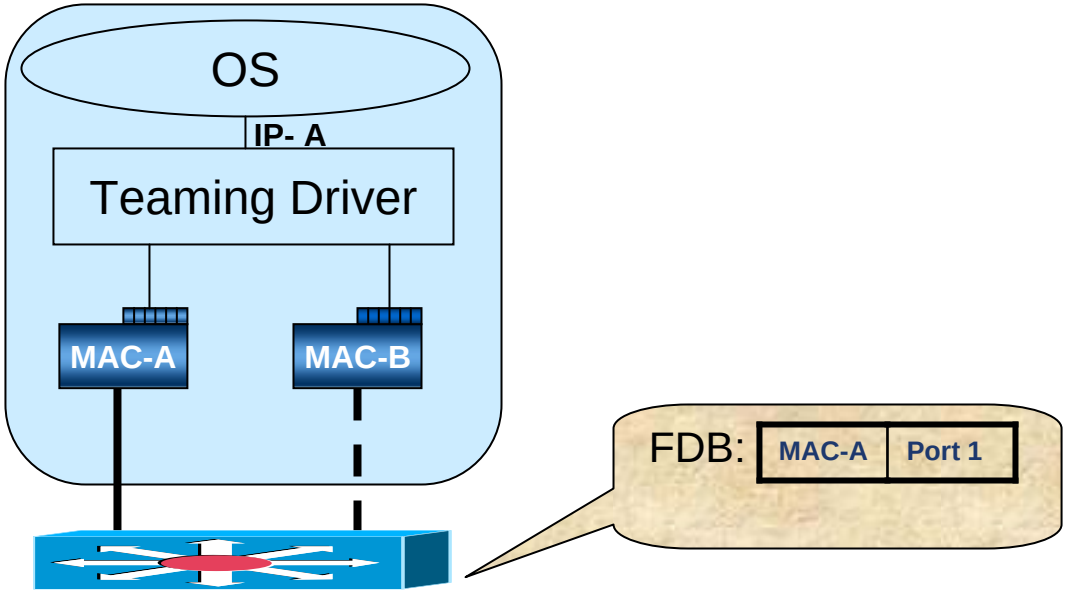
NIC Teaming and CN

Manoj Wadekar

NIC Teaming modes

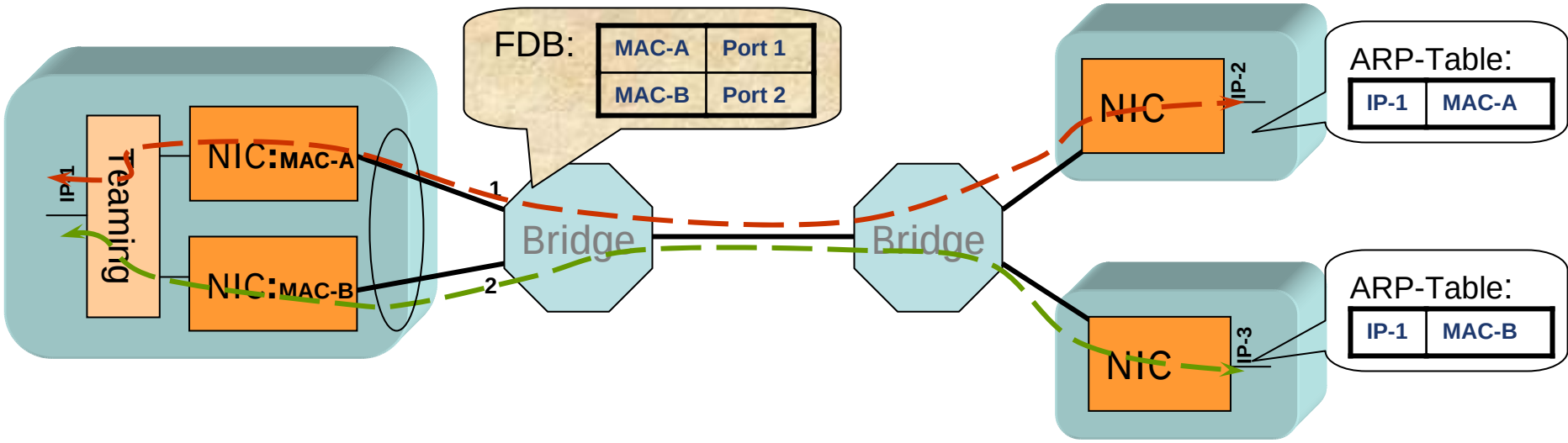
- **Goals of Teaming:**
  - Fault Tolerance
  - Load Balancing
  - Etc.
- **Mode 1: Adapter Failover Teaming**
- **Mode 2: Adapter Failover & Load balancing**
- **Mode 3: Link Aggregation (Static and dynamic – LACP)**

