Channel Bonding in Reconciliation Sublayer (Upstream Direction)

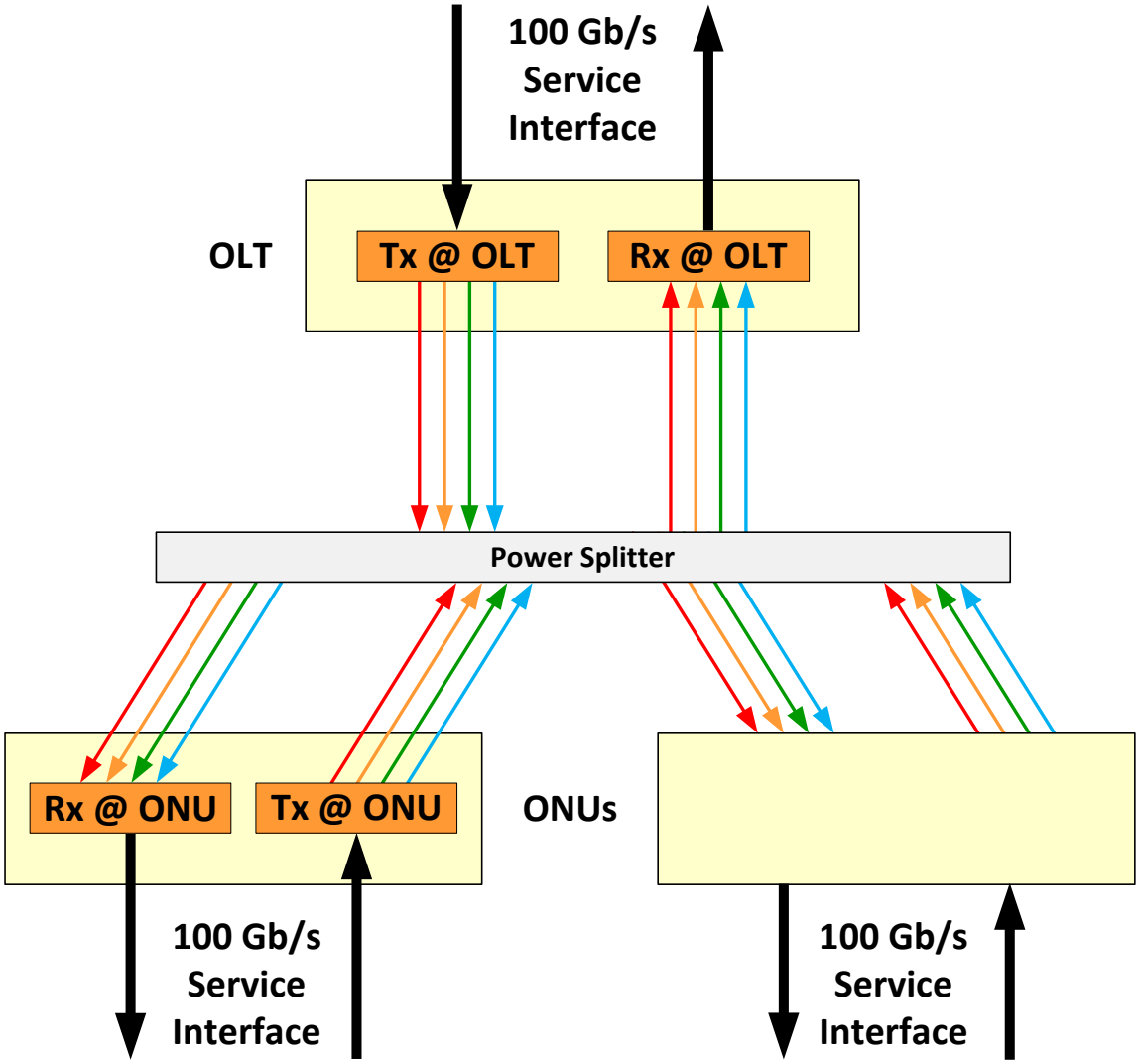
Glen Kramer, glen.kramer@broadcom.com

Scope of this proposal

This presentation proposes RS state diagrams for upstream channel bonding:

1. **Tx @ ONU**
2. **Rx @ OLT**

- Downstream channel bonding is easier and will be developed after the upstream is decided



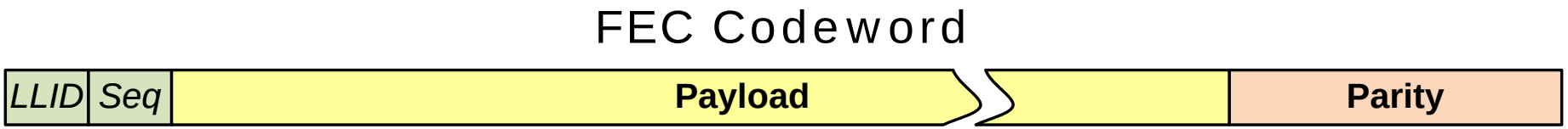
Part 1: Transmission Processes at the ONU

Basic Ideas

1) Bursts always contain integer number of FEC codewords. It makes more sense to make a codeword an atomic transmission unit.
   - Grant length is expressed in units of codewords
   - Eliminates 10G-EPON scheduling overhead due to unit conversion (Bytes → TQ → Bytes → FEC → TQ)

2) If the order of FEC codewords can be preserved, then packets can be fragmented and reassembled without any per-packet fragmentation headers

3) To restore FEC codeword order, add a header to each FEC codeword <LLID, SeqNum>



Basic Ideas

4) When RS stops asking MAC for more data bits, the rest of the frame remains in the MAC (an approach used in 802.3br)

5) MPCP can inform RS when to start transmission and for how long to transmit on each lane
   - CB_CTRL.request() primitive similar to MM_CTRL.request() using in 802.3br

Multi-Point Reconciliation Sublayer (MPRS)

- ***MPRS*** interfaces to ***M*** MAC instances above and ***L*** xMII instances below

