Architecture for a 10-Gigabit Ethernet Standard

Shimon Muller
Sun Microsystems Computer Company

IEEE 802.3 High Speed Study Group
June 1, 1999

Outline

- Introduction
  - Market Requirements
  - Customer Requirements
  - Standards Requirements
- Multi-Gigabit Transmission Options
- 10-Gigabit Ethernet Objectives
- Architecture
  - Functional Partitioning
  - Principles of Operation
  - Physical Coding and Framing Requirements
  - Media Independent Interfaces
  - Error Handling
- Summary

Introduction --- Market Requirements

- **The expectation is that scaling Ethernet to 10Gb/s will necessarily expand the scope of its application space**
  - **“Ethernet everywhere” (or at least “wherever you can get away with it”)**
    - **Everything from data center and computer room clusters, through traditional LAN backbone and desktop applications, to MAN/RAN/WAN**

Introduction --- Customer Requirements

- **The motivations and constraints for each one of the envisioned applications are quite different**
  - **Long Haul**
    - **MAN, RAN and WAN**
    - **Must operate over existing very long fiber links**
    - **Requires high coding efficiency**
    - **Not very sensitive to cost**
    - **Does not address any specific problem for traditional Ethernet users**
  - **Intermediate Haul**
    - **Traditional LAN backbones**
    - **Must operate over the existing cabling infrastructure**
    - **Coding efficiency is not an issue**
    - **Cost sensitive**
    - **Will address *future* problems of backbone congestion in Ethernet LANs**

Introduction --- Customer Requirements (continued)

- **Short Haul**
  - **Computer room clusters, server-switch and inter-switch interconnects**
  - **Cabling infrastructure is not an issue**
  - **Coding efficiency is not an issue**
  - ***Very* cost sensitive**
  - **Will address *existing* hot spots in today's networks**

Introduction --- Standards Requirements

- "One solution that fits all" for the 10-Gigabit Ethernet Standard will be sub-optimal for at least one major market segment
- The standard should be able to accommodate multiple solutions that will address the divergent market requirements
- The various options must be constrained to the Physical Layer
- Specific Requirements:
  - The MAC should be modified one more time and made "truly" speed-independent
  - Media independent interfaces should be defined below the MAC
  - Several Physical Layers can be defined, that attach to these interfaces
  - This approach is fully compatible with previous 802.3 practice

Multi-Gigabit Transmission Options

- **"Go Faster"**
  - **Architecturally straightforward**
    - **Scale everything by a factor of 10**
  - **Implementation and cost challenges**
    - **Requires very high-speed Physical Layer transmission components**
    - **Significant portions of logic need to be clocked at very high frequencies**
- **"Go Wider"**
  - **Striping of the data stream across multiple transmission channels**
    - **Can be implemented using proven existing technology**
    - **Alleviates the very high-speed logic design requirements**
    - **Will provide a much cheaper alternative in a variety of network environments**
    - **The transmission channels can be separate physical links (ribbon fiber) or a single physical link that carries multiple logical channels (WDM)**
- ***The 10-Gigabit Ethernet standard should accommodate both the serial and the parallel transmission schemes***

Multi-Gigabit Transmission Options (continued)

