PCS/PMA Architecture and OTN Support Proposal

P802.3bs 400 Gb/s Ethernet Task Force

Steve Trowbridge

Alcatel-Lucent

Key Elements of OTN Support

- See "OTN Support: What is it and why is it important?", July 2013
  - A new rate of Ethernet (e.g., 400 Gb/s) fits into the corresponding rate OTN transport signal
  - All Ethernet PHYs of a given rate are mapped the same way and can be interconnected over the OTN (e.g., same PCS for all 100 Gb/s PHYs gives a single canonical format ("characteristic information" in ITU-T terminology) that can be mapped
  - Optical modules for Ethernet can be reused for OTN IrDI/client interfaces at the corresponding rate

A new rate of Ethernet (e.g., 400 Gb/s) fits into the corresponding rate OTN transport signal

- Assumption – the OTN mapper/demapper will terminate and regenerate any Ethernet FEC code, correcting errors at the OTN ingress since the FEC is chosen to correct single-link errors but not double-link errors
- Assumption – the OTN mapper/demapper may trans-decode/trans-encode back to 64B/66B to avoid MTTFPA reduction for OTN transported signal
- Based on these assumptions, the encoded data rate of the OTN-mapped 400 Gb/s Ethernet would be no more than 400 Gb/s x 66 / 64 = 412.5 Gb/s ±100ppm. Since the 400 Gb/s OTN container would presumably be designed to also transport four “lower order” ODU4s, there should be no concern that it is large enough to carry 400 Gb/s Ethernet based on the assumption that the canonical form is near this rate.
- Any Ethernet bits in excess of this rate are likely to be part of a FEC that is not carried over OTN

Possible Canonical Forms for OTN mapping

- Serialized and deskewed 64B/66B including lane alignment markers (same as 100GBASE-R mappings)
- Deskewed and Serialized 64B/66B without lane alignment markers – remove markers from PHY at ingress after deskew and insert markers at egress when striping to lanes of Ethernet PHY. This could apply, for example, if 400 Gb/s PHYs with different lane widths are marked differently

Independence or Dependence between Ethernet and OTN frame formats using Ethernet PMA

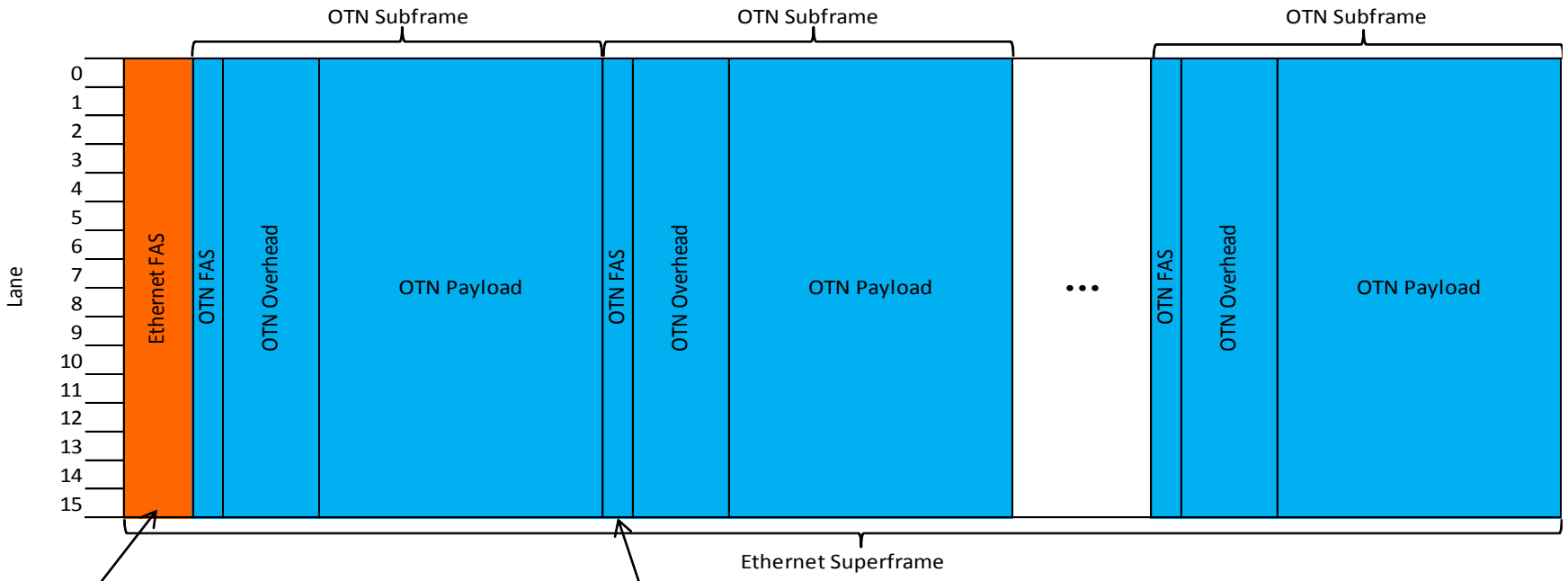
- 802.3ba lane architecture uses bit-multiplexing/gearboxing in the PMA (the only recombination of physical/logical lanes in the modules) which is completely agnostic to the bit values on lanes. As long as OTN and Ethernet were striped into the same number of lanes and each application could frame and deskew its own signal, everything was fine
- Some 400 Gb/s PCS/PMA proposals would not be agnostic to the bit values on lanes (e.g., if 10-bit interleaving is used or 66B symbol interleaving, the Ethernet lane marking is difficult to apply to the OTN frame