Mode 1: Adapter Failover Teaming



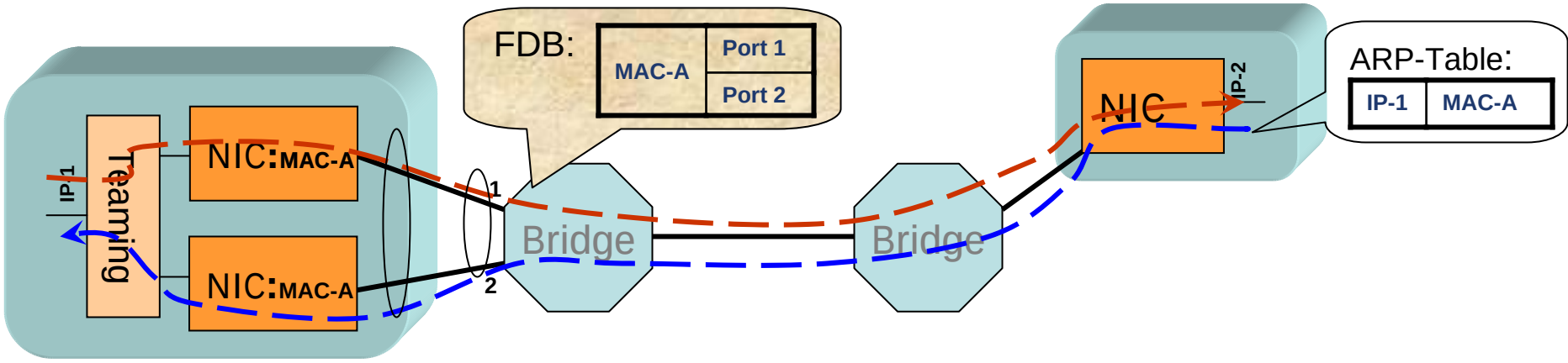
- **Multiple Adapter ports part of a “team”**
- **If one port “fails”, secondary port “takes over”**
- **NIC ports provide functionality**
- **Switch ports are unaware of “Teaming”**
- **Only one port active at a time**
  - So flows are uniquely mapped to appropriate team port

Mode 2: Load Balancing Teaming



- **Load balancing across Team of ports**
  - Single IP/MAC address to OS, however, traffic from each team ports – carries different MAC address
  - ARP responses are “trapped” by teaming driver to provide appropriate MAC address of teamed port
- **Switch ports are unaware of Teaming**
  - So flows are uniquely mapped on each team port