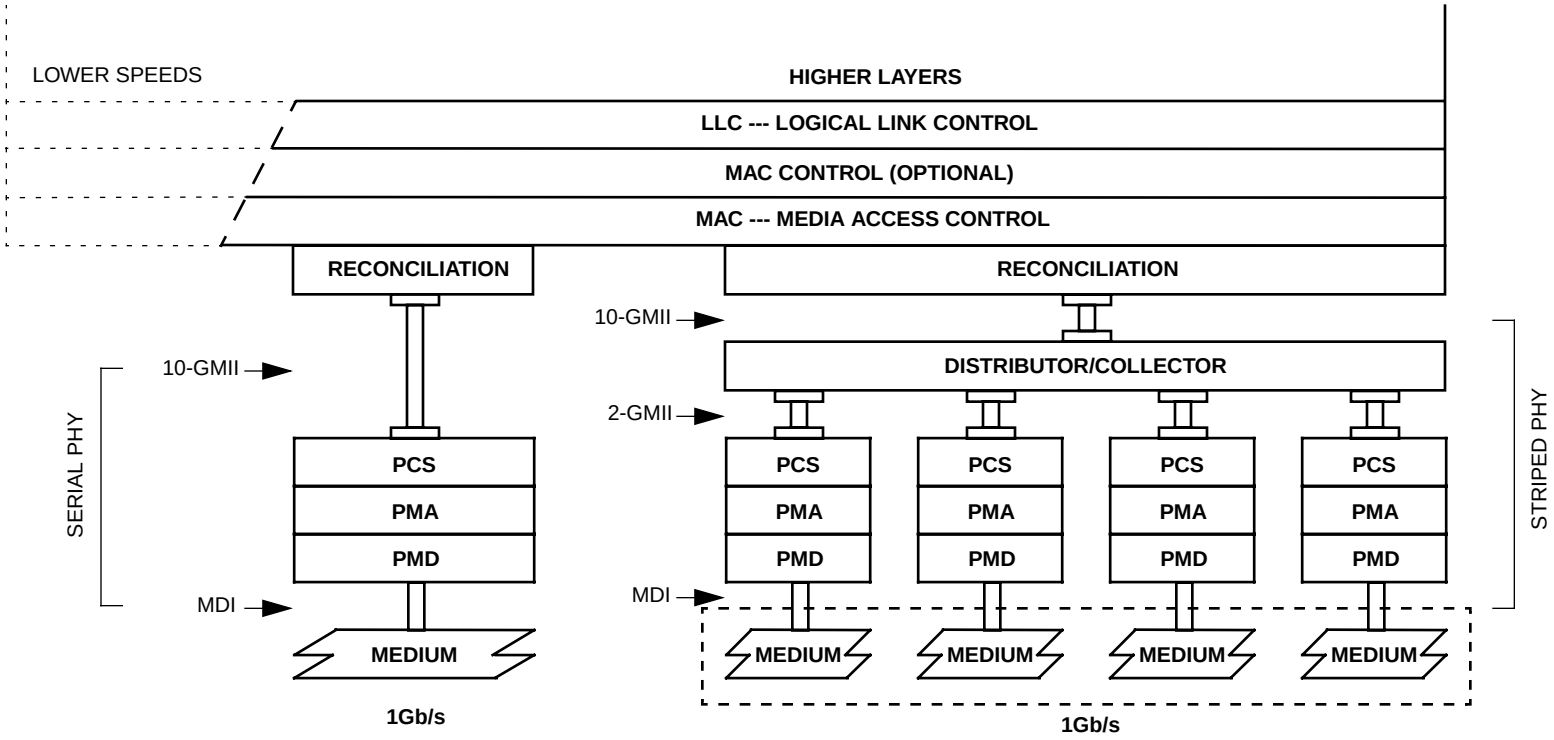
- **Coarse Granularity Striping**
  - **The channels' convergence point is *above* the MAC Layer --- 802.3ad Link Aggregation**
    - **Distribution/collection typically implemented in s/w or in the switching fabric**
  - **High-speed operation achieved only when multiple Layer 2/3/4 "flows" can be aggregated**
  - **For a given "flow", the throughput and the latency are limited by the speed of a single channel**

- **Fine Granularity Striping**
  - **The channels' convergence point is *below* the MAC Layer**
    - **Distribution/collection implemented in h/w as part of the Physical Layer**
  - **Striping of the MAC data stream performed with byte granularity**
  - **From the MAC Client's perspective the performance of the aggregate is identical to that of a single high-speed link**