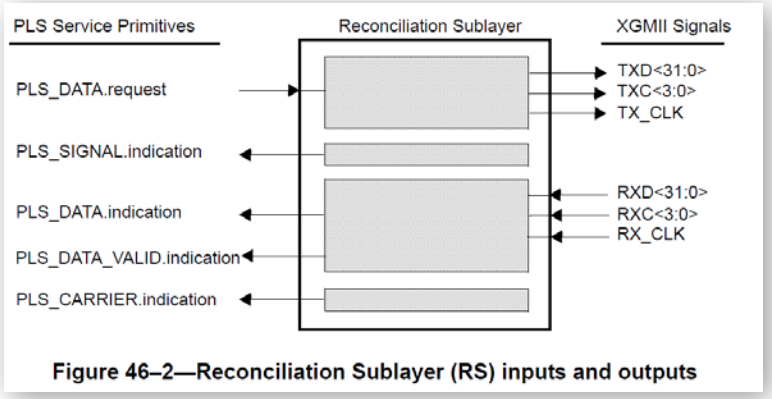
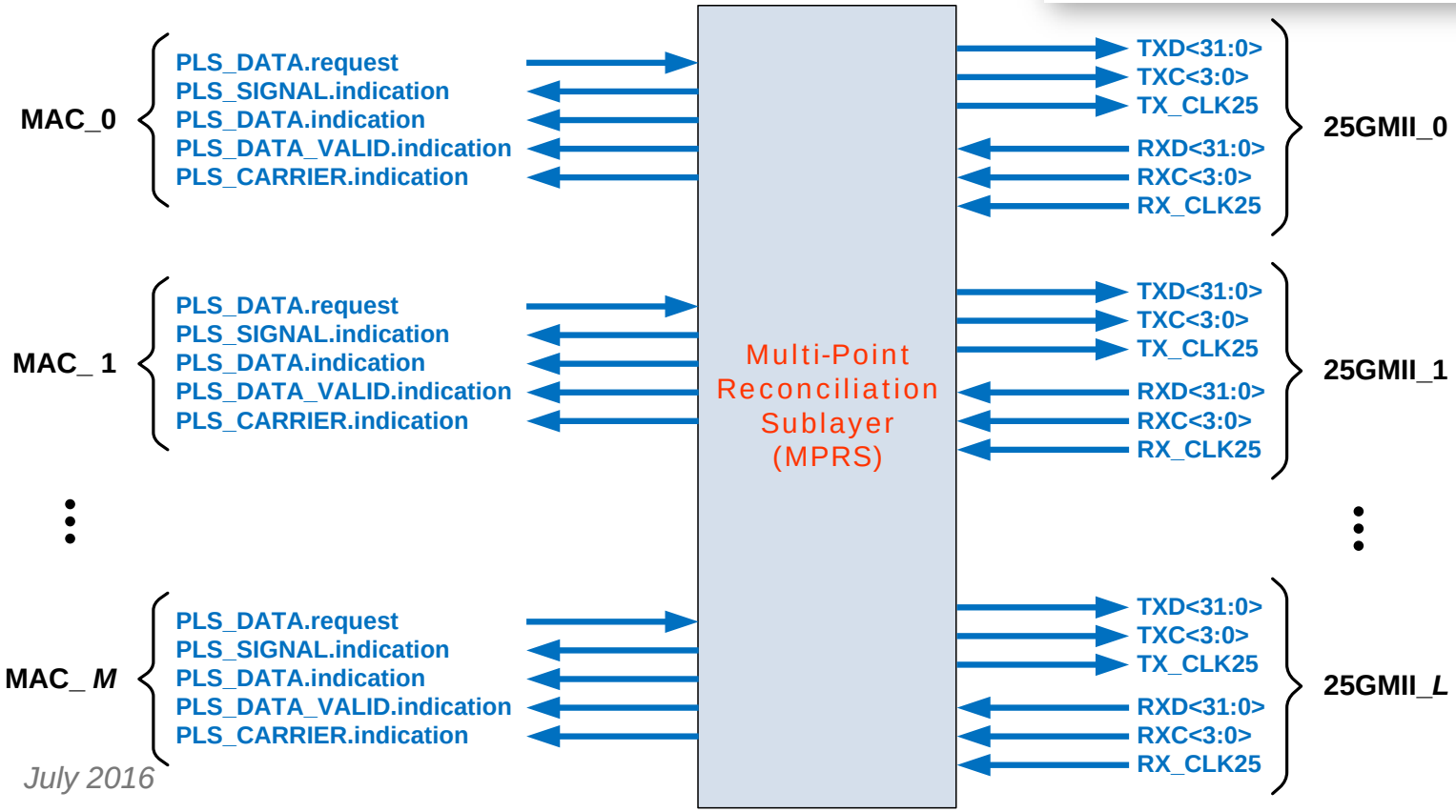


Figure 46–2—Reconciliation Sublayer (RS) inputs and outputs



Precedent for multi-interface RS

- In 802.3, we already have RS with multiple MACs above and multiple xMIIs below

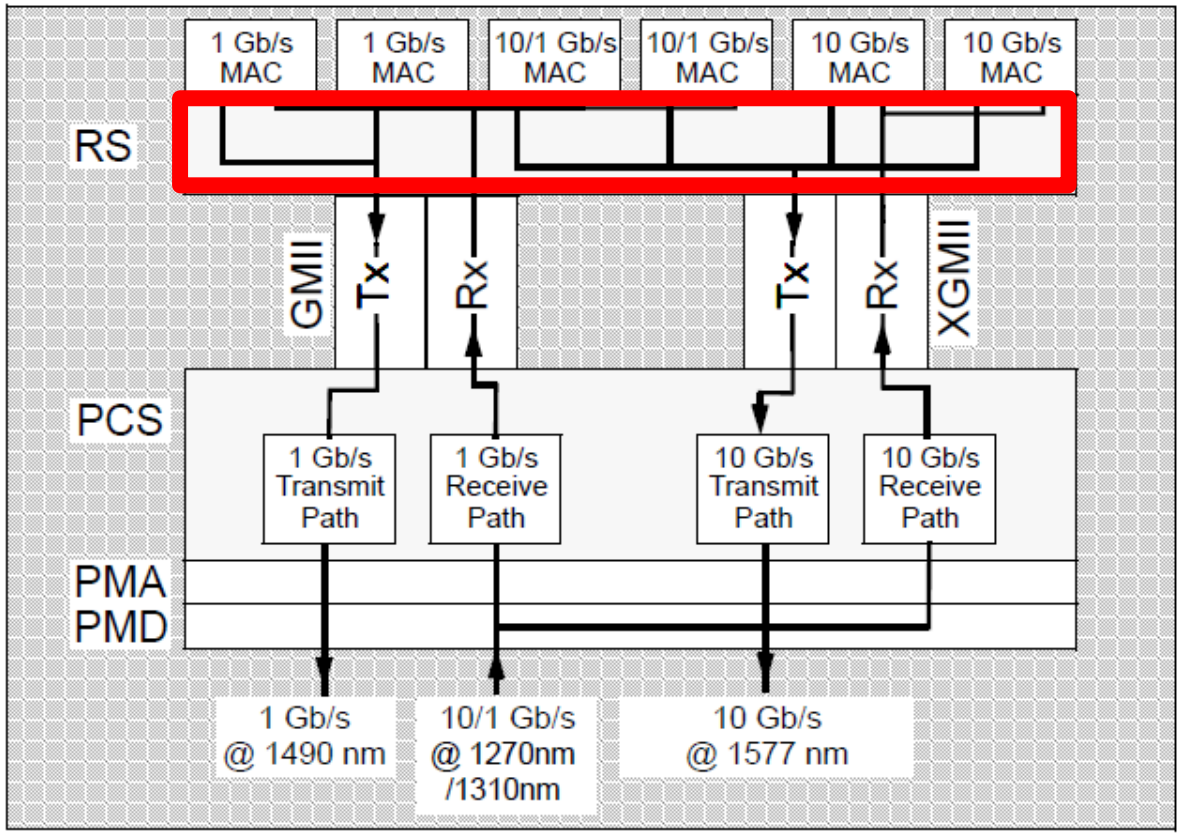
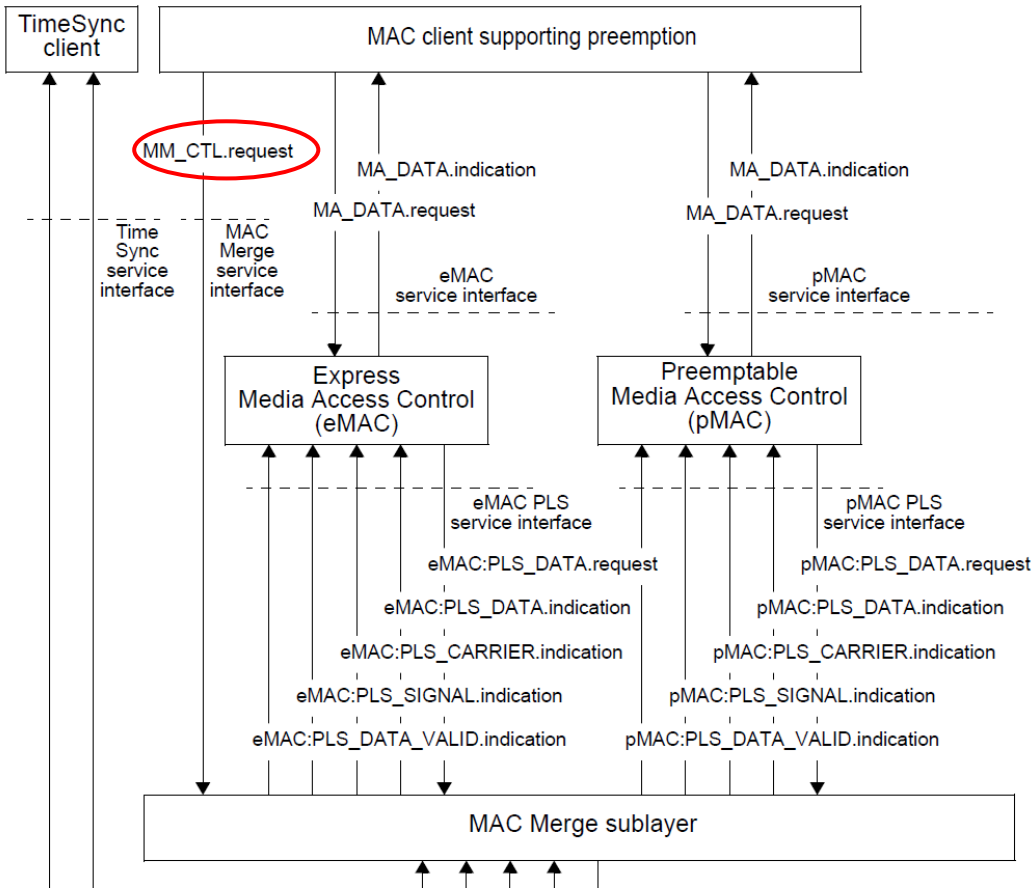


