IBM Systems and Technology Group

FC-FS-4 FC-EE Analysis & Feedback - T11 Document 14-036v0

5-Feb-2014

By: Adrian Butter
asbutter@us.ibm.com

Background

- Focussed review of energy efficient Fibre Channel (FC-EE) feature
  - FC-FS-4 Rev 0.30 (13-113v1.pdf)
    - NOTE: Comments also applicable to latest FC-FS-4 Rev 0.30 (13-445v0.pdf) release...
  - 128GFC Architecture Text (13-369v0.pdf)

- Approach – Side-by-side comparison with 802.3 energy efficient Ethernet (EEE) feature
  - Especially 802.3az-2010 Clause 49 (PCS for 64B/66B, type 10GBASE-R) and 802.3bj Clause 82 (PCS for 64B/66B, type 40GBASE-R and 100GBASE-R)

- Results -
  - 23 issues documented and posted to T11 website:

Issues Document #	T11 Upload Date	# of FC-FS-4 Issues	# of 128GFC Architecture Text Issues	Total Issues
13-501v0	12/10/13	16	0	16
13-501v1	12/10/13			
14-010v0	01/16/14	19	2	21
14-010v1	***01/24/14***	***21***	***2***	***23***

  - Source for documenting issues is the same spreadsheet template used in the T11 balloting process:
    - Comment ID, Type of Comment, Page/Line Number/Index pertaining to Comment, Comment, Proposed Solution

- 4 of 23 issues documented likely require more information to properly disposition
  - Issues IBM-23, IBM-18, IBM-21, IBM-22
  - Such additional information is provided in the remainder of this presentation...

Issue IBM-23

- FC-FS-4 LPI state diagram transitions lack clarity and precision

- Example content layout from 802.3az-2010, Clause 49
  - NOTE: Update as needed to be consistent with FC-EE requirements...

49.2.13.2.2 Variables

The following variables are used only for the EEE capability:

energy_detect
: A Boolean variable sent from the PMD that is set to TRUE when signal energy is detected at the receiver and is set to FALSE otherwise

rx_block_lock
: Variable used by the lock state diagram to reflect the status of the code-group delineation. This variable is set TRUE when the receiver acquires block delineation.

rx_lpi_active
: A Boolean variable that is set to TRUE when the receiver is in a low power state and set to FALSE when it is in an active state and capable of receiving data.

rx_mode
: A variable set to QUIET while the receiver is in the RX_QUIET state and set to DATA otherwise

tx_mode
: A variable set to QUIET when the transmitter is in the TX_QUIET state, set to ALERT when the transmitter is in the TX_ALERT state, and set to DATA otherwise. When set to QUIET, the PMD disables the transmitter as described in 72.6.5. When set to ALERT, the PMD transmits a repeating pattern of eight ones and eight zeroes as described in 72.6.2. When set to DATA the PMD passes data as normal.

scrambler_bypass
: This Boolean variable is used to bypass the Tx PCS scrambler in order to assist rapid synchronization following low power idle. When set to TRUE, the PCS will pass the unscrambled data from the scrambler input rather than the scrambled data from the scrambler output. The scrambler will continue to operate normally, shifting input data into the delay line. When scrambler_bypass is set to FALSE the PCS will pass scrambled data from the scrambler output.

scr_bypass_enable
: A Boolean variable used to indicate to the transmit LPI state diagram that the scrambler bypass option is required. The PHY shall set scr_bypass_enable = TRUE if Clause 74 FEC is in use. The PHY shall set scr_bypass_enable = FALSE if this FEC is not in use.

49.2.13.2.3 Functions

R_TYPE(rx_coded<65:0>)
: Returns the R_BLOCK_TYPE of the rx_coded<65:0> bit vector.

T_TYPE(tx_raw<71:0>)
: Returns the T_BLOCK_TYPE of the tx_raw<71:0> bit vector.

49.2.13.2.4 Counters

The following counter is used only for the EEE capability.

wake_error_counter
: A counter that is incremented each time that the LPI receive state diagram enters the RX_WTF state indicating that a wake time fault has been detected. The counter is reflected in register 3.22 (see 45.2.3.8b).

