Partial Fault Protection

Trey Malpass
Shen Yao

IEEE 802.3 Higher Speed Study Group 10-13 Sept 2007

HUAWEI TECHNOLOGIES Co., Ltd.

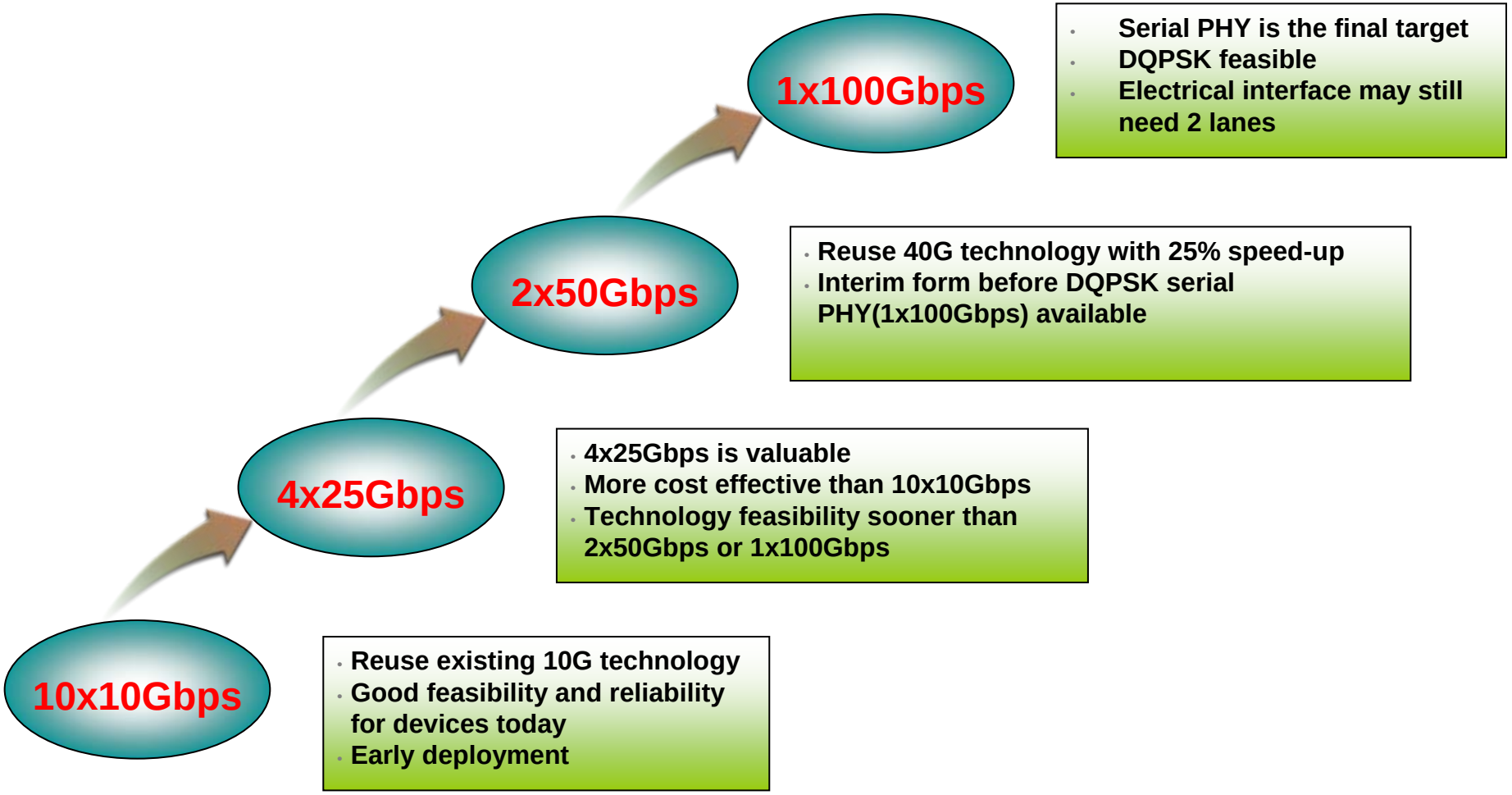
IEEE 802.3 Higher Speed Study Group, 10-13 Sept 2007