Figure 76–4—PCS and Reconciliation Sublayer for dual rate mode at OLT

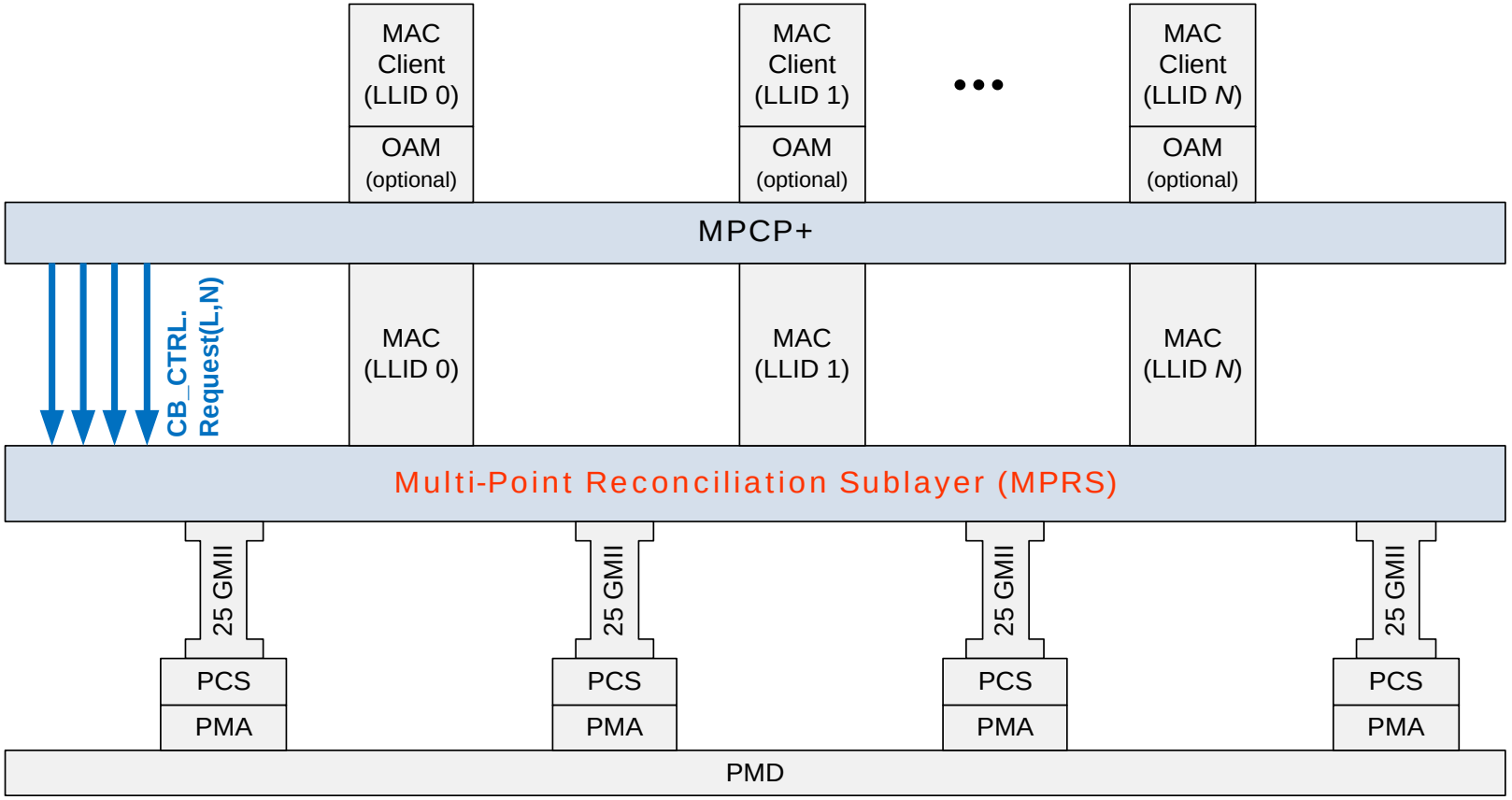
Channel Bonding Control Primitive

- **CB_CTRL.request()** primitive is similar to **MM_CTRL.request()** in 802.3br
- CB_CTRL.request() is generated in MPCP and processed in MPRS
- A separate CB_CTRL interface is instantiated for each lane



- Definition of **CB_CTRL.request(M, N)**
  - Transmission starts immediately upon reception of the primitive
  - M = MAC index (Logical link index) to source the data
  - N = number of FEC codewords to transmit on a given lane

