SOLARFLARE Communications

Fast link recovery proposal

11/18/2009 Gavin Parnaby

Overview of the problem

- Base standard 10GBASE-T link adaptation is limited since THP coefficients cannot be adapted
- In the base standard there is no method to change THP coefficients apart from a full retrain
- Full retrain is 2s
  - **plus autoneg in some cases, plus some probability the training is repeated**
- EEE mode **must** be somewhat less robust than the normal mode since only 4 out of 512 frames are useful data for adaptation
- If the receiver detects the link is bad the only option currently available is the full 2s retrain (EEE/non-EEE) or a vendor proprietary solution
- Also addresses dead link issue in rx-lpi

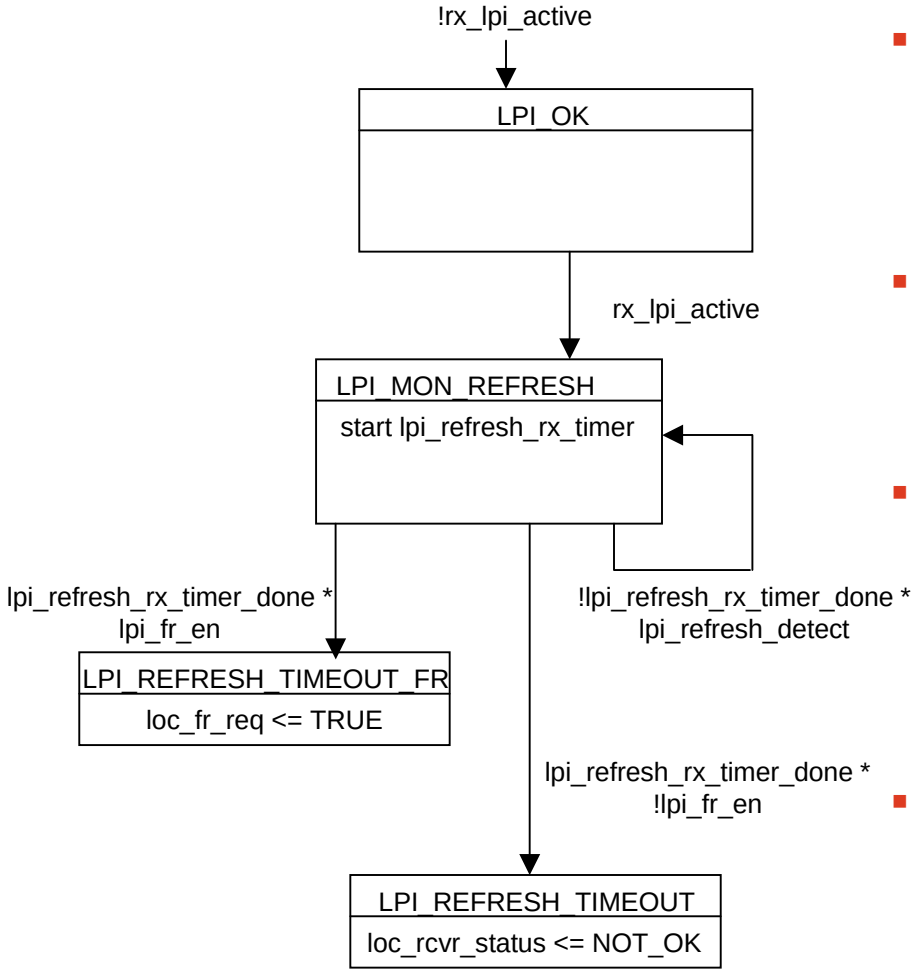
Aims

- Add a refresh monitor
  - **Avoid a dead link in rx-LPI**
- Add a fast retrain capability
  - **Improve link robustness to fast changes in link environment**
  - **Functions in LPI mode and in normal mode (must)**
  - **Interoperable**
  - **Substantially reduce link down-time from 2s+ to 30ms and lower (~ 2 orders of magnitude improvement)**
  - **Fixes a problem in the base standard**
    - THP coefficients are fixed after startup in clause 55