49.2.13.2.5 Timers

The following timers are used only for the EEE capability:

one_us_timer
: A timer used to count approximately 1 µs intervals. The timer terminal count is set to T_{1U}. When the timer reaches terminal count it will set the one_us_timer_done = TRUE.

rx_tq_timer
: This timer is started when the PCS receiver enters the RX_SLEEP state. The timer terminal count is set to T_{QR}. When the timer reaches terminal count it will set the rx_tq_timer_done = TRUE.

rx_tw_timer
: This timer is started when the PCS receiver enters the RX_WAKE state. The timer terminal count shall be set to a value no larger than the maximum value given for T_{WR} in Table 49–3. When the timer reaches terminal count it will set the rx_tw_timer_done = TRUE.

rx_wf_timer
: This timer is started when the PCS receiver enters the RX_WTF state, indicating that the receiver has encountered a wake time fault. The rx_wf_timer allows the receiver an additional period in which to synchronize or return to the QUIET state before a link failure is indicated. The timer terminal count is set to T_{WTF}. When the timer reaches terminal count it will set the rx_wf_timer_done = TRUE.

tx_ts_timer
: This timer is started when the PCS transmitter enters the TX_SLEEP state. The timer terminal count is set to T_{SL}. When the timer reaches terminal count it will set the tx_ts_timer_done = TRUE.

tx_tq_timer
: This timer is started when the PCS transmitter enters the TX_QUIET state. The timer terminal count is set to T_{QL}. When the timer reaches terminal count it will set the tx_tq_timer_done = TRUE.

tx_tw_timer
: This timer is started when the PCS transmitter enters the TX_WAKE state. The timer terminal count is set to T_{WL}. When the timer reaches terminal count it will set the tx_tw_timer_done = TRUE.

Issue IBM-23 (cont)

- Example content layout from 802.3az-2010, Clause 49 (cont)
  - NOTE: Update as needed to be consistent with FC-EE requirements...