Supporters

- **Wenbin Jiang – JDSU**
- **Frank Chang – Vitesse**

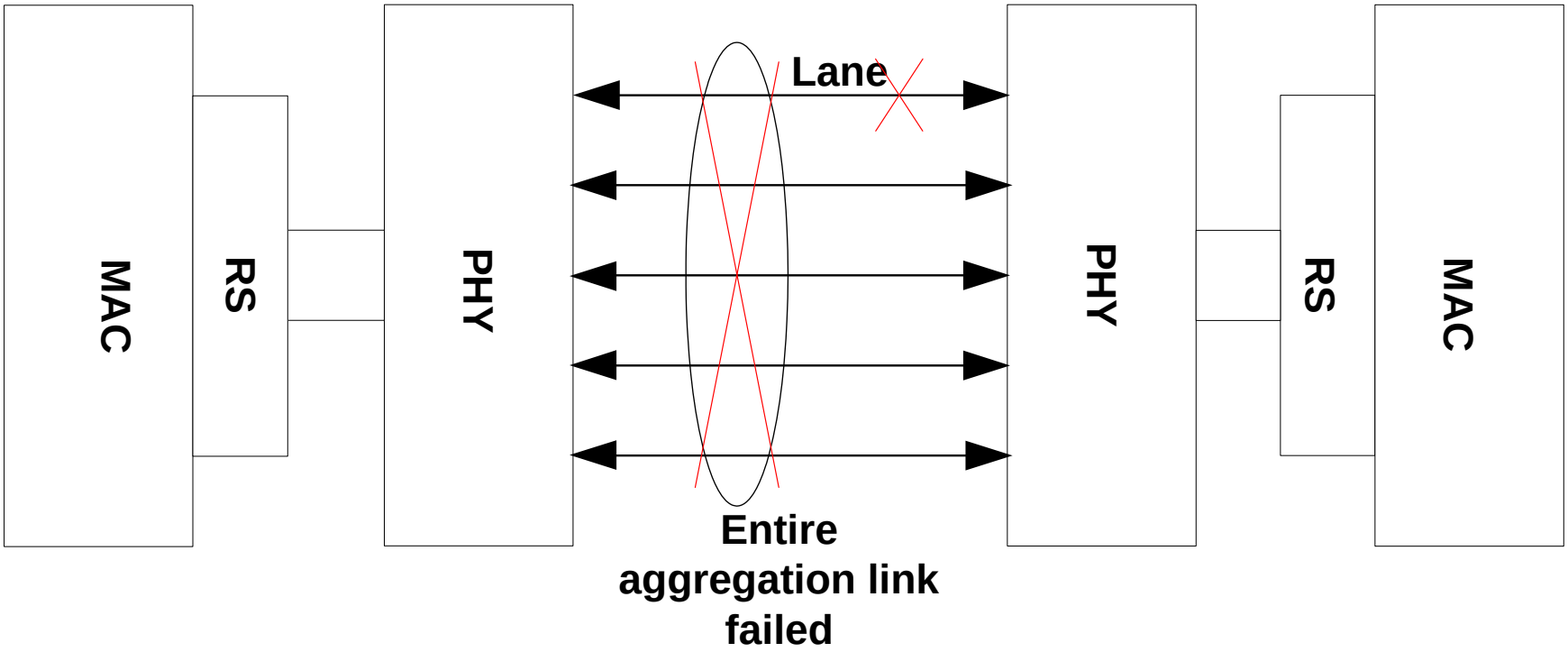
Bundle Multiple Lanes to Form Aggregated High Speed Link

Initially Use Mature and Cost Effective Technology



Individual Lane Fault in a Multi-lane Link

- **Without Partial Fault Protection (PFP)**
  - –Data normally distributed to all lanes, including failed lane
  - –Data can’t be reassembled at receiver, entire aggregation link down