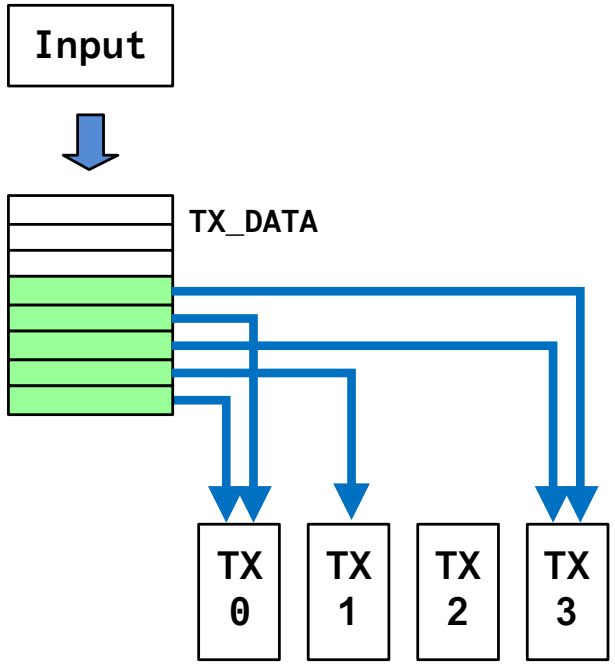
100G ONU Layering Diagram



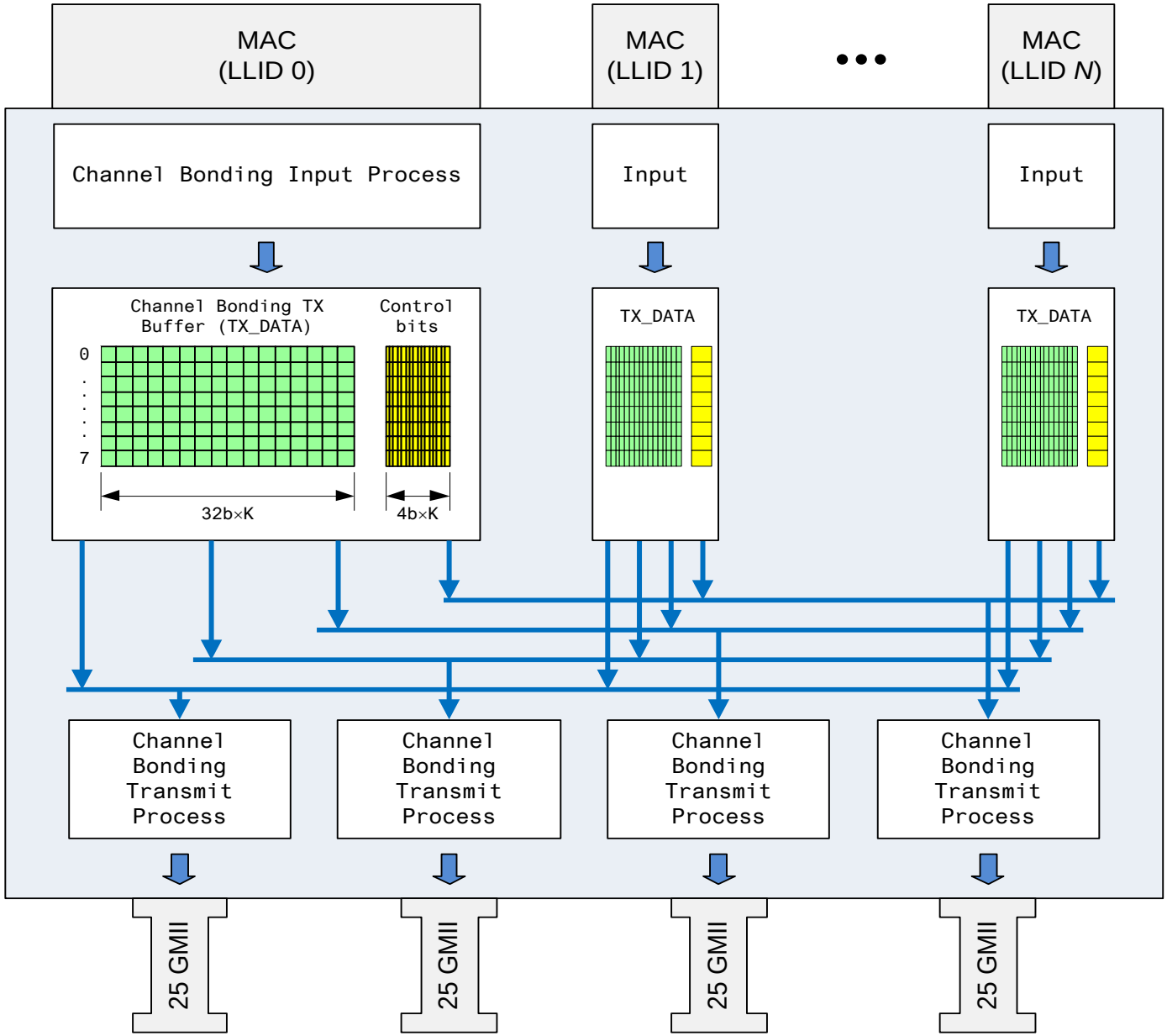
- Each MAC is a 100 Gb/s MAC
- If not all four lanes are active at a given time, the MPRS will pause MAC (i.e., not accept more bits) to equalize MAC and PHY data rates

Main blocks in ONU MPRS Tx Path

- **Channel Bonding Input Process**
  - Accepts data from MAC into *Channel Bonding TX Buffer (TX_DATA)*
  - Prepends FEC codeword header (LLID and sequence number) to each FEC payload
  - One instance of state diagram per MAC (LLID)
- **Channel Bonding TX Buffer (TX_DATA)**
  - A circular buffer sufficient to hold 8 FEC codeword payloads
  - Allows serial write and 4 parallel read streams
  - One instance per MAC (LLID)
- **Channel Bonding Transmit Process**
  - Outputs a complete FEC payload to one 25GMII.
  - One instance of state diagram per lane
- **Channel Bonding Control Process** (not shown on the block diagram)
  - Processes CB_CTRL.request() primitives
  - One instance of state diagram per lane