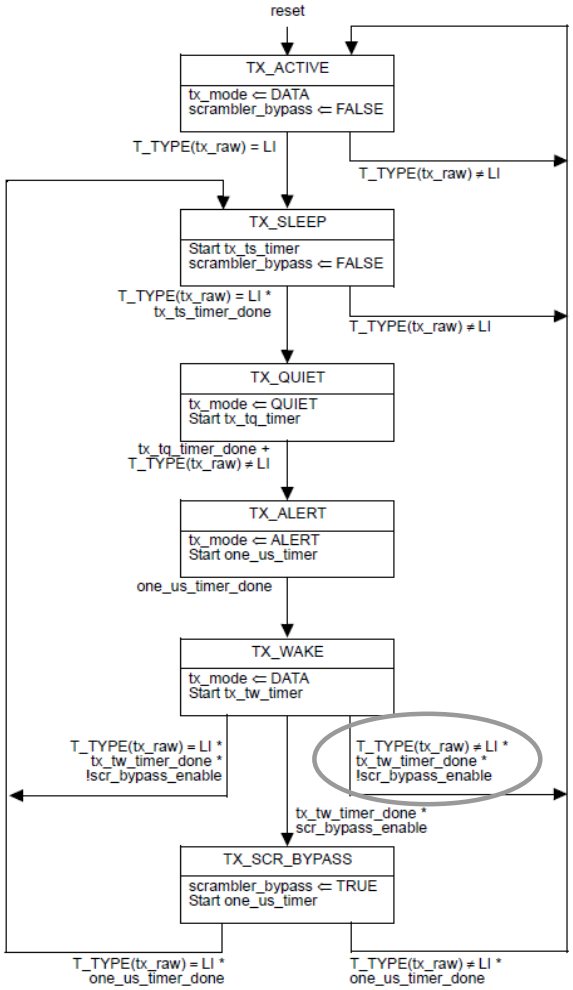

Figure 49–16—LPI Transmit state diagram

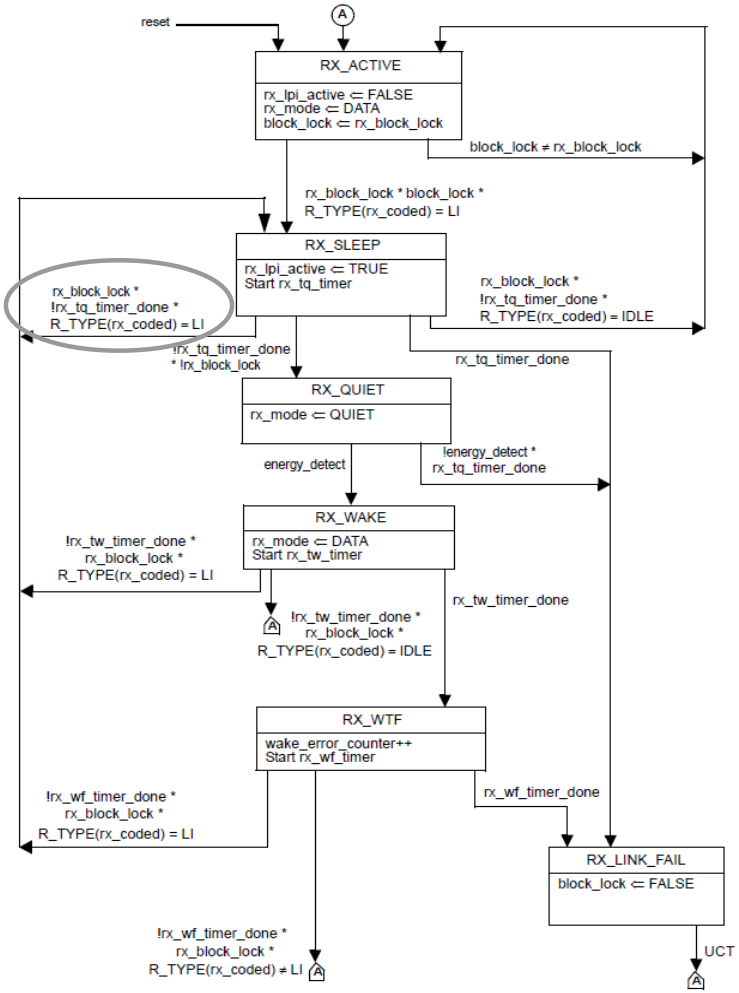

Figure 49–17—LPI Receive state diagram

Issue IBM-18

- FC-FS-4 Section 10.5 mentions an LPI Mode use case in which the transmitter continually sends LPI...

  During the quiet cycle, some transceiver types may not be capable of turning off the transmitter/receiver. In this case, LPI shall be transmitted during the LPI Mode in order to indicate low power operation, this allows the port to turn off unused capabilities to save power.

  - However, current LPI Mode Transmitter state diagram does not support this use case...