Current Method to Deal With Partial Faults

- **Usually deploy system or network level redundancy for protection**
  - Must plan a complex protection scheme in advance
    - Cost ineffective for single lane fault
    - Difficult to implementation
    - Not easy for everyone who uses 100GE
- **Or develop private protocol to protect the single lane fault**

> • Cray has built resiliency into their protocol. When a lane goes down, the data traffic continues on the other lanes until the computer can be shut down and fixed. Does not protect against a more catastrophic failure than a single lane failure, but fixing single lane failures has value. This may not be the case for applications other than Super-computing. It was indicated that it was complex to add resiliency in this application.
>
> **minutes_01_1106_approved.pdf**

- **We should add an easy and functional method to handle the single lane fault problem**

Other Individual Fault Scenarios

•802.3ad Link Aggregation

- Each link is monitored in order to confirm that the Link Aggregation Control functions at each end of the link still agree on the configuration information for that link
- If one link state has changed,
  - The link aggregation sublayer is informed to change configuration and new configuration information is communicated to the corresponding Link Aggregation Partner
  - At the remote end, the link aggregation sublayer updates the configuration according to received information
- Failed link will be removed from the aggregated links to keep aggregate link alive

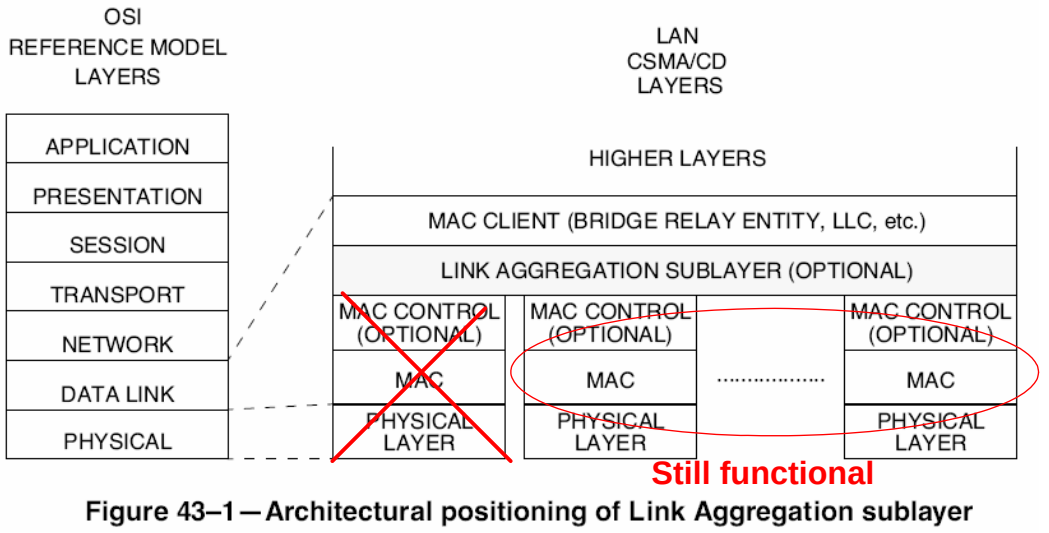


Figure 43–1—Architectural positioning of Link Aggregation sublayer

Other Individual Fault Scenarios

- **802.3ah PME Aggregation**
  - Up to 31 PME instances are aggregated at PME Aggregation block
  - Each link is monitored individually
  - Link state is notified to PME Aggregation block through γ interface, such as TC_Link_state signal, etc.
  - Link state is exchanged between partner PMEs through remote PHY register access
  - PAF only transmits on active PME (TC_link_state is OK)
  - Only active PME puts fragments to PAF
  - Individual fault won't cause an aggregation failure