Mode 3a: Link Aggregation - Static

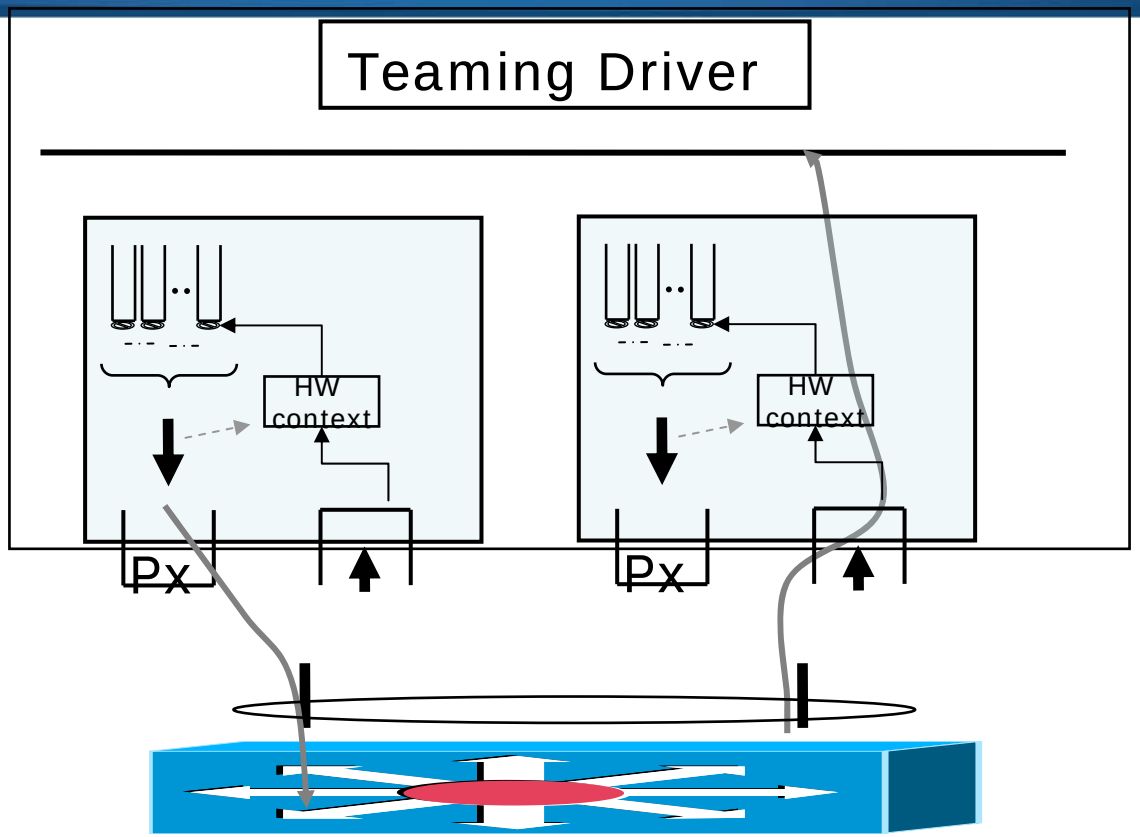


- **Traffic from all team ports carries same MAC address**
- **Adapter and Switch participate in load balancing traffic on team ports**
  - Hashing asymmetrical on two sides of the link
  - Flow mapping is not guaranteed to be symmetrical
  - E.g. IP1→IP2 traffic on teamed port 1 does not guarantee IP2→IP1 traffic will be mapped on to same port 1

Mode 3b Link Aggregation – Dynamic

- Same as Mode 3a
- LACP (Link Aggregation Control Protocol) provides ability to add/remove ports from a team

Challenges of 3a and 3b



- **Multiple ports are aggregated and HW stores context in HW**
  - E.g. Offload information (FCoE, TCP, iSCSI, iWARP) etc.
- **Multiple ports are aggregated and CN reaches wrong port**
  - CN needs to be handled expeditiously
- **Can cause Large latency to CN handling, performance impact to offload functionality**

CN Operation in Teaming Mode

Teaming Mode	Operation	Details
Mode 1 (HA)		CN Forwarded only to active port
Mode 2 (LB)		Each CN carries unique MAC address – delivered to correct port
Mode 3 (LA)		CN can be delivered to wrong port. May result in very high latency penalty

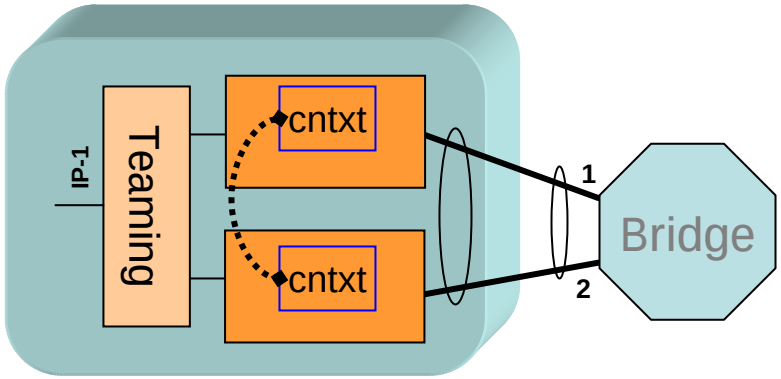
- **CN needs to be handled within "short" period**
  - Crossing PCI may result in high latency → 100s of uS
  - May need to be handled below PCI: Equivalent of "acceleration"
- **Any "context" below PCI will have similar problems**