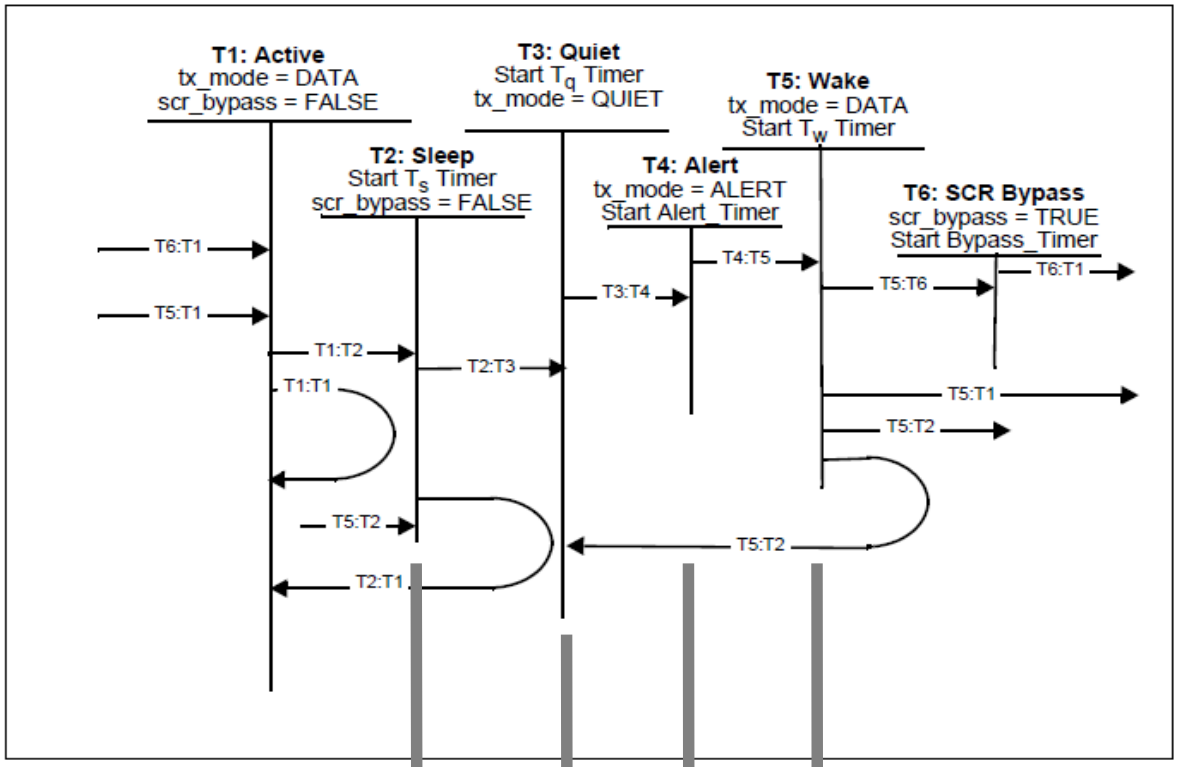

Figure 49 LPI Mode Transmitter State Diagram

Transmitted Frame: LPI LPI ALERT LPI (for Refresh sequence) or IDLE (for Wake sequence)

Issue IBM-18 (cont)

- FC-FS-4 Section 10.5 mentions an LPI Mode use case in which the transmitter continually sends LPI...
  - … and current LPI Mode Receiver state diagram does not support this use case:

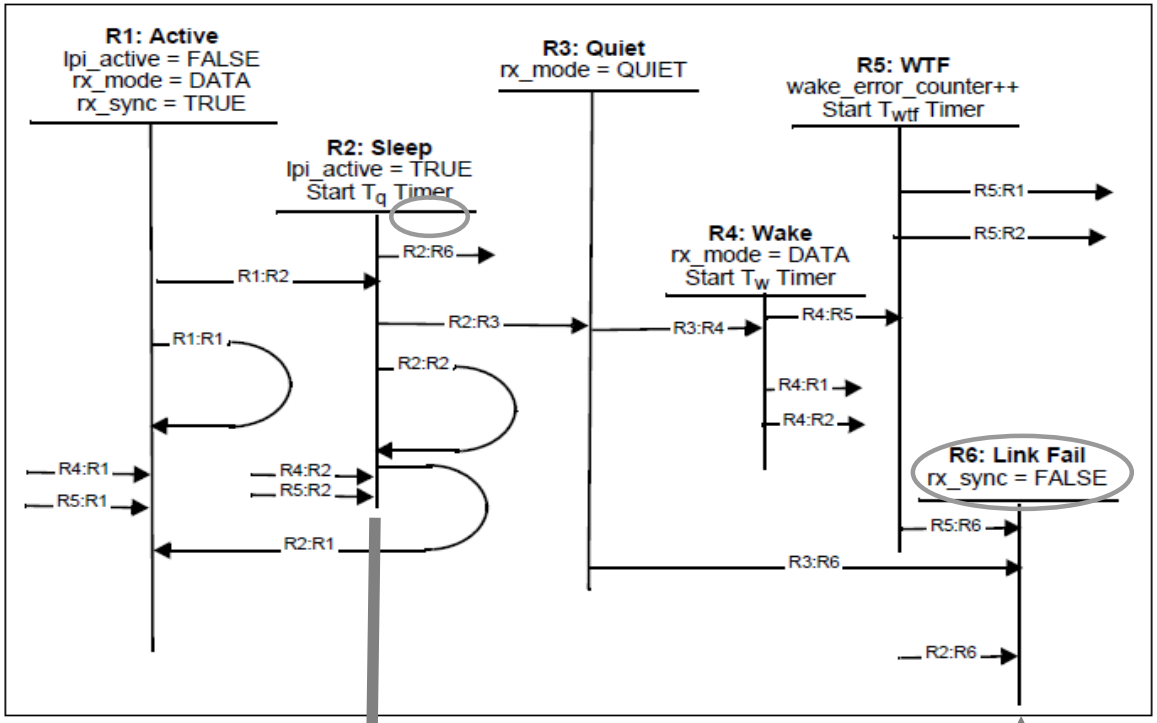

Figure 50 - LPI Mode Receiver State Diagram

Since transmitter continually sends frames, receiver maintains sync (rx_sync=TRUE) throughout R2:SLEEP:

If all frames are LPI, then only R2:R6 applies after Tq expires...