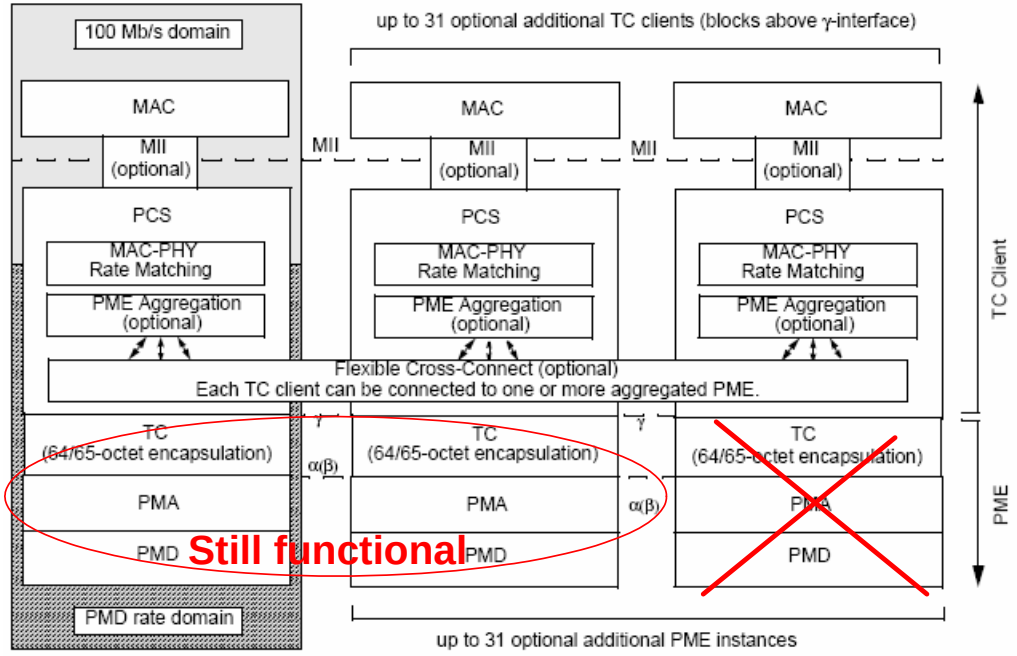
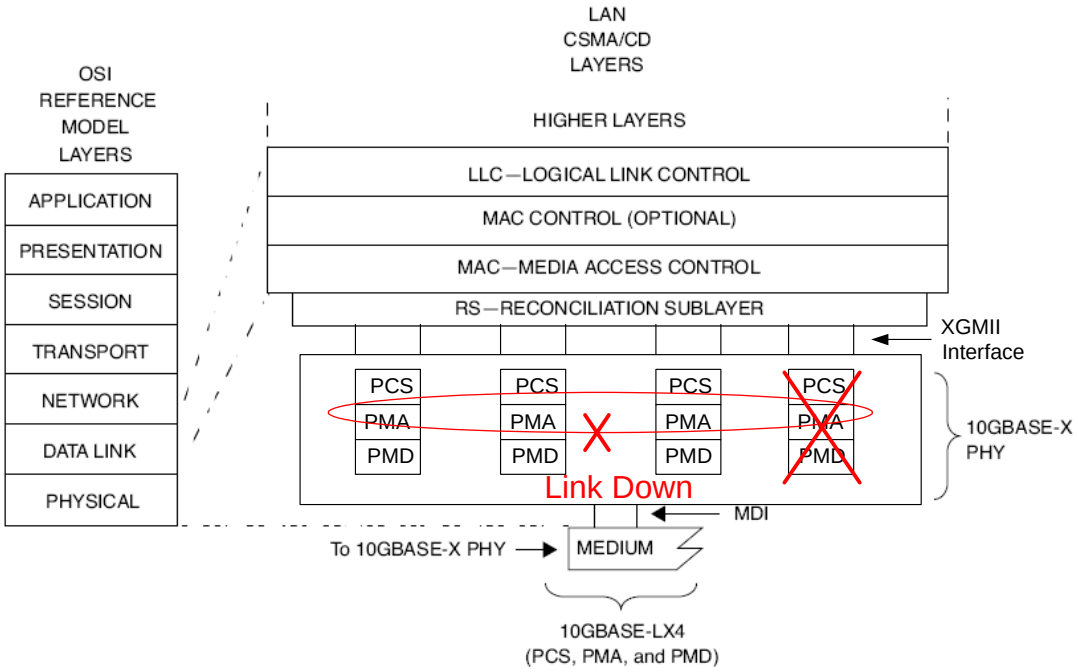