Potential Solutions

Solution approaches

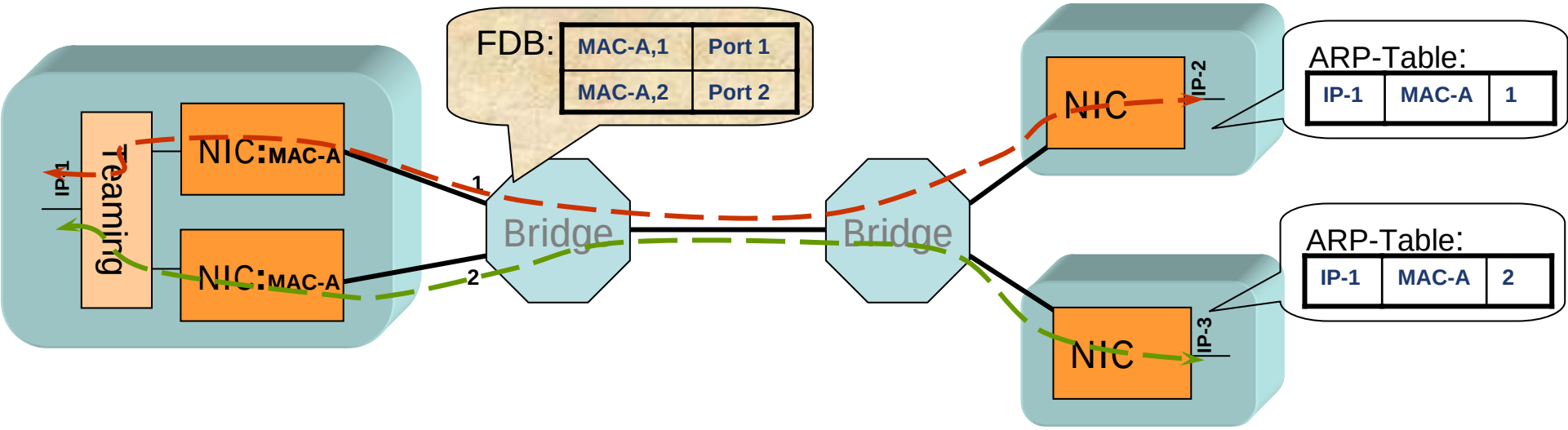
- **Path-1: Solve "HW Context" problem – addresses both data path and control path**
- **Path-2: Solve "control path" only (solves CN issue, but does not address data path)**
  - Needs discussion of use case
- **Path-3: Do nothing**
  - CN is not creating new problem – allow NIC vendors to solve with available solutions

Path 1a: Duplicate “Context”



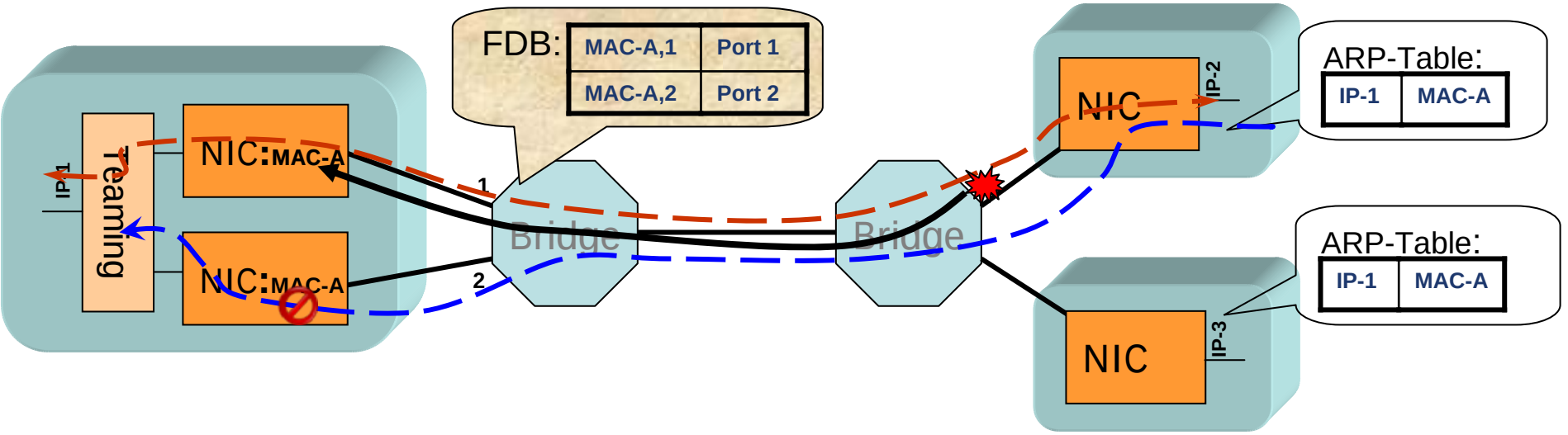
- **Duplicate “context” in all the team ports**
  - “Learn” Rx-port for a given flow and configure “context” in that port
  - Maintain “sync” between Tx and Rx ports
  - Investigate feasibility, race conditions, timeout issues, retransmissions etc.
- **Does not solve CN problem**