Issue IBM-18 (cont)

- Proposed updates to FC-FS-4 to support the LPI Mode use case in which the transmitter continually sends LPI:
  - Define Transmitter and Receiver “Fast Wake” states (similar to 802.3bj Figure 82-16 & 82-17)

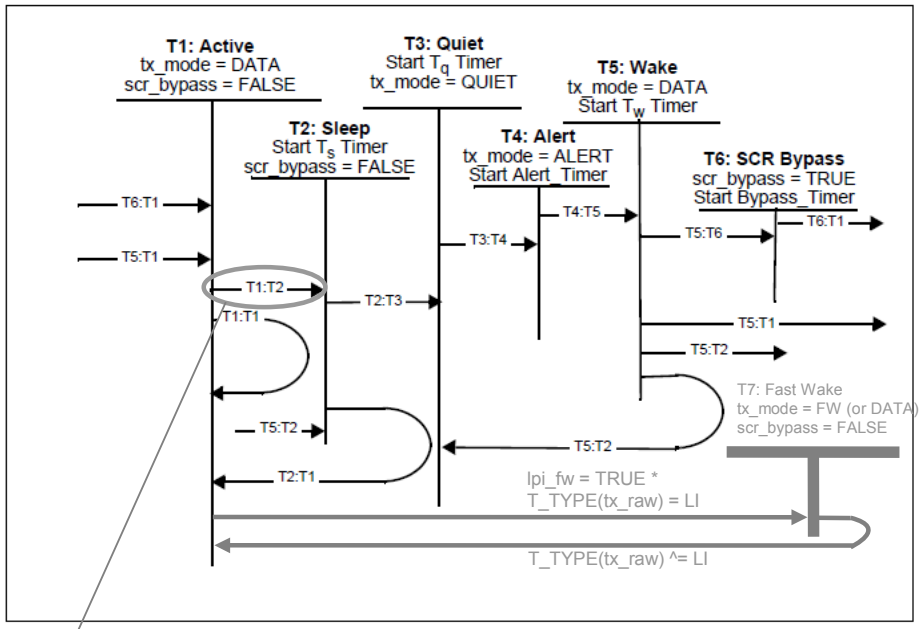

Figure 49 - LPI Mode Transmitter State Diagram

NOTE: Must also check for lpi_fw = FALSE to make T1:T2 transition...