400 Gb/s PCS/PMA Ideas so far
Idea #1

- From study group “logic” ad hoc, use 4 instances of P802.3bj clause 91 FEC to produce a 16-lane format. Logical lanes are combined to >25G physical lanes by interleaving on 10-bit RS symbol boundaries. Refer to gustlin_01_1013_logic.pdf
- Advantages: High degree of reuse for P802.3bj, P802.3bm designs, relatively simple breakout to 4x100G with FEC interfaces
- Disadvantages: Extra module complexity since the modules have to find alignment markers to do 10-bit alignment and interleaving; challenging to find a way to put this marking into a traditional OTN frame without redesigning the frame

Implications of Idea #1 for OTN

Significant re-design of OTN frame required to fit Ethernet Lane markers into OTN frame



Ethernet Alignment Markers

Divide into sub-frames to not wait 210µs to do OTN frame alignment

400 Gb/s PCS/PMA Ideas so far
Idea #2

- Use the same basic format as Idea #1, but combining logical lanes into >25G physical lanes only combine lanes of different FEC groups and can therefore use bit-multiplexing. Refer to: wang_01_0414_logic.pdf
- Advantages: allows simpler module design; no constraints on OTN framing since bit multiplexing is used.
- Disadvantages: doesn't extend as easily to 4x100G breakout applications since the lane order is different and you can't combine lanes of the same FEC group. For example, Idea #1 could allow 400G technology to provide 2x50G 100G implementations, whereas this proposal would not

400 Gb/s PCS/PMA Ideas so far Idea #3

- Consider the PCS to be a logically serial stream of 64B/66B blocks without lane alignment markers
- Marking, striping, transcoding, FEC encoding are the attributes of a physical instantiated interface (e.g., the CDAUI-n or the optical portion of the link) and are not inherently considered to be part of the information.
- FEC could be chosen sub-link by sub-link according to the error model of the connection, or could be chosen to cover a sequence of sub-links
- Advantages: completely general and extensible to future interface widths and characteristics.
- Disadvantages: slightly awkward for OTN mapping; multiple re-striping and re-FEC encoding adds complexity, power, latency across the extent of the connection; loss of end-to-end BIP monitoring capability

Idea #3 Amplification

Only verbal discussion, no formal description in a contribution

- PCS is a logical serial stream of 66B blocks. Only physical instantiations are striped over physical or logical lanes
- Maintain the principle, as in 802.3ba, that idle insertion/deletion is not done below the PCS.
- Since any physical instantiation will need to be striped with lane markers, do idle insert/delete at the PCS only so the logical stream will be at the
  nominal MAC rate x 66/64 x (1-1/16384)
  so that any physical instantiation has room to insert lane markers as needed without idle insert/delete