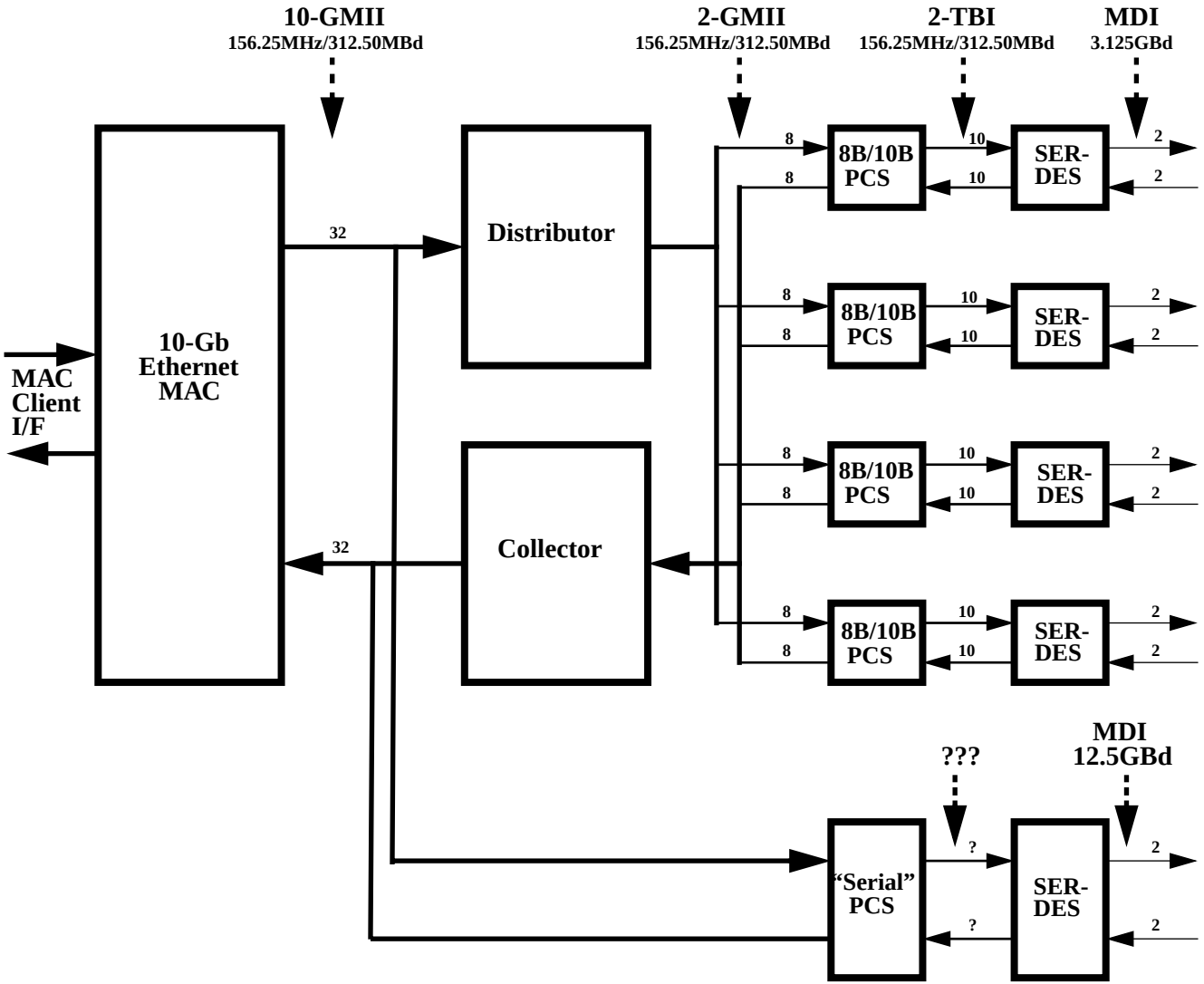
10-Gigabit Ethernet Objectives

- Support the speed of 10Gb/s at the MAC/PLS service interface
- Preserve the 802.3 frame format at the MAC Client service interface
- Preserve the minimum and maximum frame sizes of current 802.3 standard
- Support simple forwarding between 10Gb/s, 1Gb/s, 100Mb/s and 10Mb/s
- Provide support for Full Duplex operation only (no CSMA/CD)
- Meet all 802 functional requirements, with the possible exception of Hamming Distance
- Support star-wired topologies
- Support media selected from ISO/IEC 11801
- Provide a family of Physical Layer specifications which support links of:
  - At least 100m (?) over multi-mode multi-fiber bundles
  - At least 300m (?) over multi-mode single-fiber cable
  - At least 3km (?) over single-mode single-fiber cable
  - At least 50km (?) over single-mode single-fiber cable
- Provide specifications for optional Media Independent Interfaces

Architecture --- Functional Partitioning



Architecture --- Functional Partitioning (continued)



Architecture --- Principles of Operation

- **Assumptions:**
  - **A striped bundle contains a fixed number of channels**
    - **Equal to four**
  - **All the channels in a bundle have the same nominal speed**
    - **Equal to 3.125GBd**
  - **No partial operation is supported**
    - **For a striped bundle to be considered operational, all the channels in the bundle must be operational**
  - **The end points of a bundle are properly wired**
    - **In the same order and contiguous**
  - **The maximum skew between the channels in a bundle is bounded**

