Clause 82 LPI Transmit diagram

Jeff Slavick – Avago Technologies

Supporters

- Mike Bennet
- Oren Sela

FEC

- All references to FEC in this presentation refer to the Clause 74 FEC and NOT the Clause 91 RS-FEC

Problems with D1.2 scheme

- **Clause 74 FEC require unscrambled IDLE or LPI characters to be sent for Rapid Framing when exiting the RX_QUIET state.**
- **802.3az we only dealt with FEC on 1 lane and not 4+ trunked lanes**
  - 40G/100G we first have to FEC block_lock, then the PCS can begin framing and acquire align_status. PCS needs RAMs to quickly acquire align_status
- **Rapid Alignment Markers are not /I/ or /LI/ characters**
- **100GBASE-CR10 uses 20 PCS lanes operating over 10x10G PMD lanes, so you only get ½ the data in the same wall clock time**

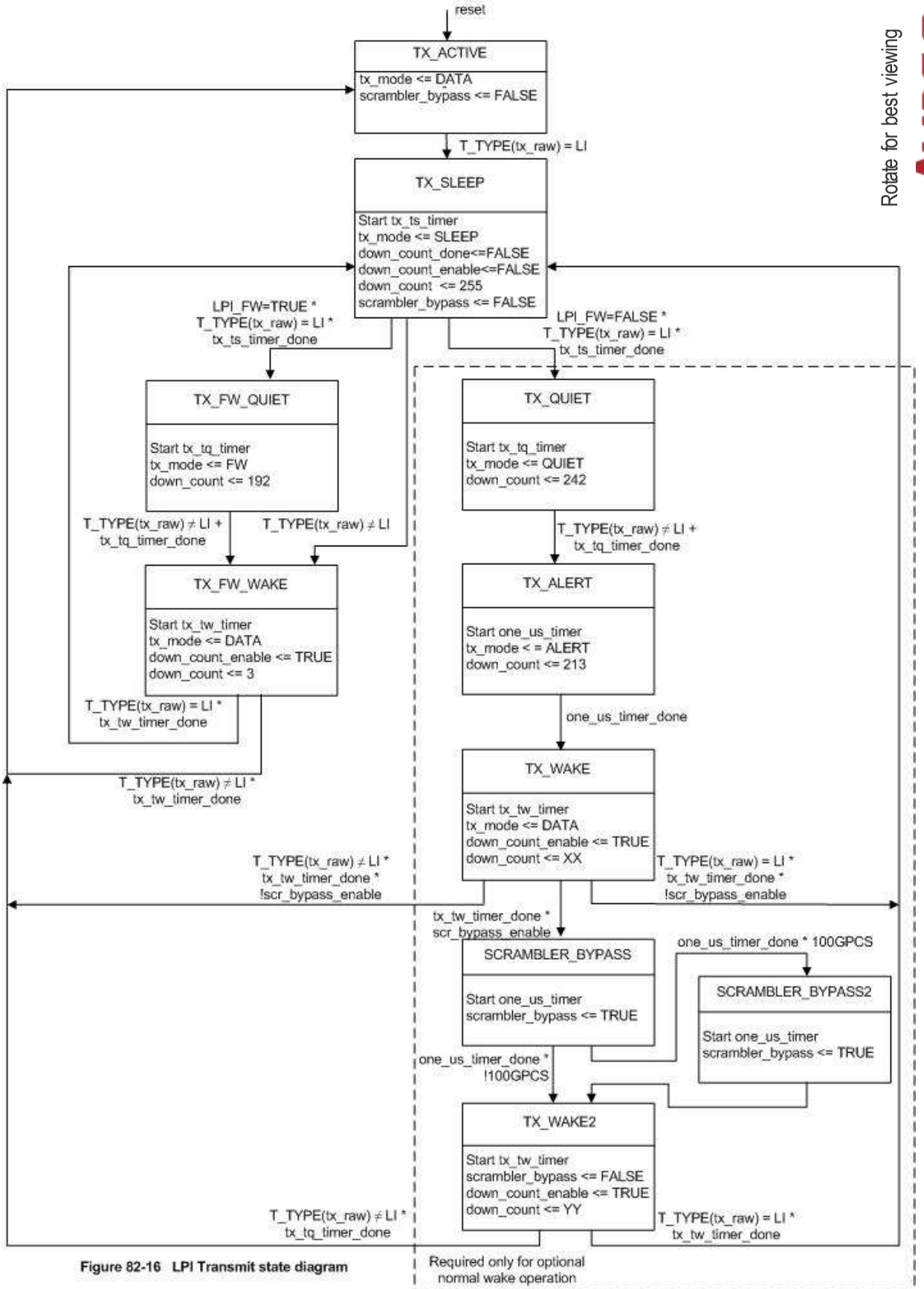
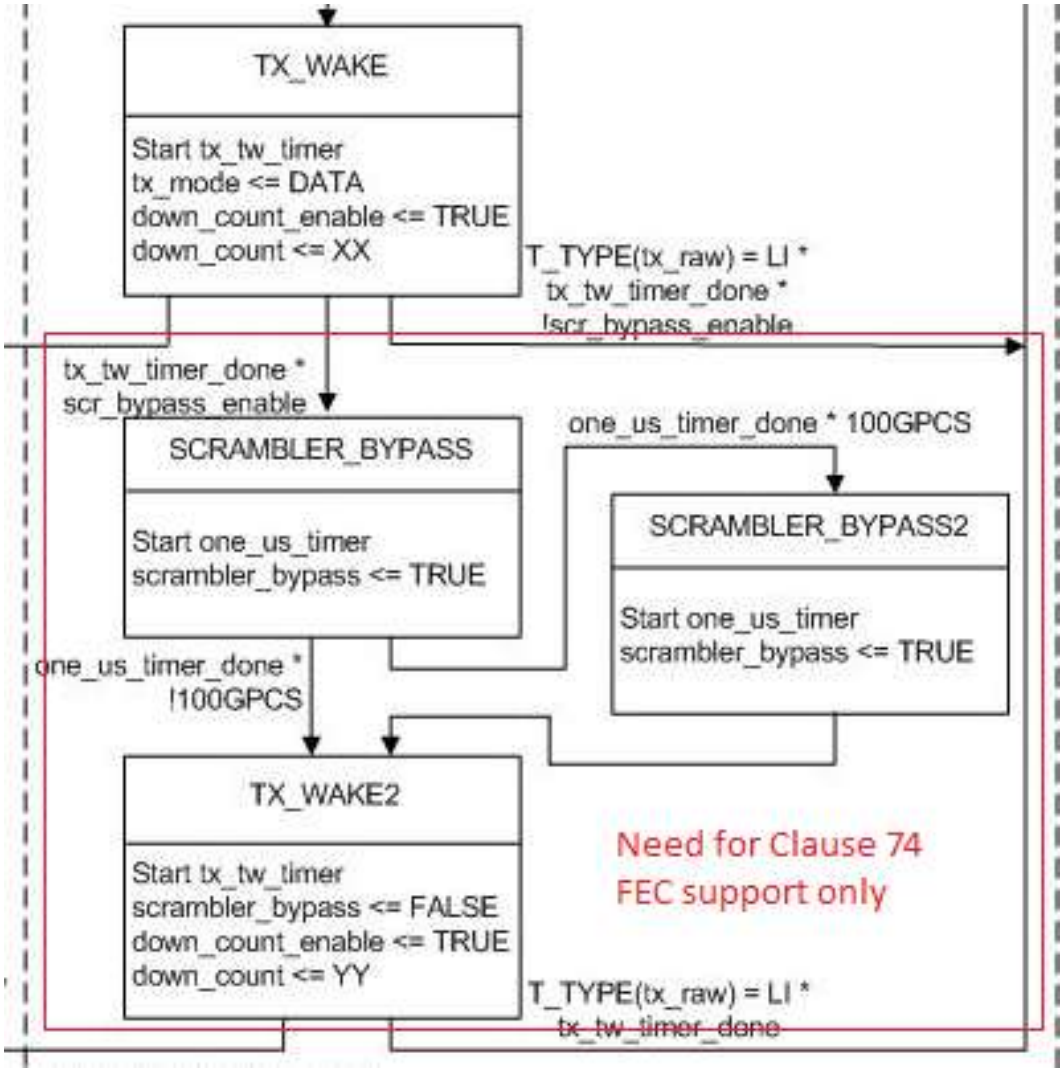


Figure 82-16 LPI Transmit state diagram

New states for FEC support for normal wake



Why do you need TX_WAKE2

- 40G and 100G PCS have multiple PCS lanes that are trunked together unlike the Clause 49 10G PCS.
- The FEC blocks run on individual PCS lanes, and the FEC scrambles the PCS scrambled data.
- So you need to achieve FEC frame lock on each FEC lane before you can begin to align the PCS lanes.