Idea #3 Amplification continued

- Example physical instantiation could be exactly the format of Idea #1, produced by transcoding 64B/66B to 256B/257B, striping first into 100G groups, striping within each 100G group into 4 logical lanes on 10-bit symbol boundaries, inserting alignment markers on each lane, and applying an RS(528,514) code based on 10-bit symbols with alignment markers appearing in the first of each of 4096 Reed Solomon code blocks

Idea #3 Implications for OTN

- Likely only possible if the same FEC code can be used for OTN applications as for Ethernet applications at about 6% higher bit-rate
- Would need to make OTN look like 66B blocks. Easiest way to do this and not lose any information in transcoding is to insert a “01” sync header after every 64 bits (all data)
- Since this is just part of the logical frame format, this doesn’t waste as many bits as it appears. 8 sync header bits are added to every 256 data bits in the “logical” frame format, but 7 of those bits are immediately recovered in 256B/257B transcoding and reused for the FEC code. So 0.39% net is added to the OTN frame to make it look like 66B blocks, then 2.724% overhead RS FEC added

Further Observations

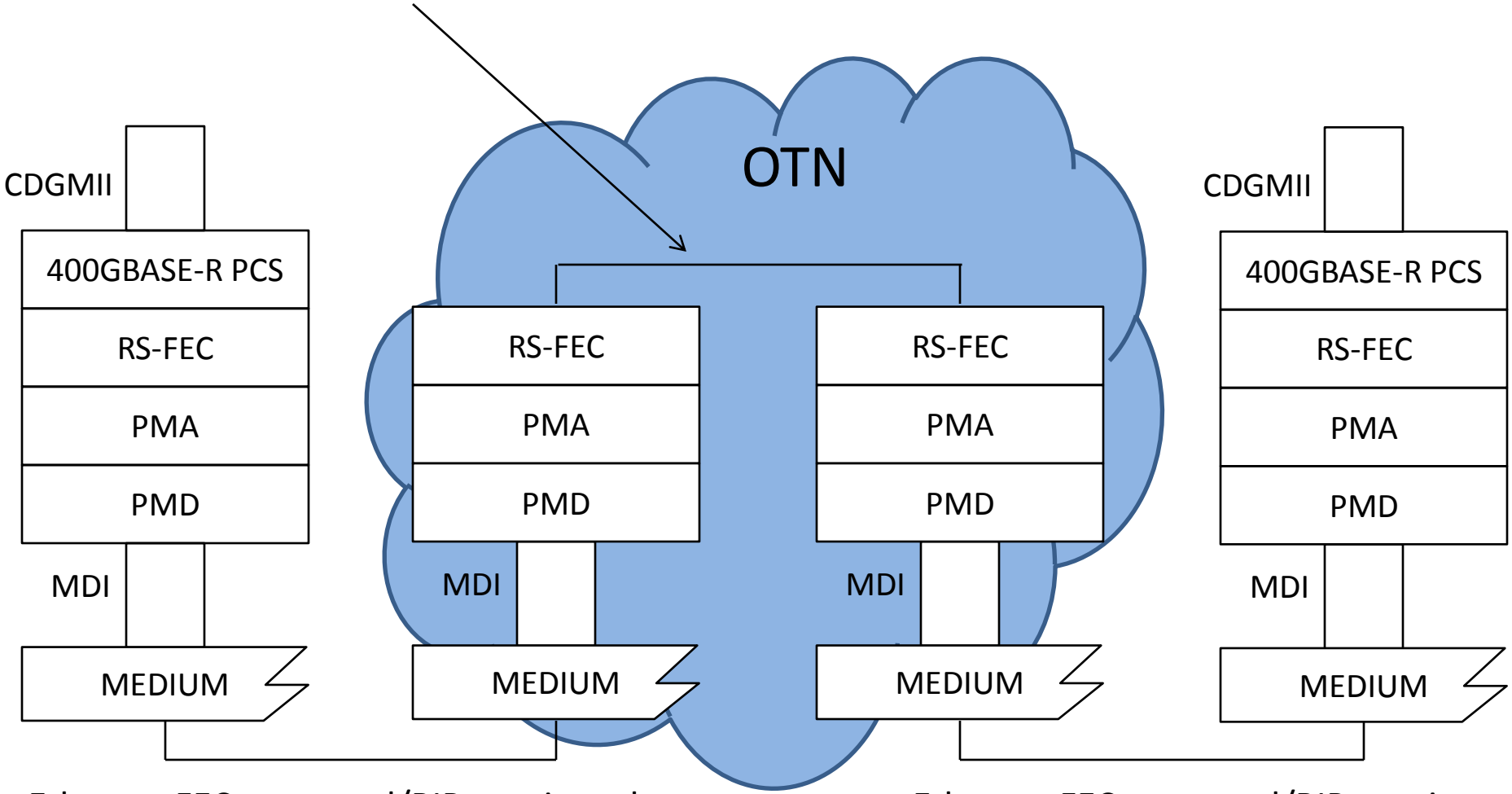
- The 40GBASE-R and 100GBASE-R PCS combines the following functions in the Tx direction and their reverse in the Rx direction:
  - Idle Insertion/Removal to provide room for alignment markers
  - 64B/66B encoding of information after Idle Insertion/Removal
  - Striping the information over logical (PCS) lanes, inserting lane alignment markers
  - Monitoring the error performance of each logical lane with BIP
- No “clean” treatment of BIP in OTN mapping:
  - Should it be used for path (end-to-end) or section (individual link) monitoring?
  - Needs to be converted to an error mask instead of a BIP with transcoding in 40GBASE-R mapper and BIP bits flipped at the egress to create aggregate error view
- While there has been discussion that various electrical and optical sections of a link might have different error characteristics, do we really imagine multiple Ethernet FECs on a single Ethernet link?
  - Complexity (overall and in the module), power, latency
  - The optical link is likely to be the dominant source of errors in any case, so even if different optics require different FEC, can we assume that the FEC for the optical link can be chosen to address errors in the electrical part of the link?