Figure 61–2—Overview of PCS functions

Other Individual Fault Scenarios

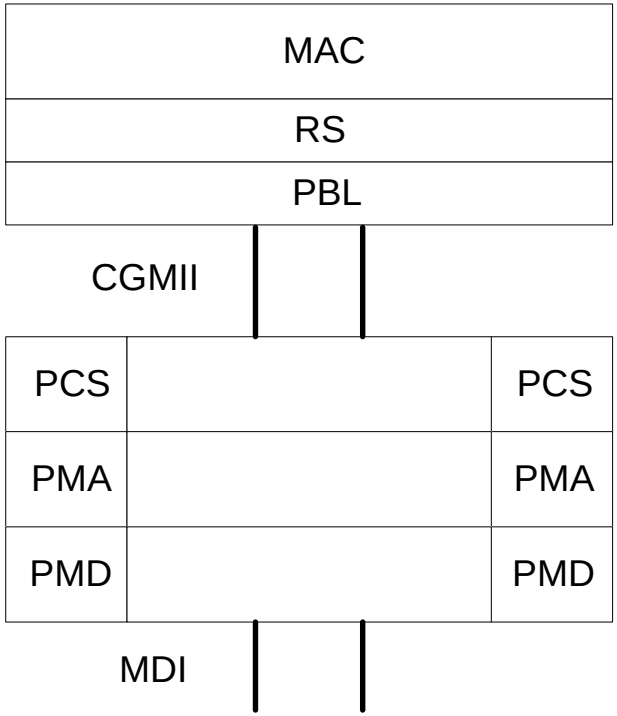
•802.3ae 10GBaseX, 8B10B coding interface



- Monitor link state at PHY layer
- When fault detected,
  - PCS notifies local RS with LF (Local Fault) signal
  - RS transmits RF (Remote Fault) signal to other end to notify link state
- At remote RS
  - Received RF signal, transmit IDLE
- Do not distinguish the single lane state and the whole link state
- Single lane fault cause whole link down

Proposed HSE PHY layer Model

- Multi-Lane Model
- Similar to 10GBaseX, deploy new distribution and combination scheme as described in malpass_01_0907
- PBL (Physical Bundling Layer)
  - Perform block distribution and combination
  - Data distribution based on 64-bit blocks
  - Multi-lane alignment