ONU MPRS TX Block Diagram

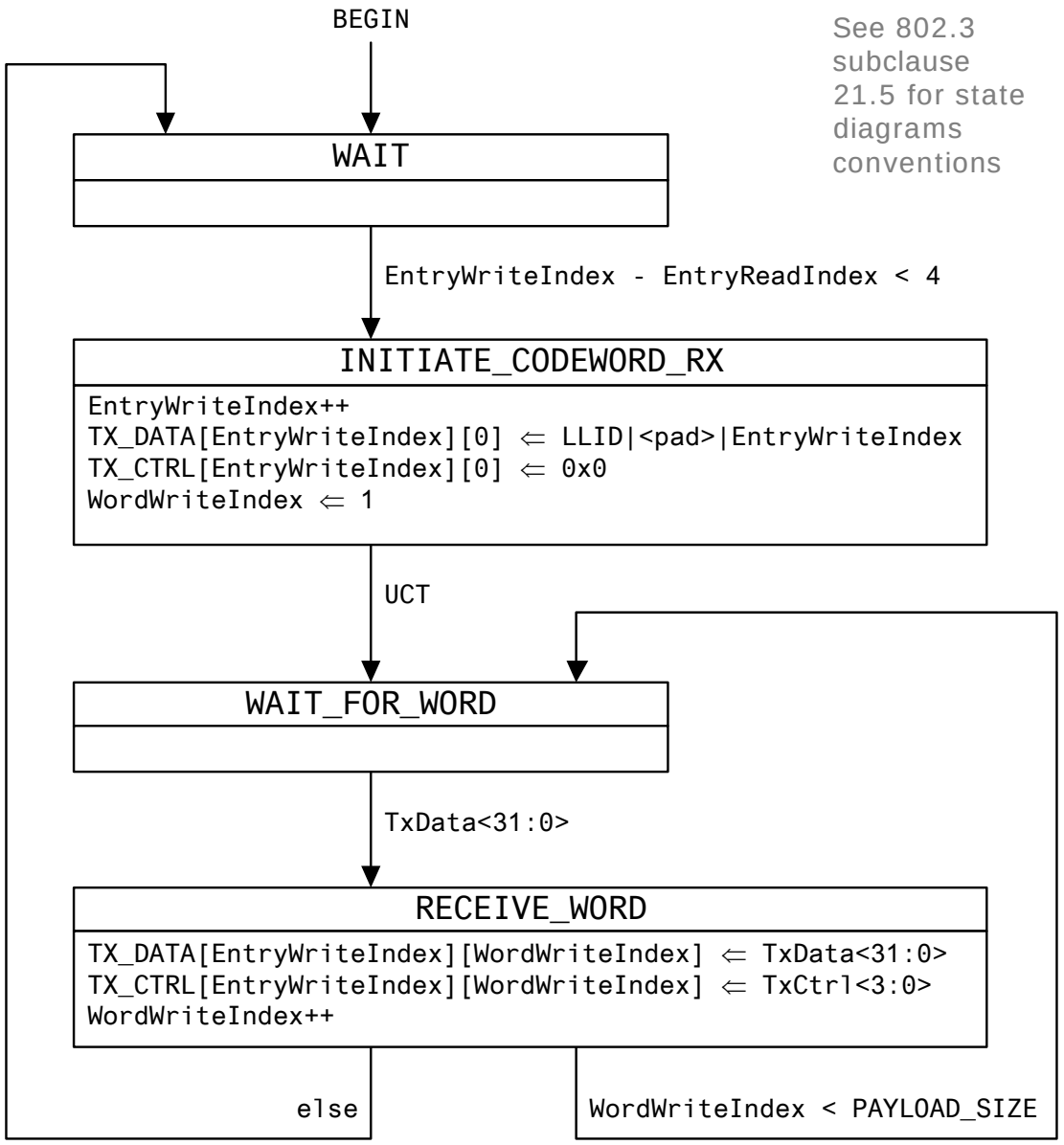


State Diagram Variables

CodewordsLeft ……	Number of codewords that remain to be transmitted on a given lane. This variable can be updated asynchronously via the Channel Bonding Control Process, resulting in a grant being extended.
EntryReadIndex …	3-bit index to the next entry in TX_DATA buffer to be transmitted. This variable is a shared semaphore variable that can be accessed (serially) by 4 output processes.
EntryWriteIndex …	3-bit index to the next available slot in TX_DATA buffer
TX_DATA_ENTRY …	A copy of (or a pointer to) an entry in the TX_DATA buffer that is currently selected for transmission. While a given instance of output process is transmitting the selected entry TX_DATA_ENTRY, other instances may be finishing transmissions of earlier entries or starting transmissions of later entries.
TX_DATA_ENTRY …	A copy of (or a pointer to) an entry in the TX_CTRL buffer that is currently selected for transmission. While a given instance of output process is transmitting the selected entry TX_CTRL_ENTRY, other instances may be finishing transmissions of earlier entries or starting transmissions of later entries.
TxData<31:0> ……	32b vector that becomes available after 32 PLS_DATA.request() calls.
TxCtrl<3:0> ……	control bits corresponding to TxData<31:0>
TXD<31:0> ……	see definition of 25GMII in 802.3by
TXC<3:0> ……	see definition of 25GMII in 802.3by
WordReadIndex …	Index to 32-bit block within a codeword to be read next
WordWriteIndex …	Index to 32-bit block within a codeword to be written next

Channel Bonding Input Process

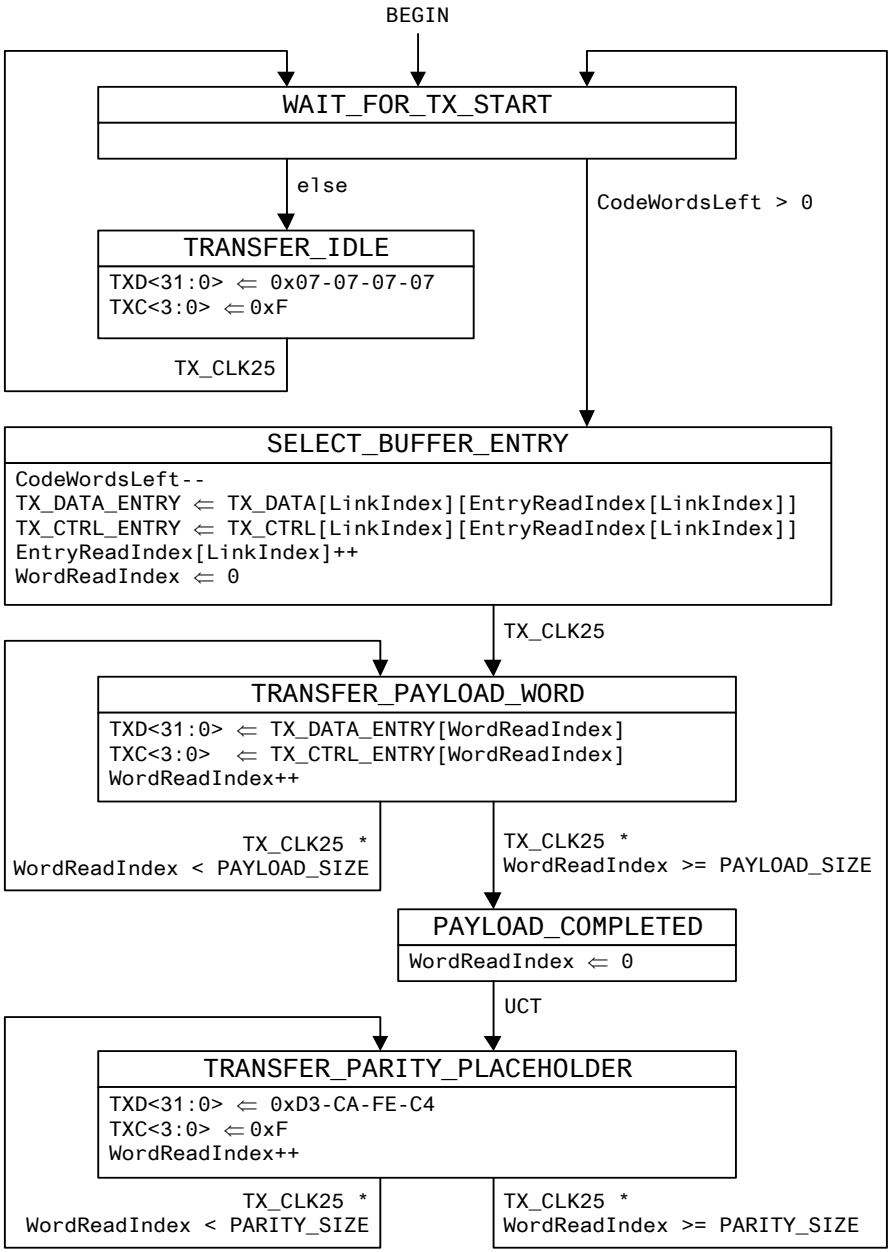
- Main function – ensure that at all times the TX_DATA buffer has 4 complete codewords (since at any time, the MPRS may need to start transmitting on 4 lanes)
- Data is taken from the MAC at 100 Gb/s one FEC payload at a time
- The interval until the next FEC payload is taken from the MAC depends on how many line are transmitting at the same time