Proposed PCS and RS-FEC Layering

- Propose that the 400GBASE-R PCS produce a logically serial stream of 64B/66B blocks, with idle insertion/deletion to allocate room for lane markers (since every physical instantiation will need to be striped). The PCS performs idle insert/delete, encode and scramble in the Tx direction and descramble, decode, and idle insert/delete in the Rx direction. The bit rate of the logical PCS is 400G×66/64×(1-1/16384)=412.474823 Gb/s ±100ppm
- No idle insertion/deletion occurs below the PCS.
- RS-FEC layer responsible for Round-robin striping of 66B blocks to FEC groups, insertion of alignment markers, 256B/257B transcoding, and FEC encode occur in the Tx direction while FEC decode, alignment lock, lane identification and deskew, 256B/257B trans-decoding, and reassembly into the logically serial stream of 64B/66B blocks occur in the Rx direction.
- RS-FEC encoded format similar to what is described in gustlin_400_02a_1113.pdf assuming the coding gain is sufficient for selected P802.3bs PMDs
- Since the boundary between the 400GBASE-R PCS and the 400GBASE-R RS-FEC is logical and not physically instantiated, no FEC, lane striping, or BIP is required at this layer.

Illustration of OTN Mapping

Serial 64B/66B, OTN FEC protected/BIP monitored



Ethernet FEC protected/BIP monitored

Ethernet FEC protected/BIP monitored

The Ethernet ingress/egress links don't need to be striped or FEC encoded the same if they are different PMDs

Possible 400G OTN Bit Stream (without FEC)
This needs to transit the module with FEC