Architecture --- Principles of Operation (continued)

- **Distributor:**
  - **Operates in an open loop**
    - **Not required to consider the skew at the receiving side or the receiver state**
  - **Accepts a contiguous byte stream from the MAC (frames and idles) and divides it into four sub-streams (“mini-frames” and “mini-idles”)**
    - **Round-robin arbitration**
    - **Byte granularity**
    - **Starting point is arbitrary, but the arbitration is contiguous afterwards**
  - **The first and/or last bytes of a packet may be sent over any channel with no restrictions**
    - **The Distributor uniquely “marks” the first and last bytes of a packet**
    - **The Distributor enforces packet sequencing during the Inter-Packet Gap**

Architecture --- Principles of Operation (continued)

- **Collector:**
  - **Contains an elasticity buffer per channel**
    - **Compensate for the worst case skew between the channels**
  - **Synchronizes the channels using sequence information received during the Inter-Packet Gap**
  - **Reassembles “mini-frames” into packets and sends them to the MAC**
    - **Reassembly occurs only if all the channels are “in synch” and contain at least a partial “mini-frame”**
    - **Uses the uniquely “marked” first and last bytes of a packet to determine packet boundaries across the channels**
      - **Packet reassembly starting point is determined by the “first” byte of a packet**
      - **After the first byte, reassembly is round-robin with byte granularity**
      - **Packet reassembly end point is determined by the “last” byte of a packet**

Physical Coding and Framing Requirements

- Leverage from 1000BASE-X to the extent possible

- Enhance the 1000BASE-X PCS to deal with following constraints:
  - 1. Preamble Shrinkage
    - The minimum Preamble of a “mini-frame” may be reduced to one byte (7/4)
    - 1000BASE-X encapsulation can extend the IPG by one symbol at the expense of the Preamble
    - 1000BASE-X encapsulation requires at least one symbol of Preamble for SPD
    - Solution:
      - Increase the Preamble field of the 10-Gigabit Ethernet frame by one byte (total of eight)
  - 2. Inter-Packet Gap Shrinkage
    - The minimum IPG between “mini-frames” may be reduced to three bytes (12/4)
    - 1000BASE-X encapsulation allows for an EPD of a maximum of three symbols
    - This can potentially eliminate the Idle signalling between frames
    - Solution:
      - Define a new EPD of no more than two symbols

Physical Coding and Framing Requirements (cont.)

- **3. Packet Sequencing Signalling**
  - **Allows for channel synchronization in the Collector**
  - **Enhances error robustness**
  - **To avoid additional overhead, packet sequencing is enforced during the IPG**
  - **Solution:**
    - **Define multiple flavors of the IDLE ordered set**
- **4. Packet and Frame Delimiters**
  - **"Mini-frame"-to-Packet reassembly in the Collector requires multiple flavors of start and end dilimiters**
  - **Solution:**
    - **Define an SPD code-group that indicates the start of a "mini-frame" on one of the channels *AND* the start of a MAC packet for the entire aggregate**
    - **Define an SFD code-group that indicates the start of a "mini-frame" *ONLY* on the remaining three channels**
    - **Define an EPD code-group that indicates the end of a "mini-frame" on one of the channels *AND* the end of a MAC packet for the entire aggregate**
    - **Define an EFD code-group that indicates the end of a "mini-frame" *ONLY* on the remaining three channels**

Media Independent Interfaces

- Standard media independent interfaces are a good thing to have!
  - Provide interoperable points of attachment between components from multiple vendors
  - Allow for clean architectural partitioning between functional modules
  - Simplify the standard's specification
- Multiple MII compliance interfaces may be needed for the 10-Gigabit Ethernet Standard
  - Allows for various levels of silicon integration in implementations
- All the interfaces are based on the same concepts that were used for Gigabit Ethernet
  - Utilize the full duplex subset of the GMII and the TBI
  - Trade-off between the absolute minimum number of signals used, without substantially increasing the baud rate
  - All the interfaces are defined such that they can be overlaid one on top of the other