Path 1b: Extend End-node identity



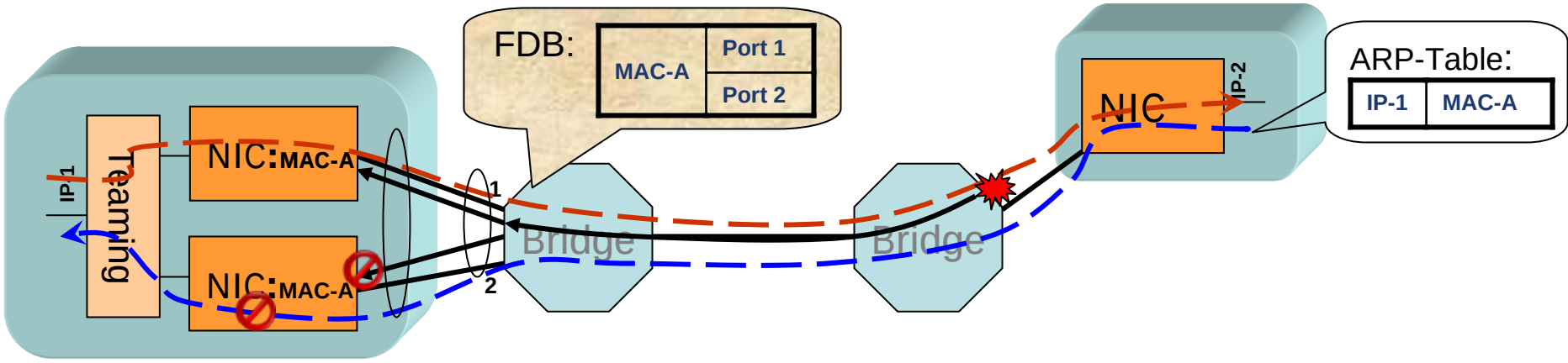
- **Extend MAC-address space**
  - Identify APID (agg-port-id) in packets
  - Edge switches and ARP tables to learn this APID
    - Allows edge switches to forward return packets to appropriate ports
  - Lot of change: NICs, Switches, IP/ARP
- **What are potential benefits over Mode-2?**

Path 2a: Solve CN problem only: APID



- **Identify APID (agg-port-id) in packets**
  - Edge switches learn APID
    - Or edge switches can modify packets with APID
  - CP turns around APID in CN packet
  - Edge switch (before providing packet to end station) uses APID for forwarding:
    - {VID, DA, APID} if EtherType = CN
    - {VID, DA} otherwise
- **How to identify "Edge Switch"?**
- **What are potential benefits over Mode-2?**

Path 2b: Solve CN problem only : Flood CN



- **Flood CN packets on aggregated links to End Stations**
  - If ((dest_port_type is edge_port) && (EtherType == CN))
    - Flood packet to all team_port_members
  - Else
    - Forward packet to dest_port
- **End Stations can potentially use Flow_id to discard CNs on "wrong" ports**
- **How to identify "Edge Switch"?**
- **Requires broadcast of a unicast packet**
- **Data path is still unsolved**

Path 3: Do Nothing

- 802.3ad problem exists prior to CN
- Look for real solution?
- Disassociate FlowId from Link Aggregation discussion


QLOGIC®