- TX_WAKE provides time for the Rx to “wake”
- SCRAMBLER_BYPASS* provides time for the FECs to frame
- TX_WAKE2 is the time for the PCS to frame and deskew

Why do you terminate RAMs before SCRAMBLER_BYPASS

- The FEC blocks rely on looking for a known data pattern.
- The known data patterns rely on the FEC frame 66b blocks being composed of IDLE or LPI characters
- If you send RAMs during SCRAMBLER_BYPASS then it's possible the RAM will be the first 66b block of a FEC frame. When that occurs you only get a 24b known data pattern, which is insufficient to prevent aliasing to scrambled data.
- Also if RAMs are sent during SCRAMBLER_BYPASS and the FEC is expected to use the RAM as a known data pattern that increases the number of patterns from 2 to 26 that the FEC has to look for. (20-100G AMs and 4-40G AMs)

Why do we need SCRAMBLER_BYPASS2

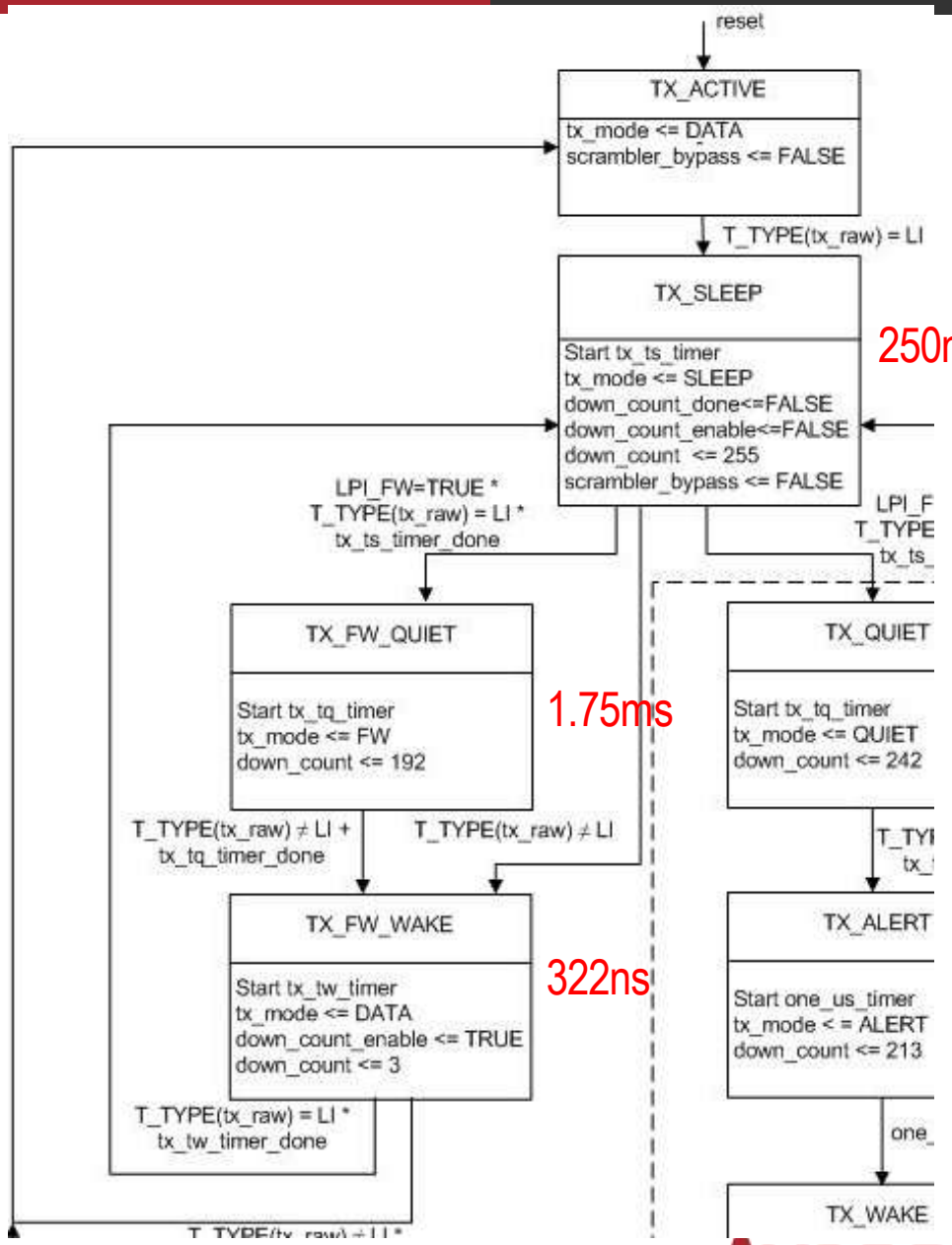
- 100GBASE-CR10 uses 20 PCS lanes running at ~5G and 20 independent FEC instances (one per PCS lane)
- 1.1 µs of scrambler_bypass duration provides 5 full FEC frames when running at ~10G
- If 100G uses a single 1.1 µs duration then each FEC will be guaranteed to see 2 full FEC frames (instead of 5)
- You need 5 FEC frames, and here's what each is for:
  - Frame 1 -> could be corrupted so you can't rely on it, and you want to avoid RX_WTF at all costs
  - Frame 2 -> match known data pattern frame
  - Frame 3 -> time to actually frame
  - Frame 4 -> 1st validly aligned frame
  - Frame 5 -> 2nd validly aligned frame, assert fec_block_lock and signal_ok. PCS waits for signal_ok assertion before it begins it's framing process.