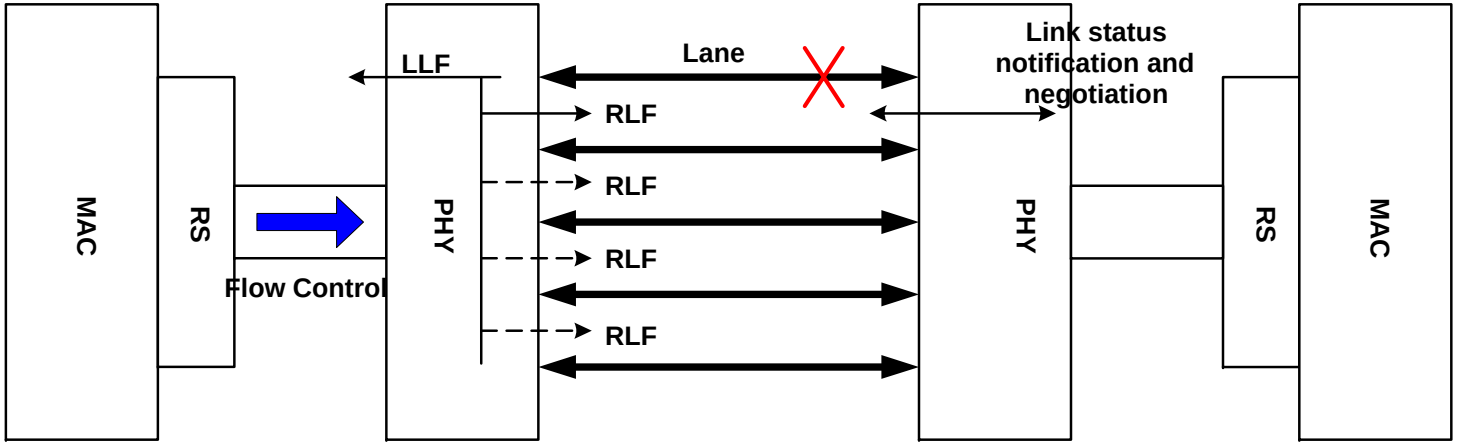


Partial Fault Protection at PHY Layer

- **Add partial fault protection at PHY Layer**
  - Deal with single lane fault, quickly recover from link down status
  - Distribute data on other available lanes to keep link alive
  - Allow users to check and resolve single lane problem
  - Hitless recovery when single lane fault is cleared

Single Lane State Monitor and Notification

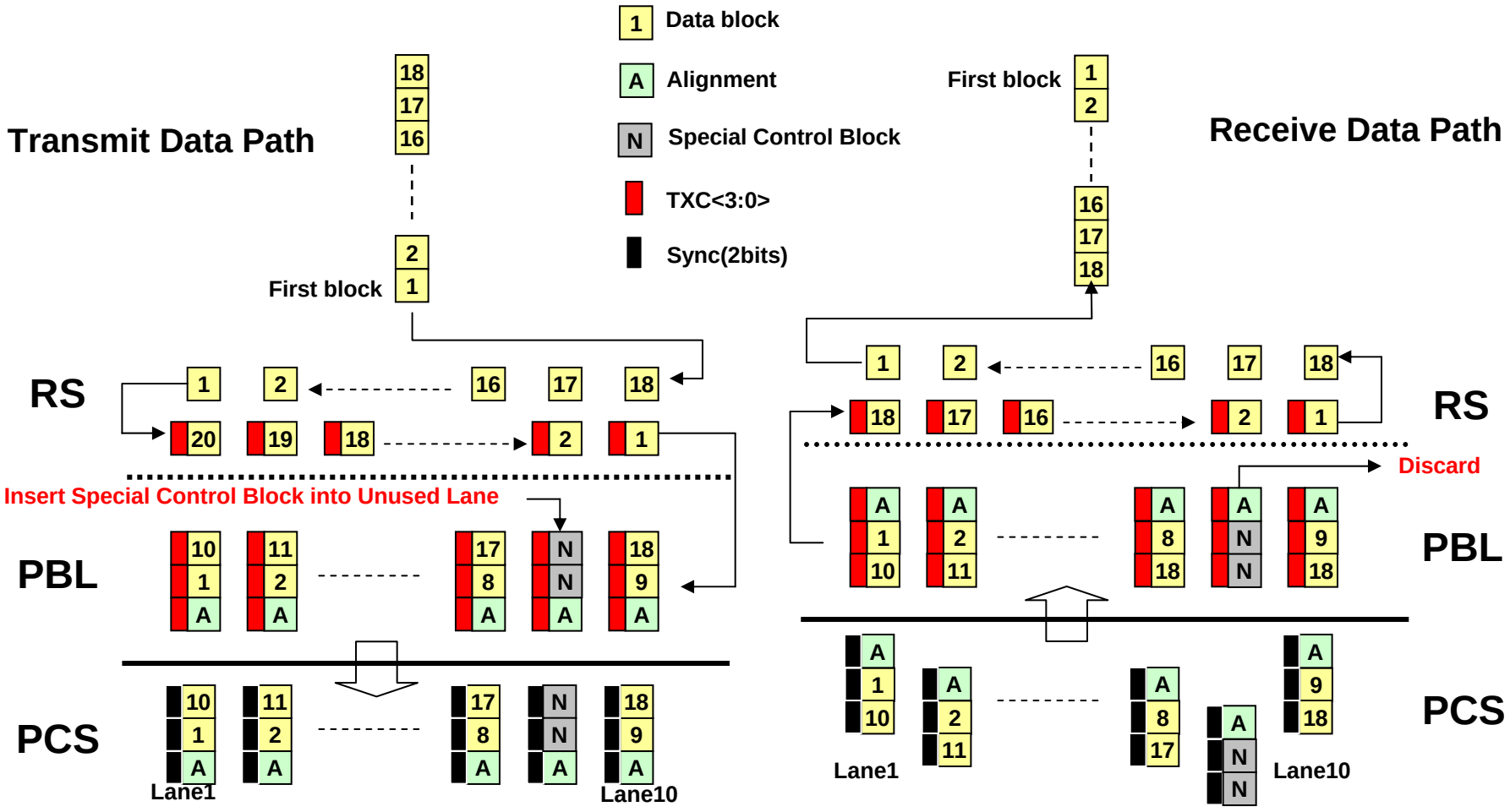
- **Single Lane State Monitor**
  - Monitor individual lane state instead of entire link state, different from 10GBaseX
  - Single lane state may include LOS, BER, Out of Sync, etc.
  - Single lane state differs from link state
- **Single Lane State Notification**
  - Notify local end entity of lane state change information on a per-lane basis
  - Communicate to remote end with the single lane state change information
    - PHY OAM, Sequence ordered_set, etc.
  - Introduce two new signal: LLF & RLF
    - LLF: Local Lane Fault, to inform local entity the lane fault state
    - RLF: Remote Lane Fault, to inform remote partner the lane fault state