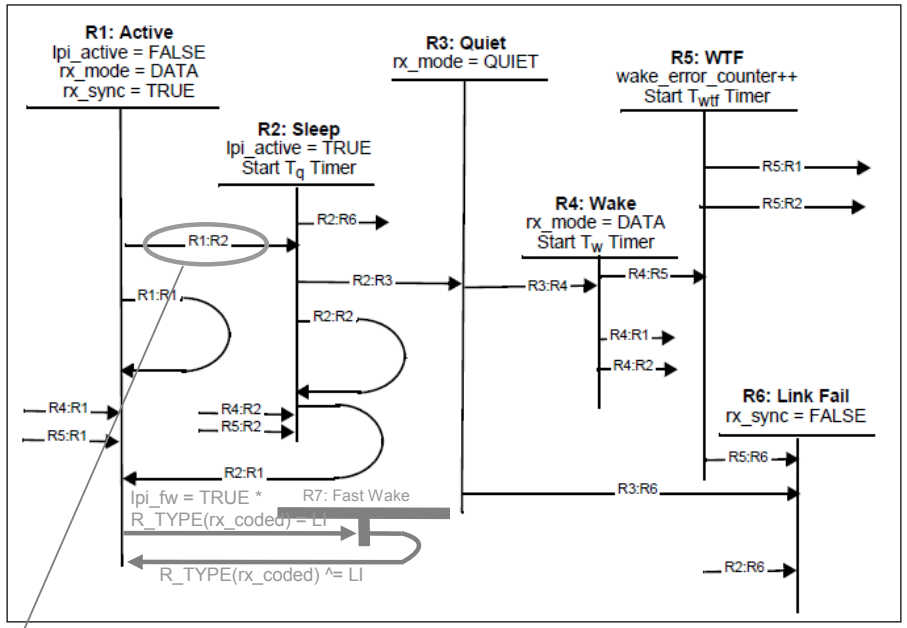

Figure 50 - LPI Mode Receiver State Diagram

NOTE: Must also check for lpi_fw = FALSE to make R1:R2 transition...

Issue IBM-21

- FC-FS-4 provides inadequate detail for FC-EE support @ 128G
- Proposed starting point for specifying FC-EE support @ 128G are LPI Mode state diagrams in 802.3bj Clause 82:
  - Note proposed changes to LPI Receive state diagram RX_FW state for 128G FC-EE...