Why do you need SCRAMBLER_BYPASS2 cont.

- Using time in TX_WAKE2 is an option, but would have to be specified how much time is allotted for the FEC to frame so PCS designers know how much time the PCS will have to achieve align_status = true.
- By extending the scrambler_bypass time to 2.2 µs for 100GBASE-CR10 operations the EEE Clause 74 FEC design is the same for 10G-KR, 40G-KR4/CR4 and 100G-CR10

Fast Wake Required implementation for EEE support

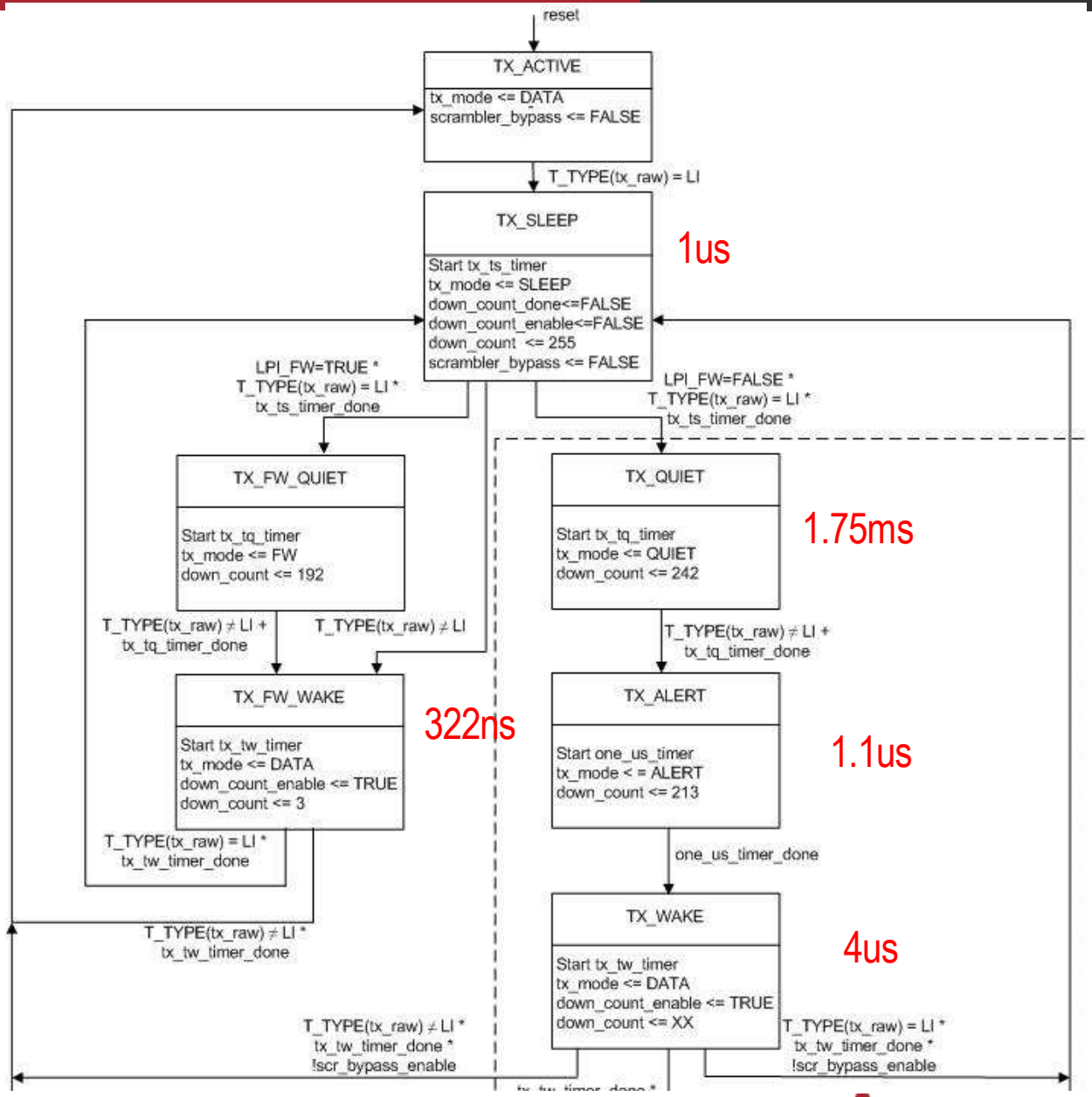
Maximum time for the Tx to exit Low Power Mode when LPI_FW = TRUE is 322ns