Capability overview

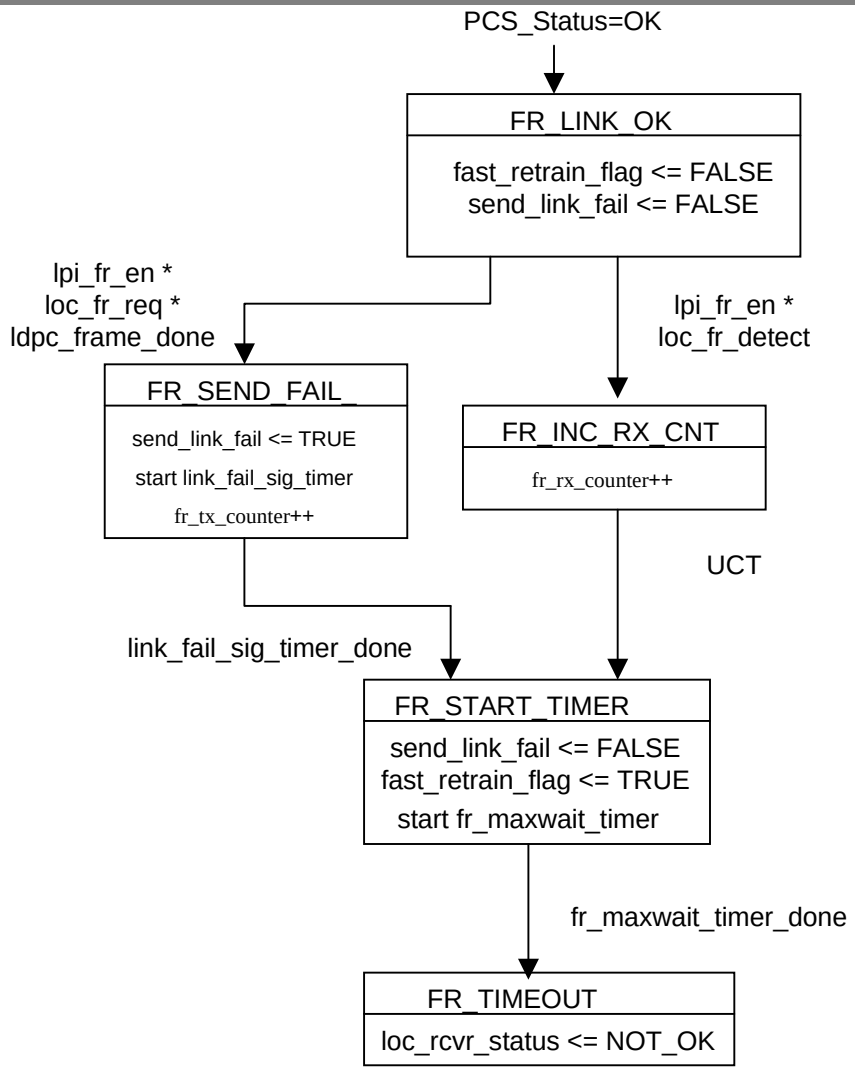
- Use autoneg bit to advertise support for fast retrain capability
  - **Both sides must have enabled fast retrain**
  - **Fast retrain can be disabled using LLDP**
- Either side can detect a link failure
  - **refresh monitor state diagram and vendor specific criteria**
    - snr degradation or missing refreshes
- Signal to the other side that the link is bad by signaling link failure, using an easily detected symbol sequence
  - **The data link goes down on both sides immediately after the alert**
- Both sides re-enter training in coefficient exchange state
  - **Use the normal startup protocol, but reduce timing allowances**
  - **Use a max_wait_timer with reduced period**
  - **Leverage existing protocol**

Add refresh monitor state diagram



- Add variable lpi_refresh_detect
  - **Set TRUE when the receiver has reliably detected a refresh signal. The exact criteria left to the implementer.**
- Add timer lpi_refresh_rx_timer
  - **Period of TBD us (330us?)**
  - **330us <-> 2 complete qr cycles**
- Add variable lpi_fr_en
  - **When TRUE the fast retrain capability is enabled. The variable is set through a management register.**
- LPI_REFRESH_TIMEOUT_FR and LPI_REFRESH_TIMEOUT both force a link retrain (fast / normal respectively)

Add fast retrain state diagram



- Add a variable **loc_fr_req**
  - when set indicates the local PHY has detected a link failure
  - set through the refresh monitor state diagram and optionally through other vendor-specific means. It causes a transition to FR_SEND_FAIL, sending the link failure signal to the link partner.
- Add a variable **loc_fr_detect**
  - set true when the link failure signal is reliably detected in the PMA
- In FR_START_TIMER both sides of the link set fast_retrain_flag<=TRUE to send the PHY control state machine back to PMA_Coef_Exch
- fr_maxwait_timer is fallback in case fast retrain fails
  - loc_rcvr_status <= NOT_OK forces full retrain
  - PCS_Status=OK sends state machine back to FR_LINK_OK