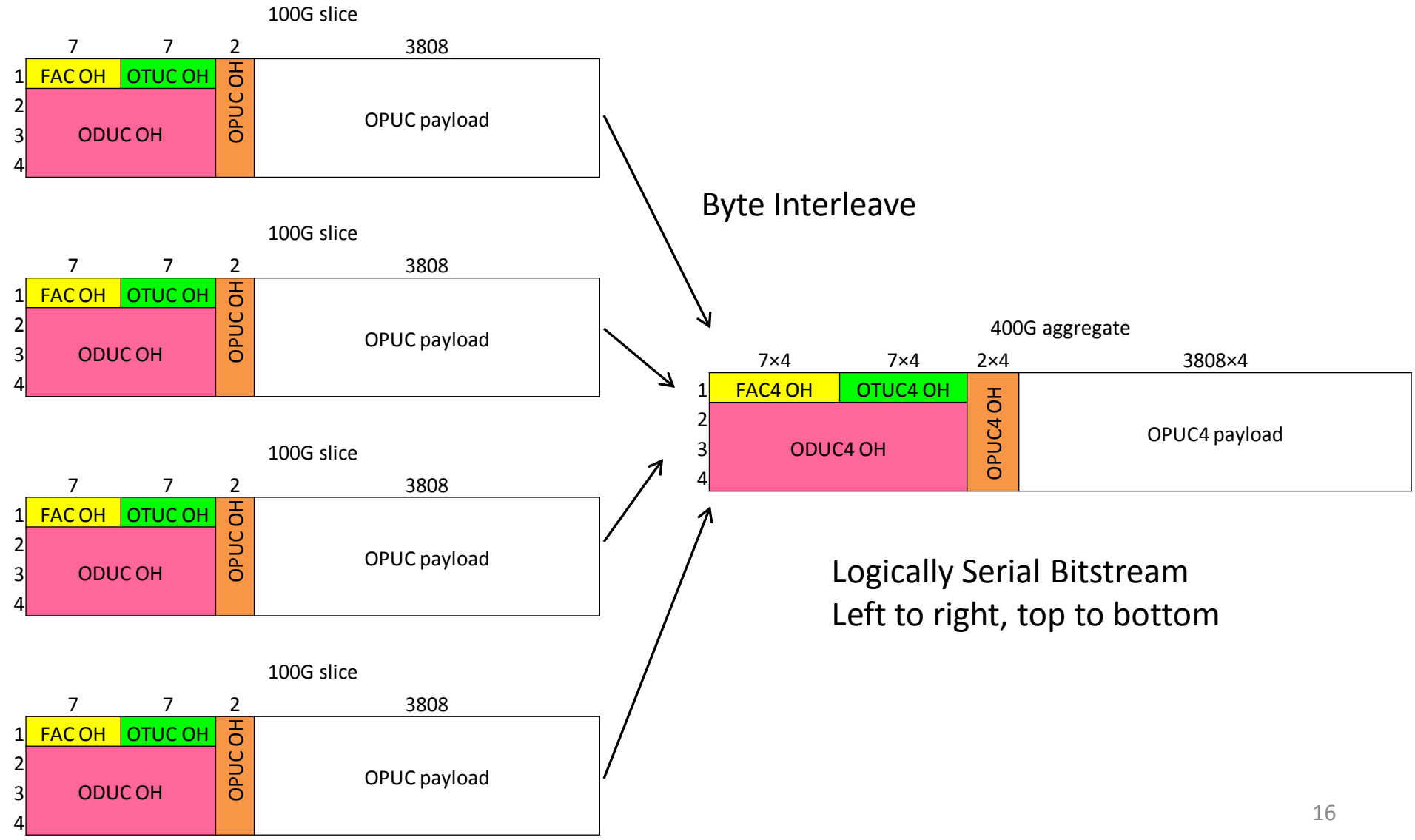
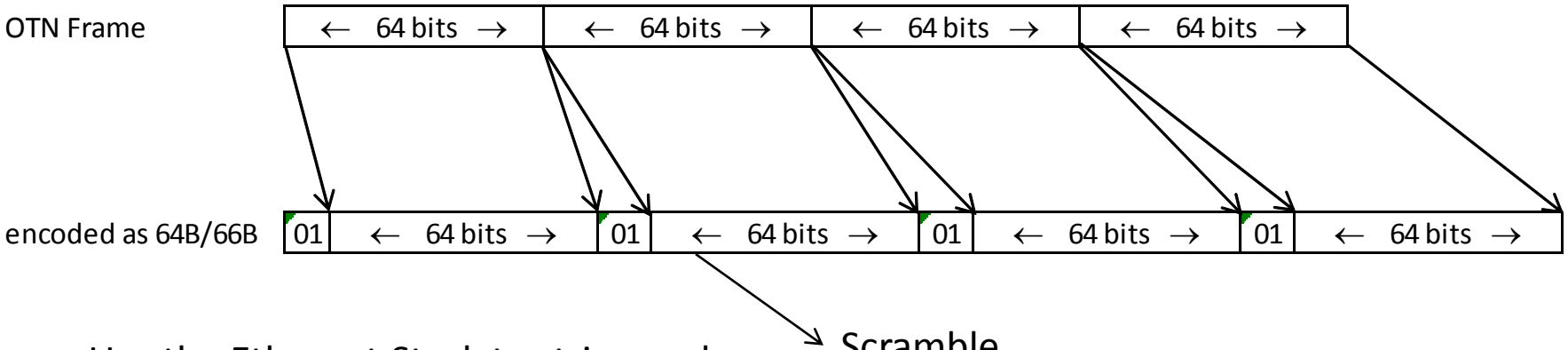
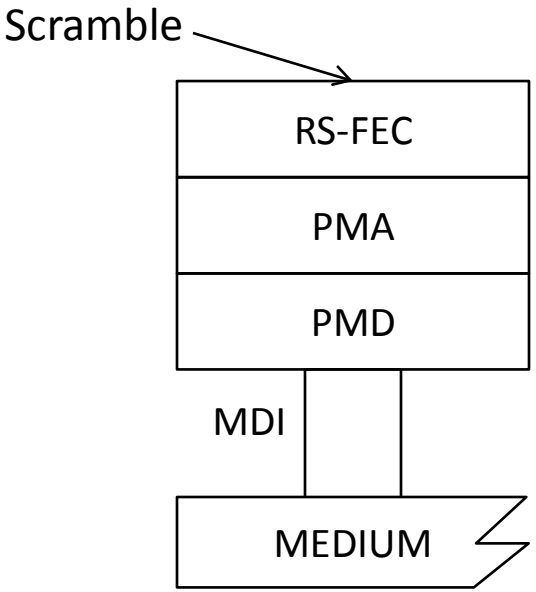