See 802.3 subclause 21.5 for state diagrams conventions

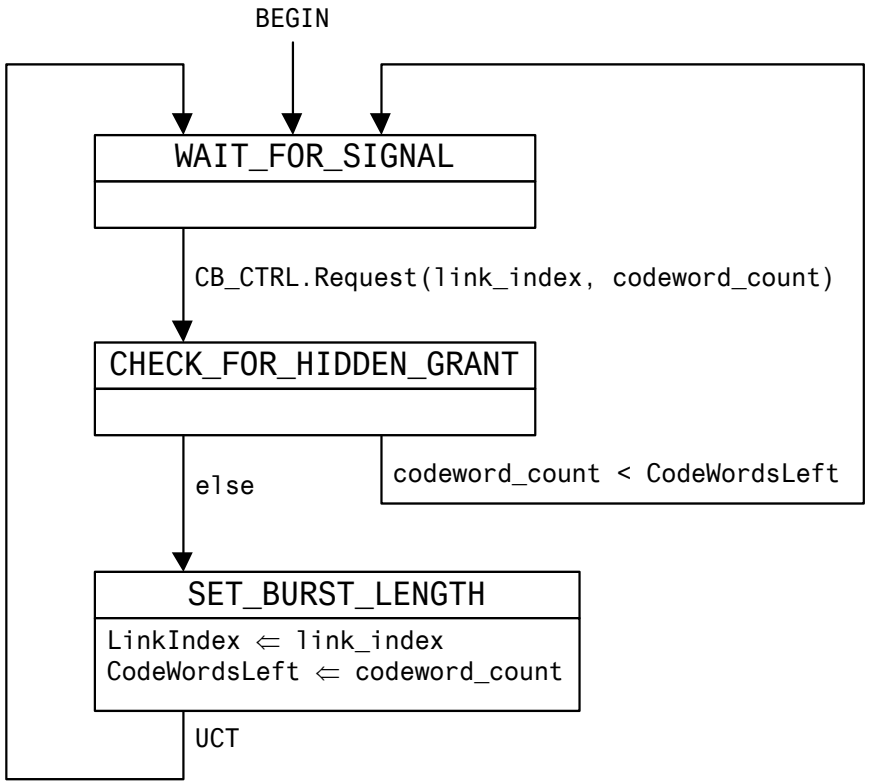
Channel Bonding Transmit Process

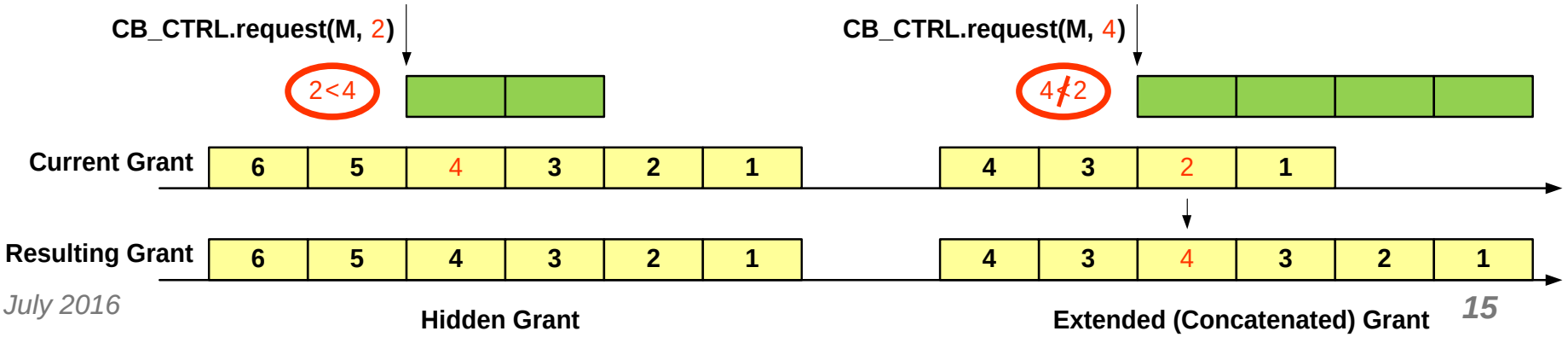
- Main function –transmit full FEC codeword to an active lane
- Transmit process is synchronized on TX_CLK25
- For each LLID, 4 separate transmit processes may be transmitting data at the same time.
- No alignment between transmit processes is required, though implementations will likely run of a single TX_CLK25, i.e., 25GMII blocks (TXD<31:0>) will be aligned.



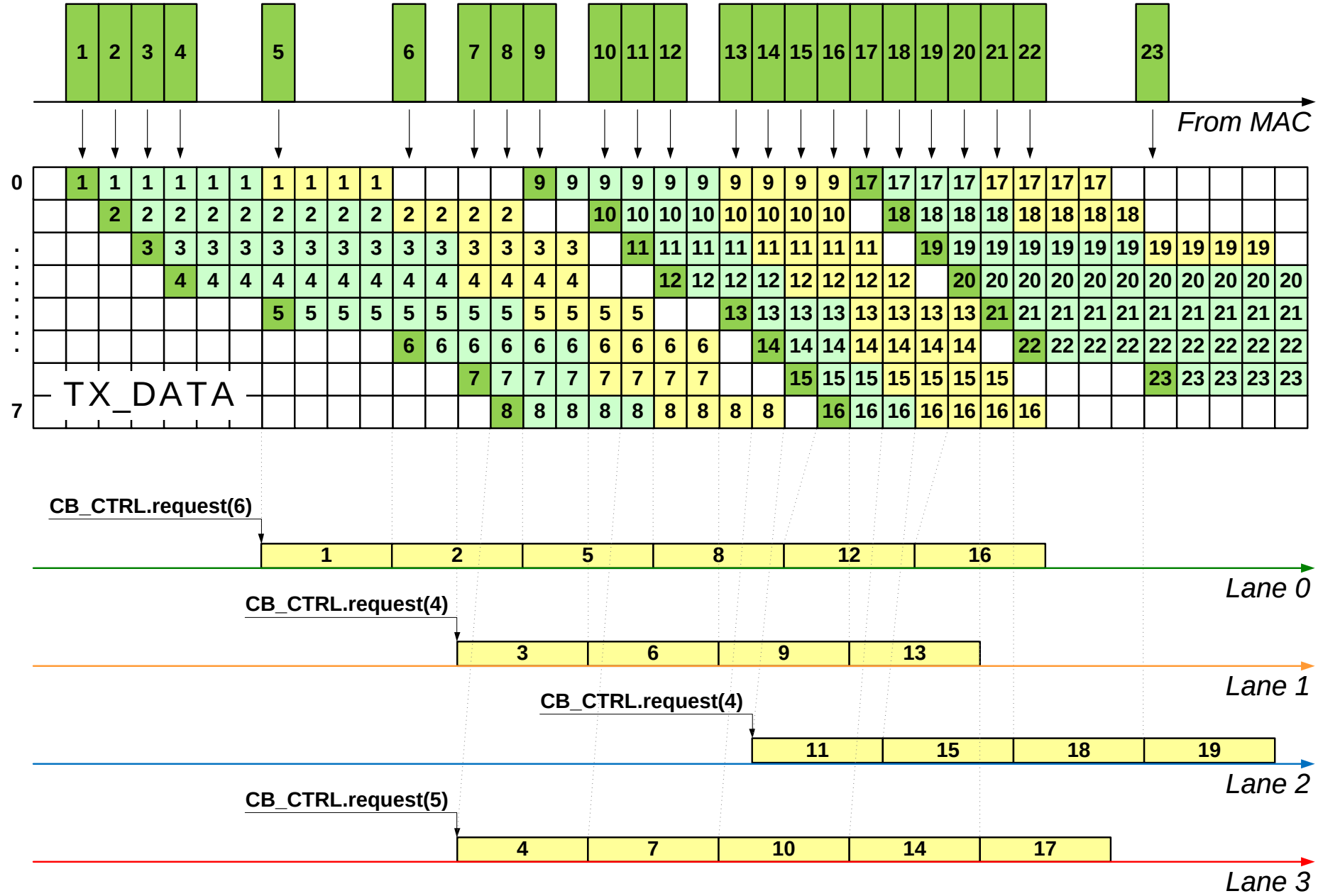
Channel Bonding Control Process

- Main function – process CB_CTRL.request(M, N) primitive
- If the new grant is not a hidden grant (i.e., does not fall entirely within another, larger grant), then the number of FEC codewords remaining to be sent is simply extended to cover the interval to the end of the new grant (concatenated grants feature)