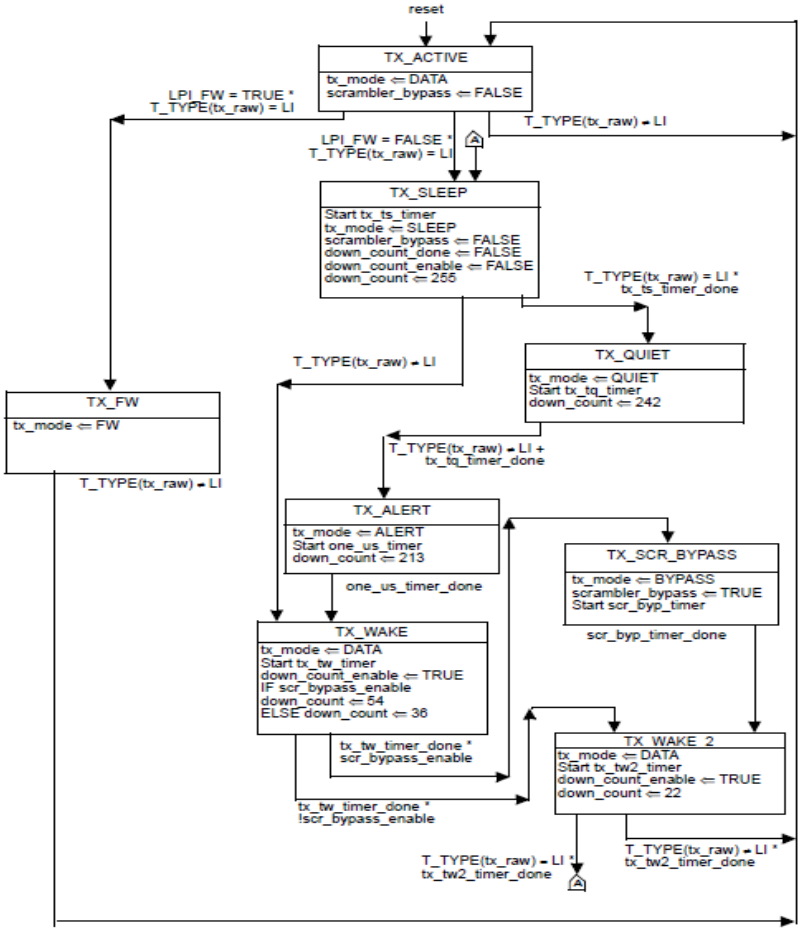

Figure 82–16—LPI Transmit state diagram

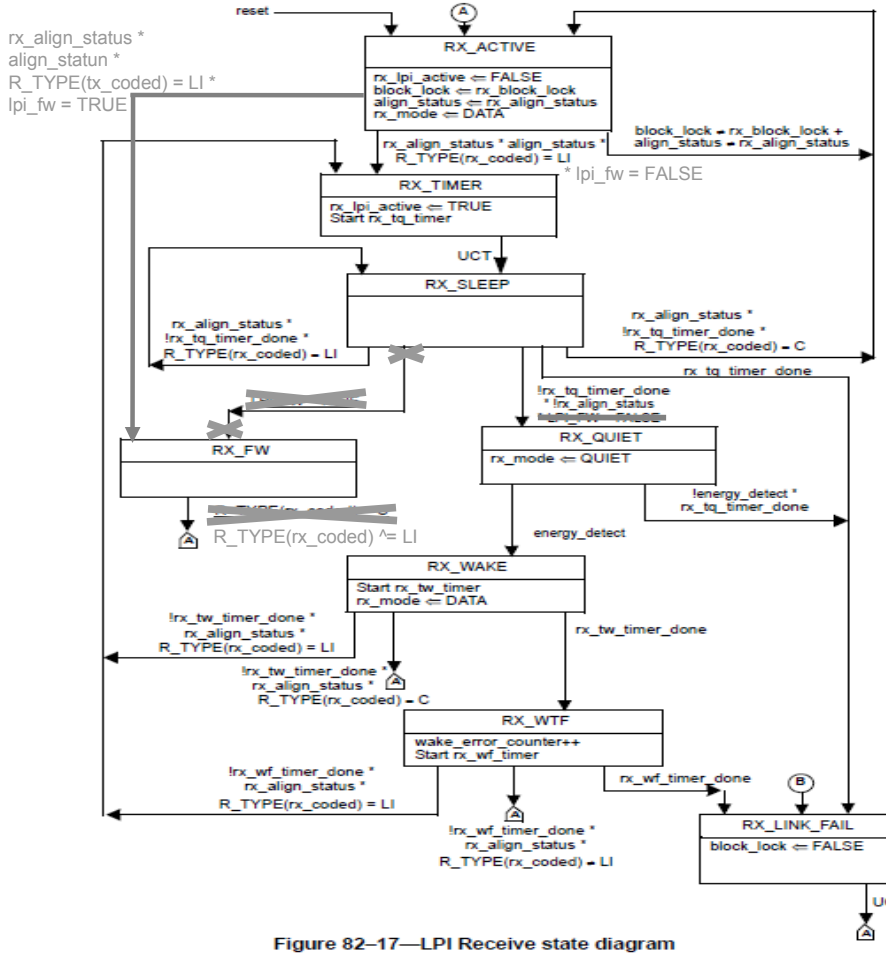

Figure 82–17—LPI Receive state diagram

IBM-18/IBM-21 Additional Consideration

- Adding Fast Wake requires adding associated Link Layer services
  - 802.3bj does this via Auto-Negotiation and exchanging TLV messages
    - References: 30.12 Layer Management for Link Layer Discovery Protocol (LLDP); 45.2.3.9 EEE Control and Capability (Register 3.20); 78.4 Data Link Layer Capabilities; 79 IEEE 802.3 Organizationally Specific Link Layer Discovery Protocol (LLDP) type, length and value (TLV) information elements
    - EEE Fast Wake TLV format and field definition:

79.3.6 EEE Fast Wake TLV

The EEE Fast Wake TLV is used to exchange information about the EEE Fast Wake capabilities. This TLV is only used by systems operating at links speeds >10 Gb/s. Figure 79–7 shows the format of this TLV.

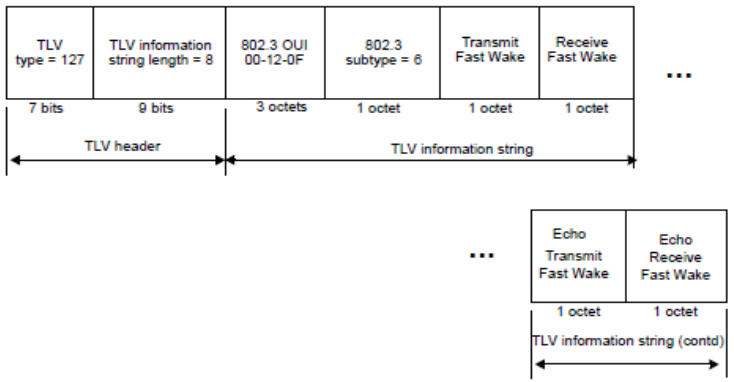

Figure 79–7—EEE Fast Wake TLV format

79.3.6.1 Transmit Fast Wake

Transmit Fast Wake (1 octet wide) is a logical indication that the transmit LPI state diagram intends to use the Fast Wake function (corresponding to the variable LPI_FW in 82.2.18.2.2). Transmit Fast Wake = 1 corresponds to LPI_FW being TRUE; Transmit Fast Wake = 0 corresponds to LPI_FW being FALSE. The default value for Transmit Fast Wake is 1 (TRUE). Transmit Fast Wake is set to TRUE unless the PHY is capable of deep sleep operation as determined by the PHY type and the results of auto-negotiation.