Illustration of turning OTN frame into 64B/66B blocks



Use the Ethernet Stack to stripe and FEC encode the OTN frame when carrying over an Ethernet Module for an OTN IrDI or client interface

Could be OTN frame aligned as an OTUC4 frame without FEC is exactly 7648×64 bits, but not essential with scrambling



OTN Bit-rates using this scheme

	Working Assumption Bit-Rate
OTUC4 bit-rate without FEC	422.904 Gb/s
64B/66B encoded	436.120 Gb/s
256B/257B transcoded	424.556 Gb/s
Insert Lane Markers	424.582 Gb/s
Add RS(528,514) FEC	436.146 Gb/s
Logical Lane Rate (well within CEI-28G)	27.259 Gb/s
Ethernet Nominal Bit-rate	412.5 Gb/s
400G OTN Increase in bit-rate	5.73 %

100G OTN Increase in bit-rate	8.42 %

Smaller increase for 400G than for 100G, mainly due to RS(528,514) FEC rather than RS(255,239) FEC

The module reuse aspect of OTN Support is satisfied if the following are true:

- There is an Ethernet sublayer reference point such as the PCS that is logically a serial stream of 64B/66B blocks
- No idle insertion/deletion occurs below the PCS (the serial stream of 64B/66B blocks), and hence the rest of the stack can deal with a constant-bit-rate (CBR) bitstream that is effectively an infinite-length packet.
- Note that any logical to physical lane interleaving that works for Ethernet also works for OTN since they are encoded the same way
- The link parameters and FEC coding gain have sufficient margin to meet the error performance target when running at approximately 5.73% higher bit-rate than necessary for 400G Ethernet. More likely to be true if all P802.3bs interfaces have FEC.