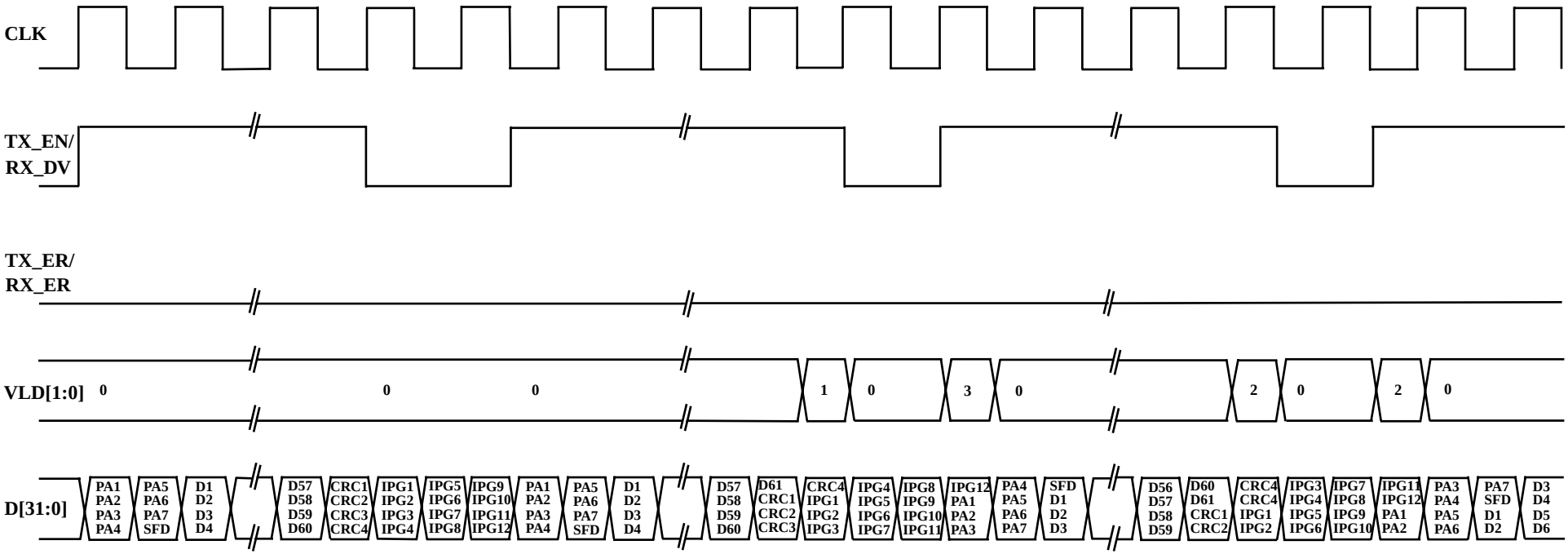
Media Independent Interfaces (continued)

- 10-GMII
  - 156.25MHz clocks
    - 25% increase compared to GMII
  - Both edges of the clocks used for data transfer
  - 32-bit data bus in each direction (TXD/RXD)
    - Increased from 8 on the GMII
  - 2-bit VLD bus in each direction
    - Indicates the number of valid bytes on TXD/RXD
  - Full Duplex operation only
    - GMII COL and CRS signals removed
    - GMII Carrier Extension encodings on TXD/RXD removed

Media Independent Interfaces (continued)

- 2-GMII
  - 156.25MHz clocks
  - Both edges of the clocks used for data transfer
  - 8-bit data bus in each direction (TXD/RXD)
    - Same as GMII
  - 1-bit Packet Delimiter (PD) signal in each direction
    - Indicates the first and the last bytes of a MAC packet
  - Full Duplex operation only
    - GMII COL and CRS signals removed
    - GMII Carrier Extension encodings on TXD/RXD removed
- 2-TBI
  - 156.25MHz clocks
  - Both edges of the clocks used for data transfer
  - 10-bit data bus in each direction (TX/RX)

Media Independent Interfaces (continued)