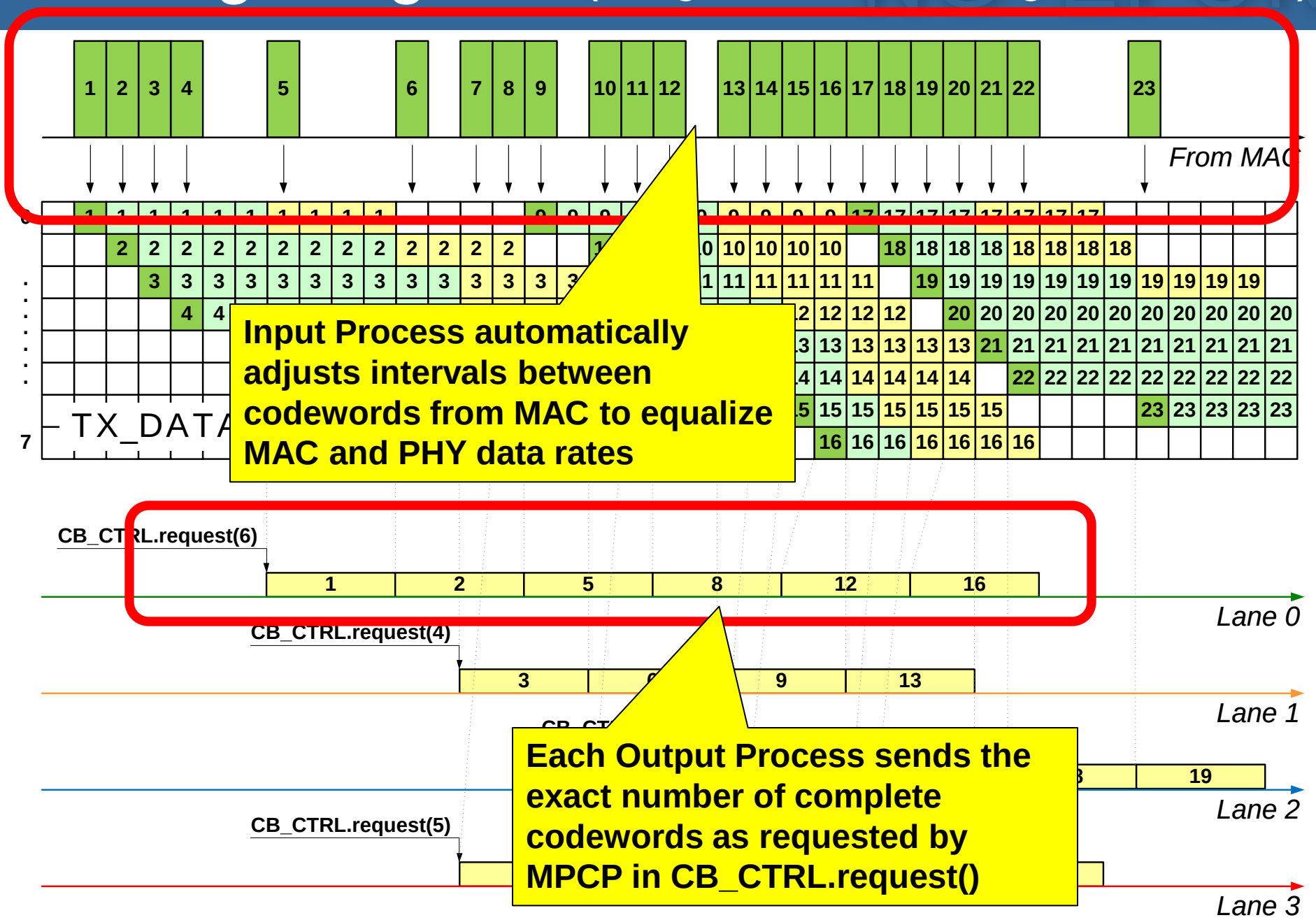




Timing Diagram (Single MAC sending on 4 lanes)



Timing Diagram (Single MAC sending on 4 lanes)



Part 2: Reception Processes at the OLT

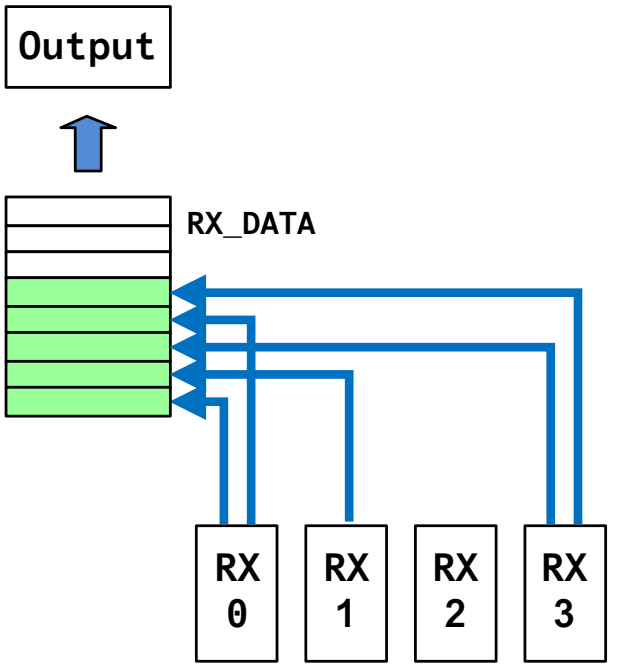
Outline of Rx @ OLT operation

- 4 separate MPRS input processes receive data from 4 lanes
- When a full FEC codeword payload is received, the data is placed into RX_DATA buffer
  - RX_DATA buffer is 8-codewords deep
  - Location in the buffer (index) corresponds to the value of ***CodewordSeqNumber*** received in the FEC codeword header
- A single MPRS output process (per LLID) takes data from RX_DATA sequentially and passes it to MAC using PLS_DATA.indication() primitives.
  - After sending a complete codeword payload to the MAC, the process pauses if the next codeword is not available.

OLT RS Rx Block Diagram

❑Channel Bonding Receive Process

- Receives a complete FEC payload from one 25GMII and stores it in RX_DATA buffer.
- One instance of state diagram per lane