Details of Proposal Assumptions

- Define a PHY link to be a point-to-point Ethernet connection over a physical cable. The 400 Gb/s PCS can be carried over either a single PHY link or a concatenation of PHY links (e.g., the OTN ingress link and egress link plus the OTN).
- The fundamental purpose of BIP is to perform fault isolation rather than service assurance and to count errors in a single PHY link. BIP errors in an OTN ingress PHY link and an OTN egress PHY link are counted separately even though they are carrying the same Ethernet PCS client – Note, however that if the purpose of BIP is for service assurance rather than fault isolation, this could be done by splitting the proposed LLS sub-layer and having a transcoded and non-transcoded version of the alignment markers, and carrying the serialized alignment markers along with 64B/66B data as in the current 100GbE mapping into ODU4. If used as a path monitor, the BIP should probably be calculated on the blocks and alignment markers pre-transcoding.
- Note that this could be combined with either proposal #1 (simple reuse of P802.3bj FEC), or proposal #2 (swap around lanes of 100G FEC groups to allow bit-muxing in the module. There may be other good reasons to allow bit-muxing even if not needed for OTN support

Details of Proposal Nomenclature

- Propose to use interface and sublayer names to describe what they do, not to “force fit” to an historical set of names. Specifically:
  - CDGMII – 400 Gb/s Media Independent Interface
  - 400GBASE-R PCS – Physical (64B/66B) coding sublayer – creates logically serial 64B/66B stream. Scrambled idle test pattern generation/detection is the responsibility of this sublayer. Idle insertion/deletion is performed in this sublayer to produce a PCS rate of 400 Gb/s x 66/64 x (1-1/16384) ±100ppm. All 400 Gb/s Ethernet PHYs will use this same PCS
  - LLS – Logical Lane Striping sub-layer – Adapts from logically serial 64B/66B to a logical lane striped format by round-robin distribution of 64B/66B blocks and insertion of lane alignment markers. This sublayer is responsible for 256B/257B transcoding as the size of the alignment markers may be chosen to match transcoded stream. There MAY be more than one kind of LLS if the striping for different PHYs is fundamentally different. Logical lane BIP is calculated in the Tx direction and errors counted in the Rx direction. FEC parity, calculated below this sublayer, is not included in logical lane BIP

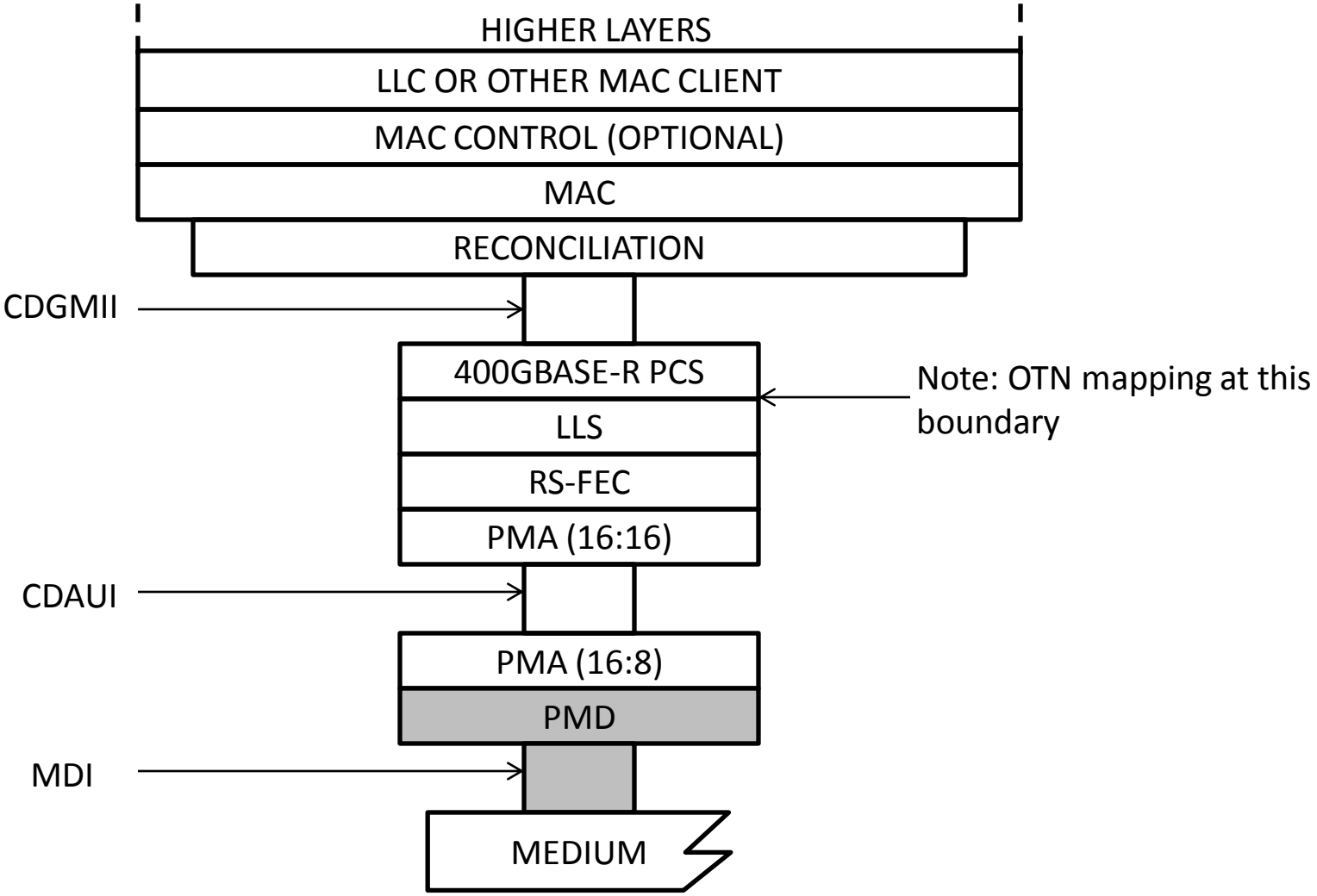
Details of Proposal Nomenclature - Continued