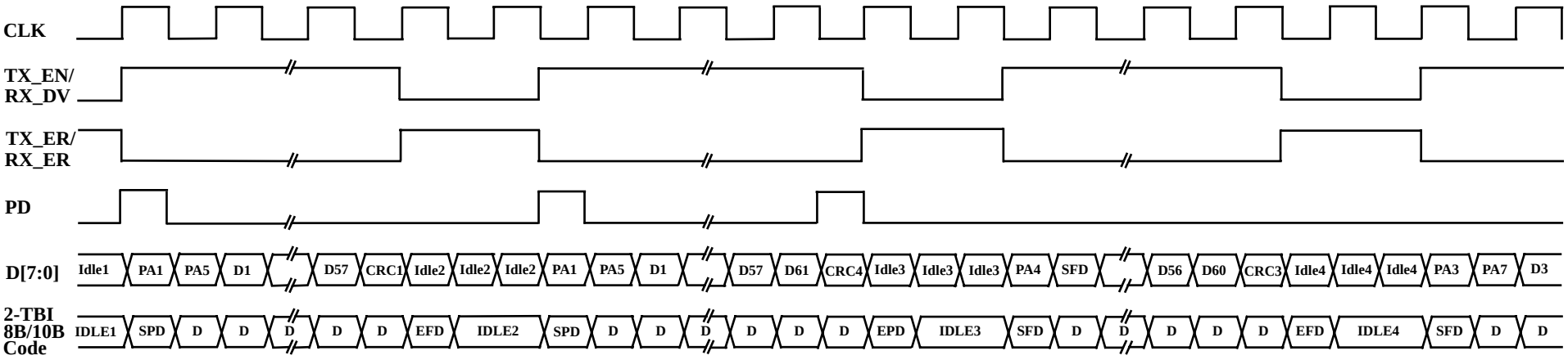


Notes: - When TX_EN/RX_DV is asserted, an encoding of VLD=0 indicates four valid bytes on D[31:0].

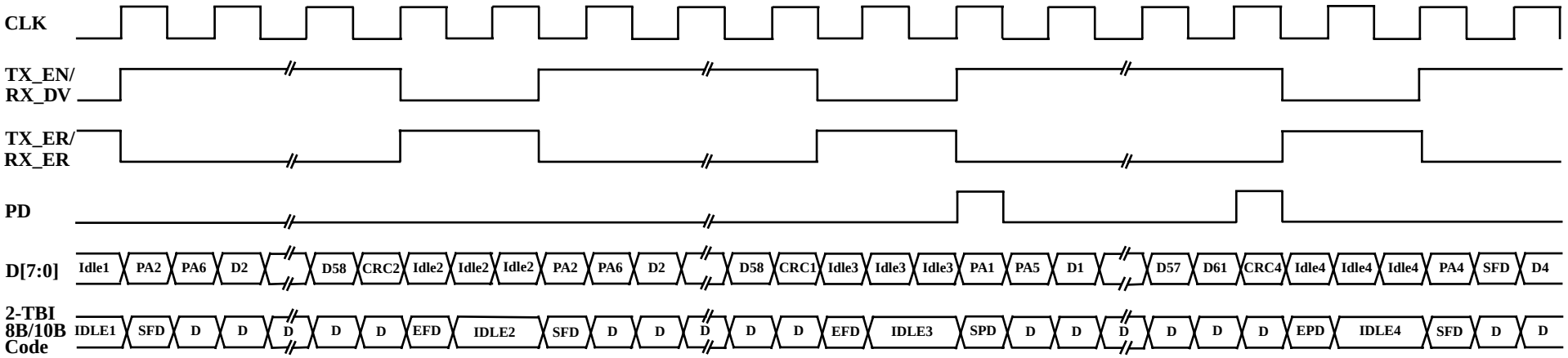
- A non-zero encoding of VLD is only relevant when TX_EN/RX_DV is asserted.

10-GMII

Media Independent Interfaces (continued)