Normal Wake not using Clause 74 FEC (optional EEE mode)

Minimum time for the Tx to exit Low Power Mode when LPI_FW = FALSE is 322ns

Maximum time for the Tx to exit Low Power Mode when LPI_FW = FALSE and scr_bypass_enable = FALSE is 5.1 µs



Normal Wake With a Clause 74 FEC Option 1

Minimum time for the Tx to exit Low Power Mode when LPI_FW = FALSE is 322ns

Maximum time for the Tx to exit Low Power Mode when LPI_FW = FALSE and scr_bypass_enable = TRUE is 6.1 µs for 40G

Maximum time for the Tx to exit Low Power Mode when LPI_FW = FALSE and scr_bypass_enable = TRUE is 7.2 µs for 100G

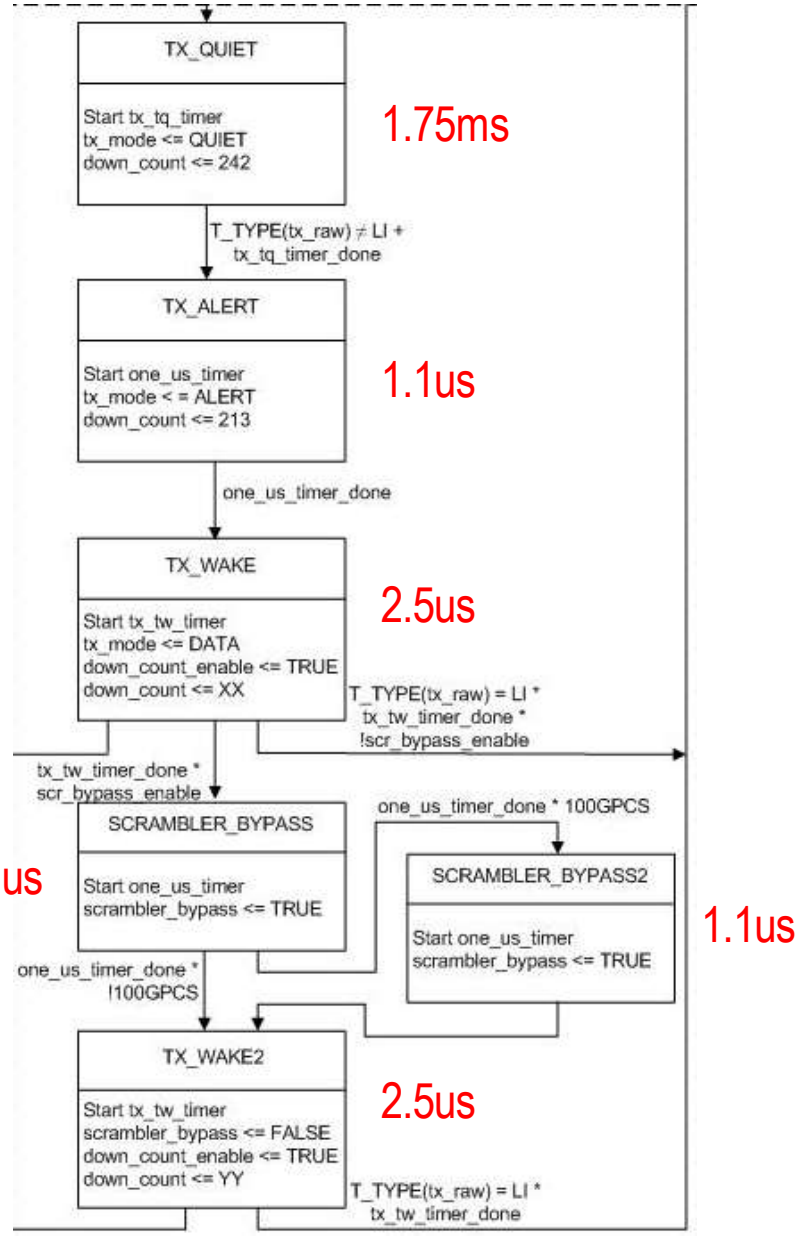


Table 82-5a Option 1

Parameter	Description	Down count	Min	Max	Units
Twl	Time spent in TX_FW_WAKE state	3	312	332	ns
Twl	Time spent in the TX_WAKE state when scr_bypass_enable = FALSE	36	3.9	4.1	µs
Twl	Time spent in the TX_WAKE and TX_WAKE2 states when scr_bypass_enable = TRUE	20	2.4	2.6	µs

- Using these durations extends wake for PHYs that include a Clause 74 FEC from 5.1µs to 6.1 µs (40G) and 7.2µs (100G)

- The wake time for 100GBASE-KR4, 100GBASE-CR4, 100GBASE-KP4 remain the same as D1.2 at 5.1µs when normal wake is enabled.

Normal Wake With a Clause 74 FEC Option 2

Minimum time for the Tx to exit Low Power Mode when LPI_FW = FALSE is 322ns

Maximum time for the Tx to exit Low Power Mode when LPI_FW = FALSE and scr_bypass_enable = TRUE Is 5.1 µs for 40G

Maximum time for the Tx to exit Low Power Mode when LPI_FW = FALSE and scr_bypass_enable = TRUE is 6.2 µs for 100G