- RS-FEC – in the Tx direction, computes and inserts FEC parity. In the Rx direction, performs error correction based on the FEC parity. There may be more than one kind of FEC sublayer if different FECs are required in different contexts, and there may be decode of one kind of FEC and re-encode of another, e.g., below a CDAUI, if a particular PMD requires a stronger FEC
- PMA – FEC lane striping sub-layer. This is the equivalent of the 100GBASE-R PMA in that it is capable of multiplexing, demultiplexing, or grearboxing of FEC lanes to change physical lane widths and rates. The FLS is responsible for CDR and meeting of timing and electrical specifications if the interface below or above is physically instantiated as a CDAUI. Optional test pattern generation and detection (other than scrambled idle, which would be in the PCS) is the responsibility of the FLS. The FLS will interleave physical lanes to logical lanes as required by the FEC

Details of Proposal Nomenclature - Continued

- CDAUI – 400 Gb/s Attachment Unit Interface (chip-to-chip or chip-to-module, a physical instantiation of the FLS service interface)
- PMD – one of the specified 400 Gb/s PMD types
- MDI
- Medium

Details of Proposal
Example of proposed layering



Details of Proposal
OTN Mapping

- Proposed that the 400GBASE-R PCS (logically serial 64B/66B at rate of 400 Gb/s x 66/64 x (1-1/16384) ±100ppm) is mapped regardless of 400 Gb/s Ethernet PHY type. The OTN ingress PHY link and OTN egress PHY link can be different 400GBASE-R PHY types
- The OTN mapper uses the Ethernet stack (LLS and below) to convert from the physical ingress PHY link to the mapped bit-stream
- The OTN demapper uses the Ethernet stack (LLS and below) to convert from the demapped bit-stream to the physical egress PHY link (which may be a different PHY type than the ingress PHY link)

Details of Proposal

Enabling 400 Gb/s Ethernet Module reuse for 400 Gb/s OTN IrDI/client interfaces

- No Idle Insertion/Deletion below the PCS
- The OTN frame is adapted to a CDAUI-like format by encoding as 64B/66B data blocks and using the Ethernet LLS, RS-FEC, and FLS sub-layers at 5.73% higher bit-rate
- All 400 Gb/s Ethernet PHYs are assumed to use FEC
- The FEC code is selected to have sufficient margin to meet the error performance target at 5.73% above the Ethernet bit-rate

THANKS!