New variables

- **lpi_fr_en**
  - Set TRUE through a management register. Advertised/resolved during autoneg.
- **loc_fr_req**
  - Set TRUE when the receiver has detected a link failure condition and is requesting a fast retrain, set FALSE otherwise
- **loc_fr_detect**
  - Set TRUE when the receiver has reliably detected the link failure signal. It is highly recommended that loc_fr_detect is qualified with the reception of errored blocks at the LDPC decoder output. Set FALSE when the link failure signal is not detected.
- **send_link_fail**
  - When TRUE indicates that the PMA should send the link failure signal during the current LDPC frame period (in the same manner as the LPI alert signal)

New timers

- **link_fail_sig_timer**
  - Determines the length of time the PHY sends the link failure signal. Has a period of 4 LDPC frame periods.
- **fr_maxwait_timer**
  - Determines the period of time the PHY has to set PCS_Status = OKAY following a fast retrain before the fast retrain is aborted and a full retrain performed. Has a period of TBD ms (30?).

New counters

- **fr_tx_counter**
  - Counts number of transmit link failure signals
- **fr_rx_counter**
  - Counts number of receive link failure signals
  - Both counters need management registers

Link failure signaling – add to PMA Alert clause

- **Link retrain request signaling is generated when send_link_fail is TRUE**
  - Has priority over LPI signaling (replaces LPI alert/refresh/quiet)
- **Inverted alert sequence used as 'link failure signal**
  - Four frames of inverted (multiplied by –1) LPI alert signaling on the alert pair
    - No new hardware needed, can use existing alert generator/detector
    - Recommend that it is qualified with repeated frames of receive errors at LDPC decoder output

Fast retrain changes to Fig 55-24

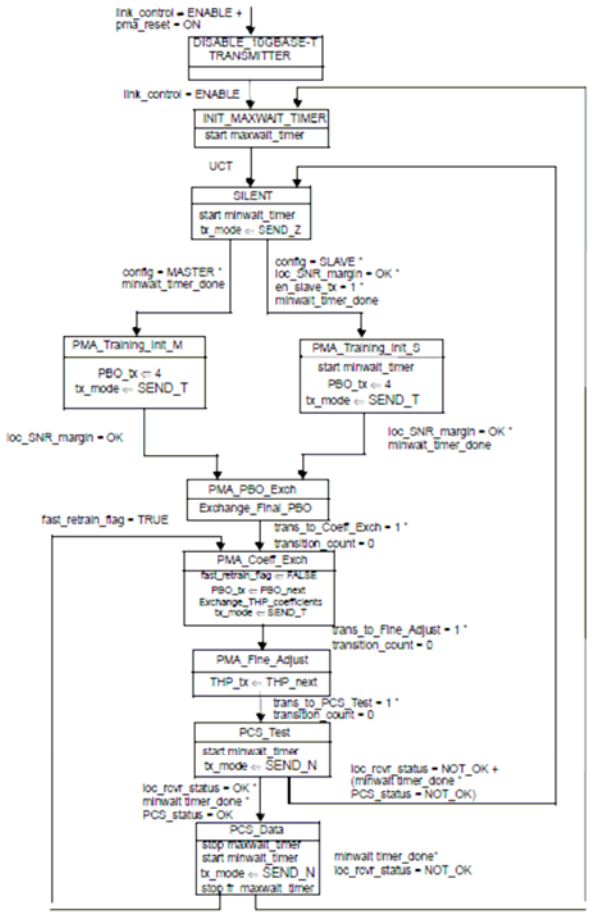
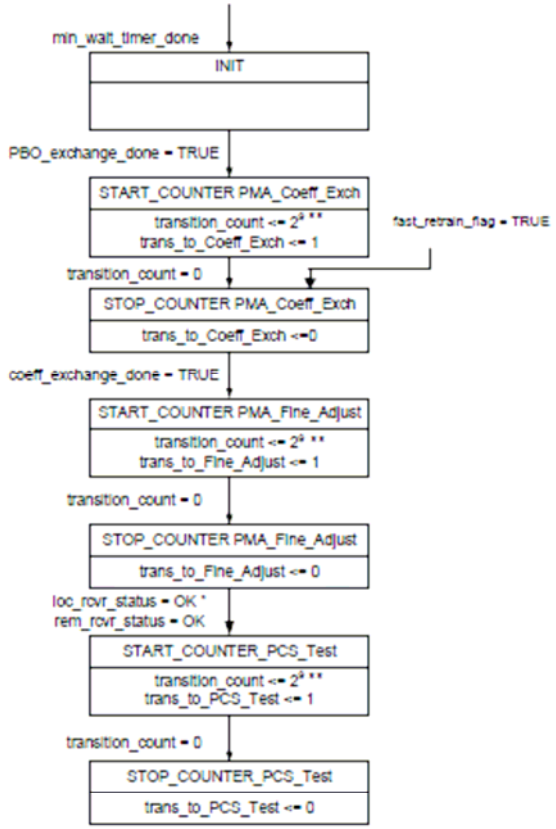