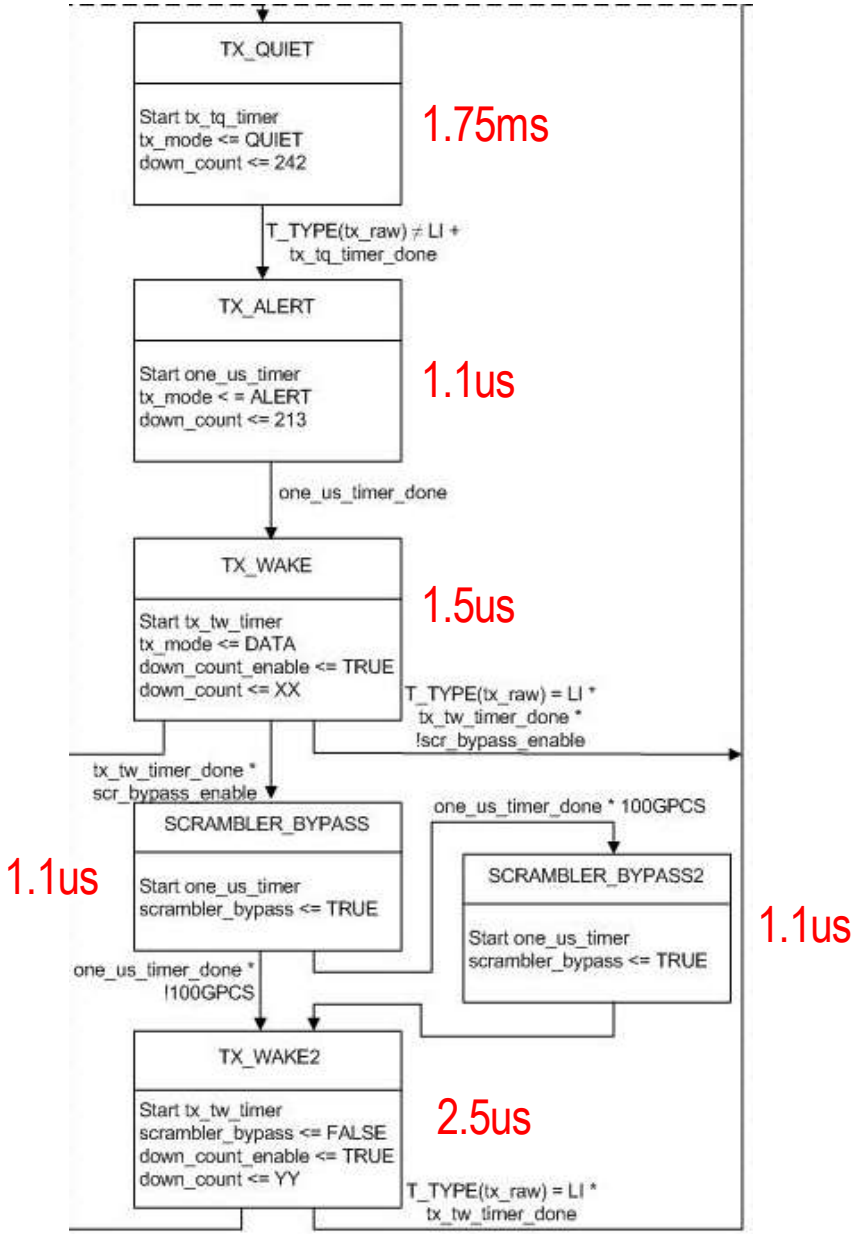


Table 82-5a Option 2

Parameter	Description	Down count	Min	Max	Units
Twl	Time spent in TX_FW_WAKE state	3	312	332	ns
Twl	Time spent in the TX_WAKE state when scr_bypass_enable = FALSE	36	3.9	4.1	µs
Twl	Time spent in the TX_WAKE state when scr_bypass_enable = TRUE	12	1.4	1.6	µs
Twl	Time spent in the TX_WAKE2 state when scr_bypass_enable = TRUE	20	2.4	2.6	µs

- Using these durations extends wake for PHYs that include a Clause 74 FEC from 5.1µs to 6.2µs for 100G only (40G remains at 5.1µs)
- But this reduces the duration the Rx gets to "wake" and provide "good" data from <3.6µs to <2.6µs (TX_ALERT + TX_WAKE) for 40G and 100G PHYs that are using Clause 74 FEC

Recommendation for D1.2 comment #202

- **Adopt the diagram shown in slide 5 as the new Tx LPI Transmit state diagram**
  - This requires the creation of a state variable denoting if the PCS is a 100GPCS in clause 82
- **Adopt the changes presented in slide 16 (Option2) for the Twl time durations of TX_FW_WAKE, TX_WAKE and TX_WAKE2 states**
  - Normal wake times are 5.1 µs, 5.1 µs, 6.2 µs for non-FEC, 40G+FEC, 100G+FEC
  - The value down_count is initialized to is included in the table, but should not be inserted into Table 85-2a, however Figure 82-16 diagram needs to be updated to reflect the values provided in the table.

EEE Wake time timeout for RS-FEC

- Only needed for detached RS-FEC since co-located has tx_mode as a interface signal.
- PCS Rx uses a 2-3ms timeout timer to enter RX_WTF state
- Tx LPI turns data off for maximum 1.8ms

Recommendation for D1.2 comment #208

- **Set minimum timer to 1.8ms for Transmit path**
- **Set maximum timer to 2.0ms for Transmit path**
  - Add 0.2ms to min/max from what the PCS Tx does

- **Set minimum timer to 2.0ms for Receive path (same as PCS)**
- **Set maximum timer to 2.8ms for Receive path**
  - Remove 0.2ms from PCS Rx max timer

- **These numbers cause the RS-FEC to wait just a little bit longer in the Tx and to breakout slightly earlier on the Rx from what the PCS does.**