkAg packing

- Conversely, if many end stations have only one RP, and all are labeled, “1”, a means to avoid piling all traffic on one physical link of each Aggregation must be found.
- If the Bridges use the source MAC address, as well as the RPID, then Issues #1, #2, and #3 are solved, but not #4 (CNM returned to wrong NIC), because the CNMs’ source addresses are different than the data frames’ source addresses.
- So, perhaps edge Bridges use the destination MAC address with the RPID for CNMs, or perhaps each station adds in a MAC address hash to its RPIDs, or perhaps an additive value is assigned the end station by the network or ...?

Further notes: Provider Bridges



- Link Aggregation in the Service Provider world needs flow distribution on a VLAN ID basis, with the same distribution in the reverse direction, in order to ensure that each customer (or tunnel) uses a single physical link. This improves the coverage of Connectivity Fault Management and facilitates error diagnosis.
- The RPID is the same size as the Q- or S-tag payload.
- This is perhaps a happy coincidence of needs: Link Aggregation (or EoNECMP) based on VLAN ID and based on RPID, both on the same sized tag.

Further notes: New flows

- The end station is required to assign an RPID to each frame in a CN priority, and thus to an RP, even for new flows that have not experienced congestion, yet. Otherwise, there is no RPID for the CP to put in the CNM.
- The editor believes that this will simplify the document, as uncontrolled (yet) CN flows will not take a separate path from controlled CN flows.

Opportunities

Opportunities

- Add a CN tag to every frame transmitted using a CN priority value.
  - We can do this, now.
  - We need to pick a solution to the “all RPIDs = 1” issue.
- Refine Link Aggregation to include a means for coordinated use of the CN tag for flow distribution.
  - This is a job for later.