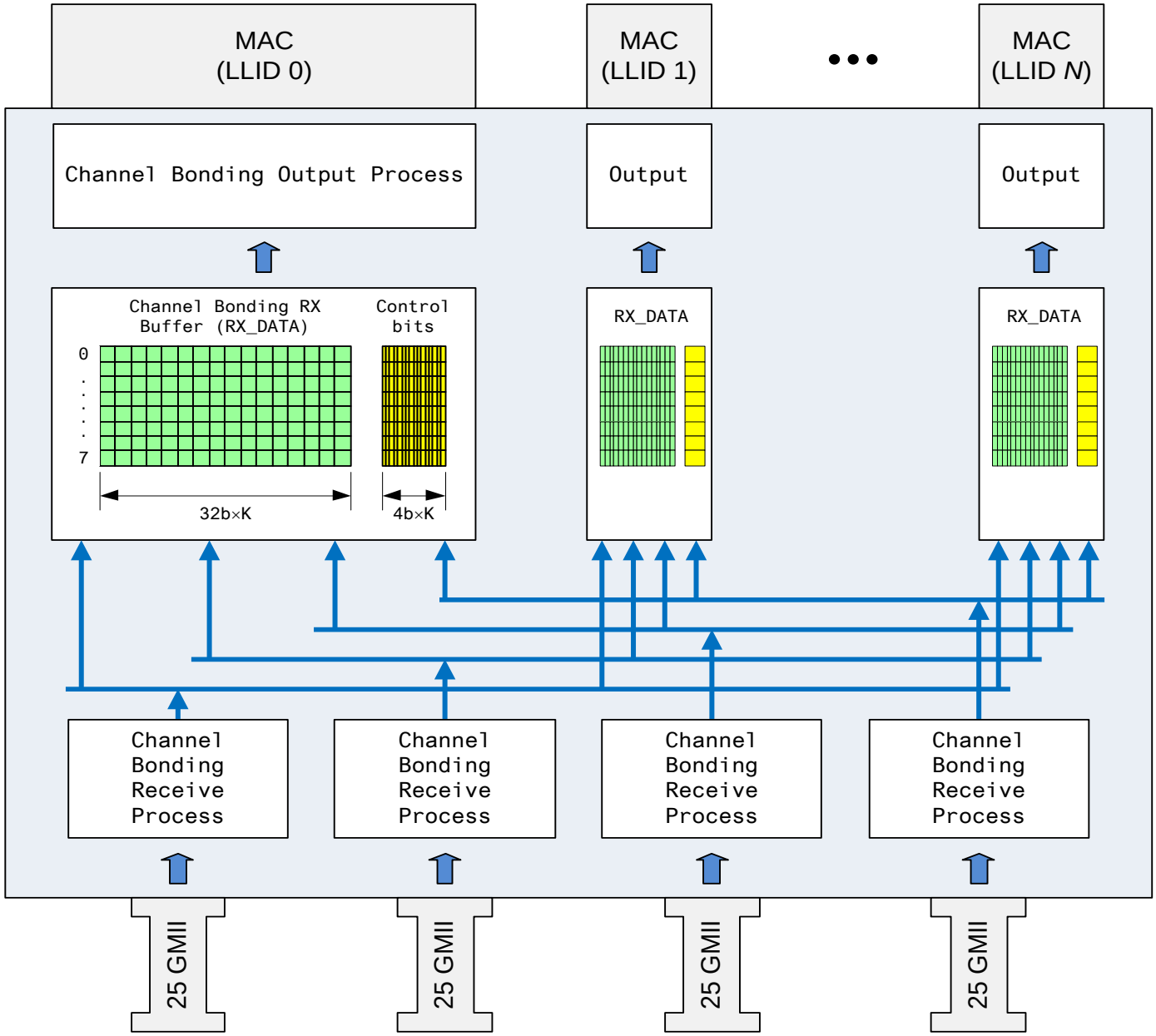
❑Channel Bonding RX Buffer (RX_DATA)

- A circular buffer sufficient to hold 8 FEC codewords payloads
- Allows 4 parallel write streams and a single read stream
- One instance per MAC (LLID)

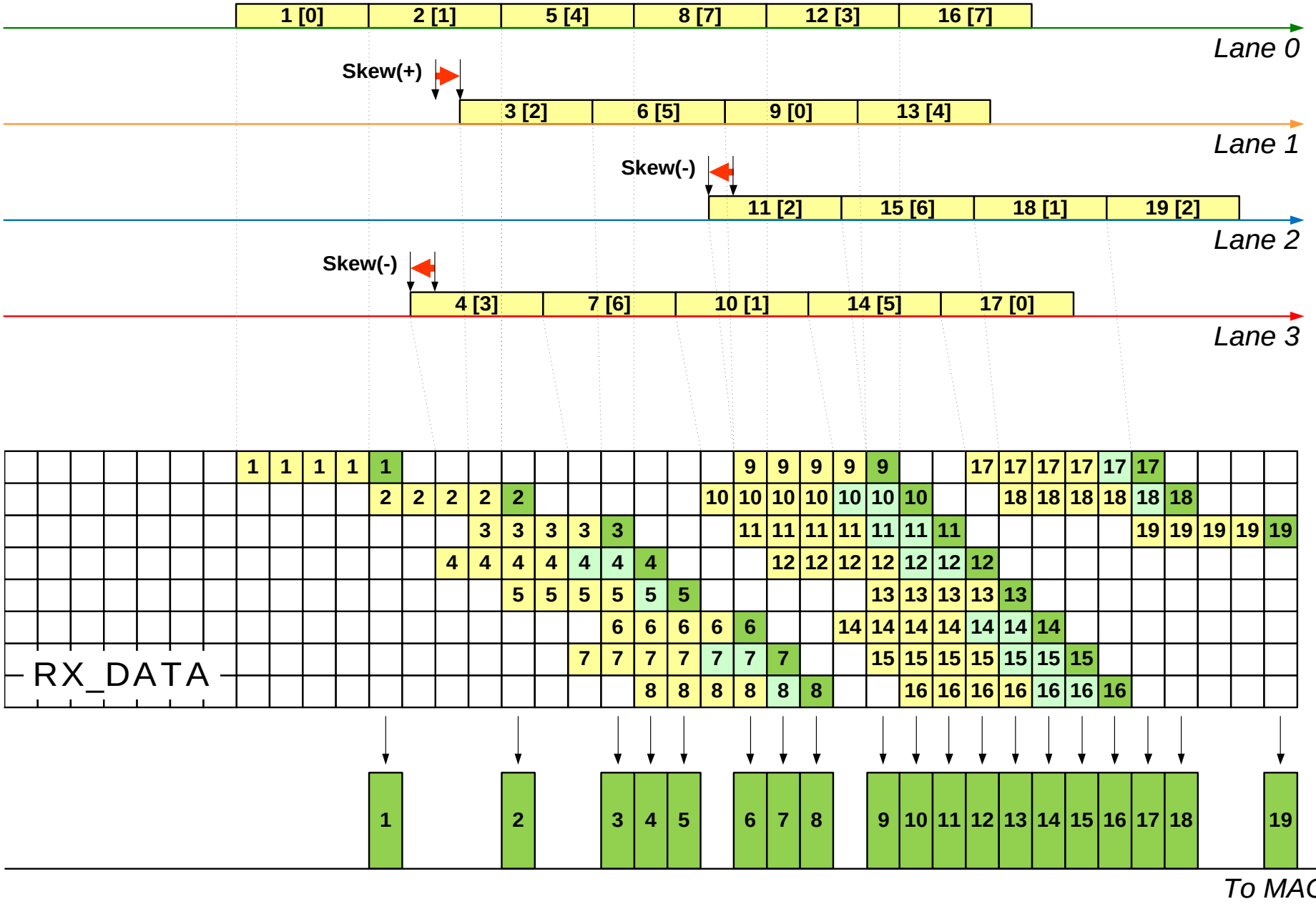
❑ Channel Bonding Output Process

- Transfers full FEC codeword payload from MAC into *Channel Bonding RX Buffer (RX_DATA)* to a MAC
- One instance of state diagram per MAC (LLID)

OLT RS RX Block Diagram



Timing Diagram (Single MAC receiving from 4 lanes)



Next steps

- OLT MPRS RX state diagrams are to be developed next
  - Does not seem to be difficult

- Upstream is more complicated than downstream. Once we figure out the upstream operation, downstream will be trivial(-ish).

Advantages: Simplified MPCP

- An ability to pause the MAC after a certain number of codewords allows for a much simpler MPCP operation
  - **No need to report or grant on packet boundary.**
    - ONU reports just the total queue length.
    - OLT grants a fair share rounded to whole FEC codewords
  - **MPCP does not need to mess with data stream**
    - No need to do idle insertion or deletion
    - No need for MPCP to pre-calculate the location of FEC parity in the data stream
    - No need for MPCP to check whether the next frame fits in a given grant
  - **Built-in support for jumbo frames**

Thank You