Distribution & Combination Scheme

- **Distributes data to working lanes at PBL using round robin scheme**
- **Unused lanes filled with special control block, such as Sequence Ordered_set, NULL, ERROR etc. For example, using new block type NULL indicates to receiver to discard blocks during reassembly**
- **Combine discrete data blocks at receive PBL, discard special control blocks or other non-data block on unused lanes**

Distribution & Combination Scheme



Partial Fault Protection

- **Single Lane Fault**
  - At transmitter, distribute data on available lanes, and special control blocks on failed lanes
  - At receiver, discard anything received on failed lanes during reassembly

- **Single Lane Fault Recovery**
  - Continue distribution of special control blocks on recovered lane until lane state is confirmed
  - Then distribute and reassemble data blocks on recovered lane

- **Use LLF&RLF signal to exchange lane state information between two endpoints**

Benefits

- **The aggregation layer (PBL, etc.) on both ends can flexibly deal with single lane state changes differently from aggregation link state changes**
- **Individual lane state changes do not impact whole link**
- **Scalable and resilient to lane faults at PHY layer**
  - avoid link down during single lane fault
  - hitless recovery when single lane fault cleared
  - Additional resilient bandwidth advantage
  - Based on current PHY layer monitor mechanism (Refer to 802.3ae-2002, 46.3.4 Link fault signaling), with minor changes
  - Easy implementation
  - Do not need complex configurations in advance, plug & play operation
  - Available for everyone and everywhere who deploys multi-lane Ethernet

Summary

- **Partial fault protection can be implemented when deploying multi-lane Ethernet**
- **We propose to:**
  - Monitor individual lane state at the PHY layer
    - Distinguish the individual lane state and the whole aggregation link state
    - Notify lane state information to local and remote ends
  - A distribution and combination scheme that can adapt to a dynamic number of lanes
    - Consider the lane state during distribution and reassembly
    - Only distribute and reassemble data on available lanes
    - Hitless switching when lane fault is cleared
- **Flow control may need more consideration**

Thank You

HUAWEI TECHNOLOGIES Co., Ltd.

IEEE 802.3 Higher Speed Study Group, 10-13 Sept 2007