79.3.6.2 Receive Fast Wake

Receive Fast Wake (1 octet wide) is a logical indication that the receive LPI state diagram is expecting its link partner to use the Fast Wake function (corresponding to the variable LPI_FW in 82.2.18.2.2). Receive Fast Wake = 1 corresponds to LPI_FW being TRUE; Receive Fast Wake = 0 corresponds to LPI_FW being FALSE. The default value for Receive Fast Wake is 1 (TRUE). Receive Fast Wake is set to TRUE unless the PHY is capable of deep sleep operation as determined by the PHY type and the results of auto-negotiation.

79.3.6.3 Echo of Transmit Fast Wake and Receive Fast Wake

The respective echo values are the local link partner's reflection (echo) of the remote link partner's respective values. When a local link partner receives its echoed values from the remote link partner, it can determine whether or not the remote link partner has received, registered and processed its most recent values. For example, if the local link partner receives echoed parameters that do not match the values in its local MIB, then the local link partner infers that the remote link partner's request was based on stale information.

IBM-18/IBM-21 Additional Consideration (cont)

- Adding Fast Wake requires adding associated Link Layer services (cont)

  - For FC-LS-3, propose adding Transmit and Receive Fast Wake (FW) to the Exchange Energy Efficient Parameters Extended Link Service (EEEP ELS) Descriptor field after the Transmit and Receive Wake time (Tw) fields:

Table 8 – EEEP Descriptor Format

Bits Word	31 .. 24 Byte 0	23 .. 16 Byte 1	15 .. 08 Byte 2	07 .. 00 Byte 3
0	Descriptor Type = <TBD>h			
1	Length = X 8			
2	Transmit T_w		Receive T_w	
3	Transmit FW	Echo Transmit FW	Receive FW	Echo Receive FW

  - For FC-SW-6, propose adding those same parameters to the EEEP Switch Fabric Internal Link Service (SW_ILS) Descriptor field after the Transmit and Receive Wake time (Tw) fields:

Table 11 – EEEP Descriptor

Item	Size Bytes
Tag Value = <TBD>h	4
Length = X 8	4
Transmit T_w	2
Receive T_w	2
Transmit FW	1
Echo Transmit FW	1
Receive FW	1
Echo Receive FW	1

Issue IBM-22

- In FC-FS-4, provide more clarity regarding Wake time negotiation
  - Current text in Section 10.1 -

    Energy Efficient operation is negotiated per link using a login bit either in the FLOGI/PLOGI, for N_Ports, or in the ELP for E_Ports (see FC-LS-3). Wake parameters may be set using the EEEP ELS, for N_Ports (see FC-LS-3), or the EEEP SW_ILS, for E_Ports (see FC-SW-6).

  - Proposed text update in Section 10.1 -

    FC-EE specifies means to exchange capabilities between link partners to determine whether Energy Efficient operation is supported and to select the best set of parameters common to both devices. Energy Efficient operation is negotiated per link using a login bit either in the FLOGI/PLOGI, for N_Ports (see FC-LS-3), or in the ELP for E_Ports (see FC-SW-6). Devices use the Exchange Energy Efficient Parameters Extended Link Service defined in FC-LS-3 or Switch Fabric Internal Link Service defined in FC-SW-6 to negotiate system wake-up times and fast wake/deep sleep support from the transmitting link partner via wake time (Tw) and Fast Wake (FW) parameter exchange, respectively. The result of this negotiation determines the extent to which Energy Efficient operation is supported across a given link, and allows the system to select more or less aggressive energy saving modes.

  - ***NOTE:***

    Neither includes the Exchange Energy Efficient Parameters service described in FC-EE Section 2 (Negotiation), so further update of those documents to include this service is required...