Figure 55–24—PHY Control state diagram

- Re-enter at PMA_Coeff_Exch
  - PAM2 signaling eliminates slicer errors
  - Robust training
  - Reuse existing states and transition protocols
  - Minimize new text

SOLARFLARE Communications

Figure 55-25 changes



** - devices supporting fast retrain use 2^5

Figure 55–25—MASTER transition counter state diagram

SOLARFLARE Communications

Figure 55-26 changes

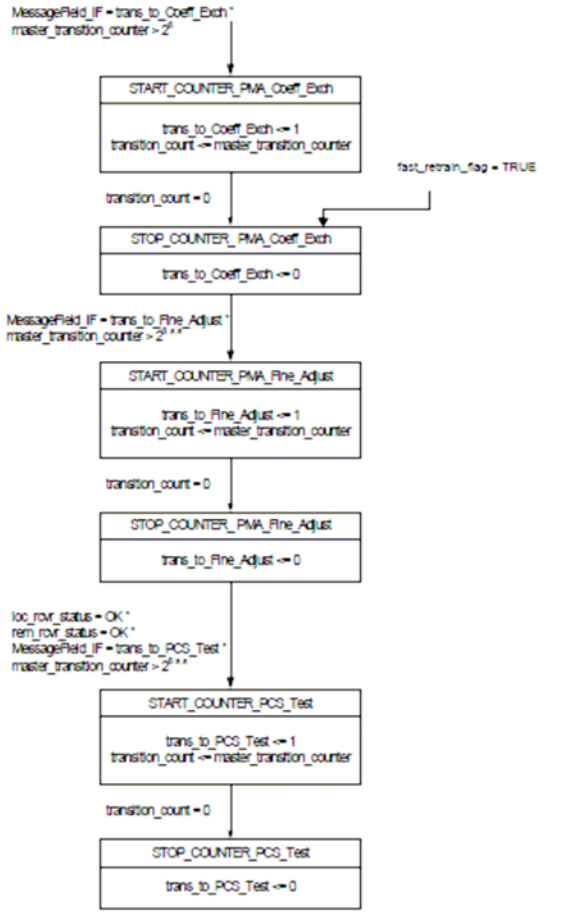


Figure 55–26—SLAVE transition counter state diagram

SOLARFLARE Communications

Existing startup timing

Table 55–6—Recommended startup sequence timing

Master	Recommended maximum time (ms)	Recommended average time (ms)	Slave
SILENT plus (PMA_Training_Init_M state AND en_slave_tx = 0)	350	315	SILENT
(PMA_Training_Init_M state AND en_slave_tx = 1) plus PMA_PBO_Exch state	480	432	PMA_Training_Init_S state plus PMA_PBO_Exch state
PMA_Coeff_Exch state	520	468	PMA_Coeff_Exch state
PMA_Fine_Adjust state	650	585	PMA_Fine_Adjust state
Total	2000	1800	

Reduce these times significantly

In fast retrain receiver is relatively converged, but needs to be fine-tuned for new environment

Shortened state timing

- Reduce infofield countdown from 10ms to <1ms
  - **Minimize transition sync. overhead**
  - **Set fr_maxwait_timer to 30ms**

State	**Recommended Maximum time (ms)**
PMA_Coeff_Exch state	20
PMA_Fine_Adjust state	10

Conclusion

- Improves link robustness and uptime
  - **Link is down for <30ms, instead of 2s+ if a normal retrain were used**
  - **Vendors can optionally provide faster retrains**
  - **Link is much more robust to changes in the link environment**
  - **Interoperable**
- No new hardware blocks required
  - **retrain request generator/detector is the same block/function as eee alert generator/detector**
  - **link training is firmware-controlled in 10GBASE-T PHYs**
  - **Adds two simple state machine and small changes to PHY control**
- EEE frame period counters reset at re-entry to PAM16
  - **no effect on existing eee protocol**

Next steps

- Add proposed mechanism to the draft
  - **Concept is well defined, details need more work**
  - **Adds two new state diagrams, small modifications to 3 PHY control state diagrams**
- Use 10GBASE-T ad hoc to refine approach
  - **Provide a complete solution for the January meeting**