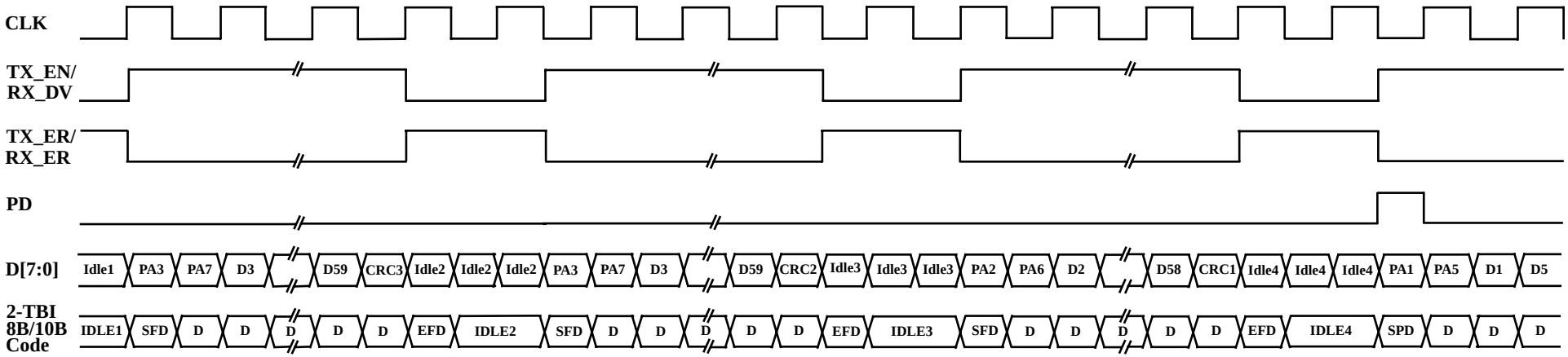


2-GMII/2-TBI --- Channel 1

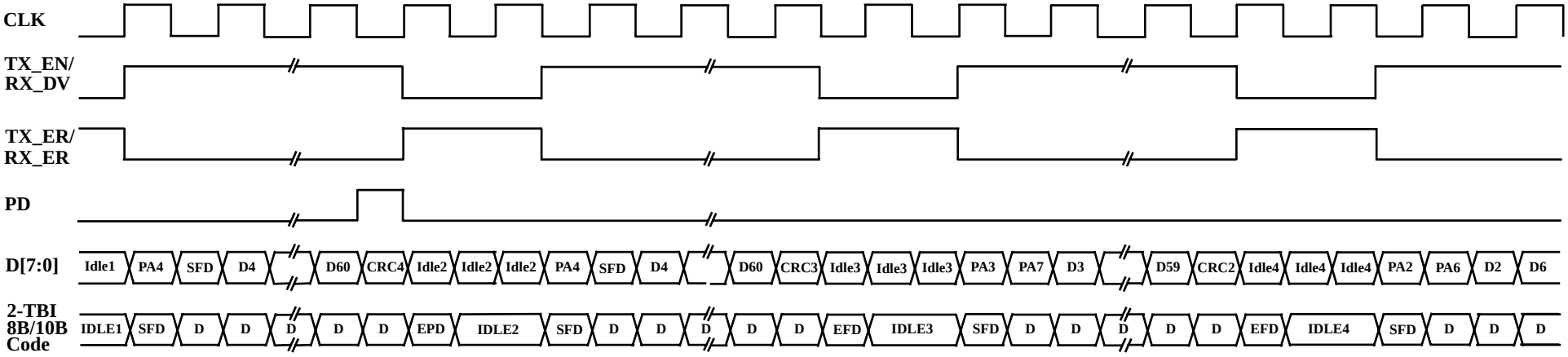


2-GMII/2-TBI --- Channel 2

Media Independent Interfaces (continued)



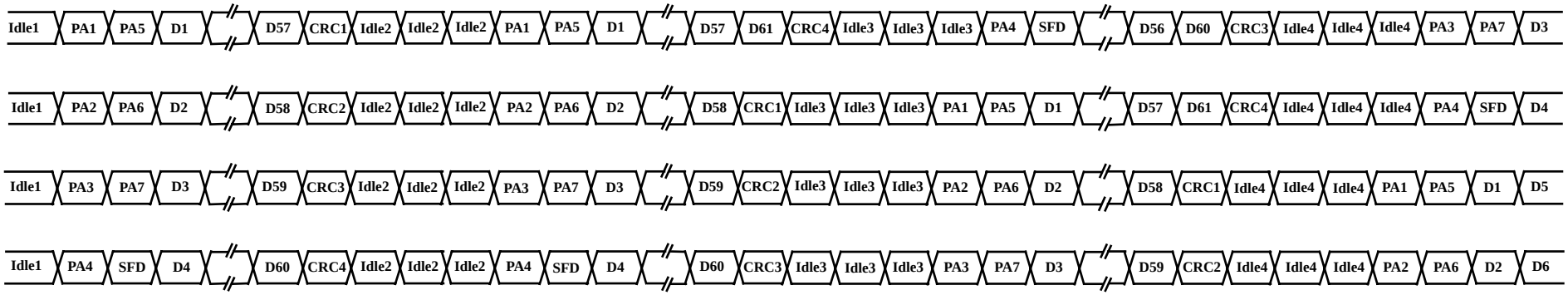
2-GMII/2-TBI --- Channel 3



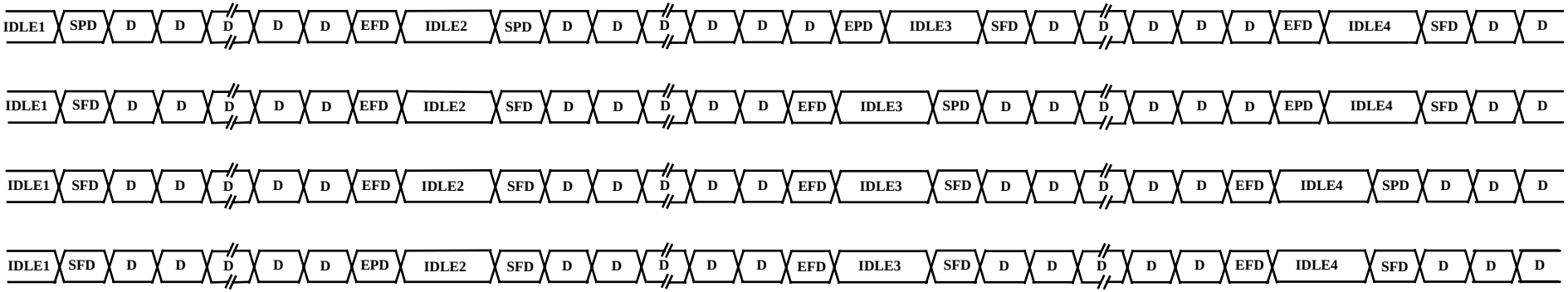
2-GMII/2-TBI --- Channel 4

Media Independent Interfaces (continued)

10-GMII



2-GMII --- Channels 1-4



2-TBI --- Channels 1-4

Error Handling

- **Data Corruption Errors**
  - **Data symbol converted to another data symbol**
    - **Detected and handled by the MAC Layer, analogous to a single-channel transmission scheme**
- **Per Channel Errors**
  - **Code violations, Framing errors, Disparity errors, etc.**
    - **Detected by the PCS in each channel**
    - **Propagated to the Collector over the 2-GMII using the RX_ER handshake**
    - **Propagated to the MAC over the 10-GMII using the RX_ER handshake**
    - **Affect the entire packet in progress**
- **Distribution/Collection Errors**
  - **Length mismatch between “mini-frames” that belong to the same packet greater than 1 byte**
    - **Detected by the Collector**
    - **Propagated to the MAC over the 10-GMII using the RX_ER handshake**

Error Handling (continued)

- **Channel Skew Errors**
  - **Overflow of at least one per channel elasticity fifo in the Collector**
    - **Detected by the Collector**
    - **The entire packet is dropped in the Collector**
- **Channel Synchronization Errors**
  - **At least one channel is out of synch**
    - **The channel received a “mini-frame” with a sequence number that does not match the expected one**
    - **The Collector drops packets until synchronization is reestablished**
  - **Collector cannot reestablish synchronization**
    - **Timer based**
  - **Collector reestablished synchronization incorrectly**
    - **Can happen if multiple “mini-frames” on a single channel “vanished”**
    - **Results in a very high rate of CRC errors, with no other errors apparent**
    - **Flush the pipe (link initialization or 802.3x)**

Summary

- **“One solution that fits all” for the 10-Gigabit Ethernet Standard will not appropriately address the customer needs**
- **The standards effort should be prioritized as follows:**
  - **Phase 1:**
    - **Overall architecture and structure of the standard**
    - **Changes to the MAC Layer**
    - **Media Independent Interfaces**
    - **Channelized transmission scheme**
    - **PCS definition based on the 8B/10B coding scheme**
    - **Physical Layer for ~100m over MMF bundles using SX lasers**
    - **Physical Layer for ~300m over MMF using LX lasers and WDM**
  - **Phase 2:**
    - **Serialized transmission scheme**
    - **New PCS definition based on a high efficiency encoding method**
    - **Physical Layer for ~3km over SMF using LX lasers**
    - **Physical Layer for ~50km over SMF using LX lasers**