# 10GBASE-T Low-Power Idle Proposal


| | |
|---|---|
| Dimitry Taich | Teranetics Inc |
| Jose Tellado | Teranetics Inc |
| George Zimmerman | Solarflare Communications |

# Supporters

- Robert Hays — Intel
- Wiren Perera — Plato Networks
- Hugh Barrass — Cisco
- Dan Dove — HP ProCurve
- Mario Träber — Infineon Technologies
- Brad Booth — AMCC
- Joseph Chou — Realtek Semiconductor Corp.

# Agenda

- LPI Mode concept
- Leveraging accepted terminology and new 10GBASE-T terms definition
- Active-to-Quiet transitioning process
- Quiet State parameters
- Quiet-to-Active transitioning process
- Wake-up time estimation and Analysis
- Estimated power saving in various power-saving scenarios
- Proposed parameters
- Summary and conclusions

# LPI Concept

- Save power by entering a LPI (Low Power Idle/State) state when there is no data to be transmitted
- While at LPI stage:
  - **All transmit and receive data path circuits can be turned off**
  - **All adaptive coefficients are saved and stored**
  - **Timing circuits free run with acquired frequency**
    - *Only fraction of the nominal power to be consumed*
  - **Periodically refresh local/remote timing so they remain locked**
  - **Periodically refresh all coefficients**
  - **PMA and PCS maintain synchronization**
    - *To enable fast return to full mode of operation*
- Merged 10GBASE-T LPI proposal (zimmerman_2_0308.pdf, see back-up slide) allows transition to Active state throughout super-frame via Alert signal
- MAC requests PHY to enter or exit LPI
  - **If the remote PHY initiates exit, local PHY immediately signals to the MAC**

# Terminology definition

| Time | Description |
|---|---|
| *Active* | Legacy operating state where data or idle are transmitted. |
| *Low Power* | New operating state used during periods of no data transmission, enabling system power reduction between data bursts. |
| *Sleep* | Frames that transmitted to inform the link partner that the local transmitter is entering the low-power state |
| *Quiet* | Transmitters are off. |
| *Refresh* | Frames that are periodically transmitted during the low-power state to allows the link partner to refresh timing and equalization |
| *Alert* | Signal that transmitted to inform the link partner that the local transmitter is returning to the active state. |
| *Awake* | Signal transmitted to inform the link partner that the local transmitter is returning to the active state. |
| *Wake* | Negotiated period when no data to be transmitted to allow receiver PHY to resume normal operation. |

# Terminology definition

- Synchronizing with hays_01_0308 and healey_01_0308.

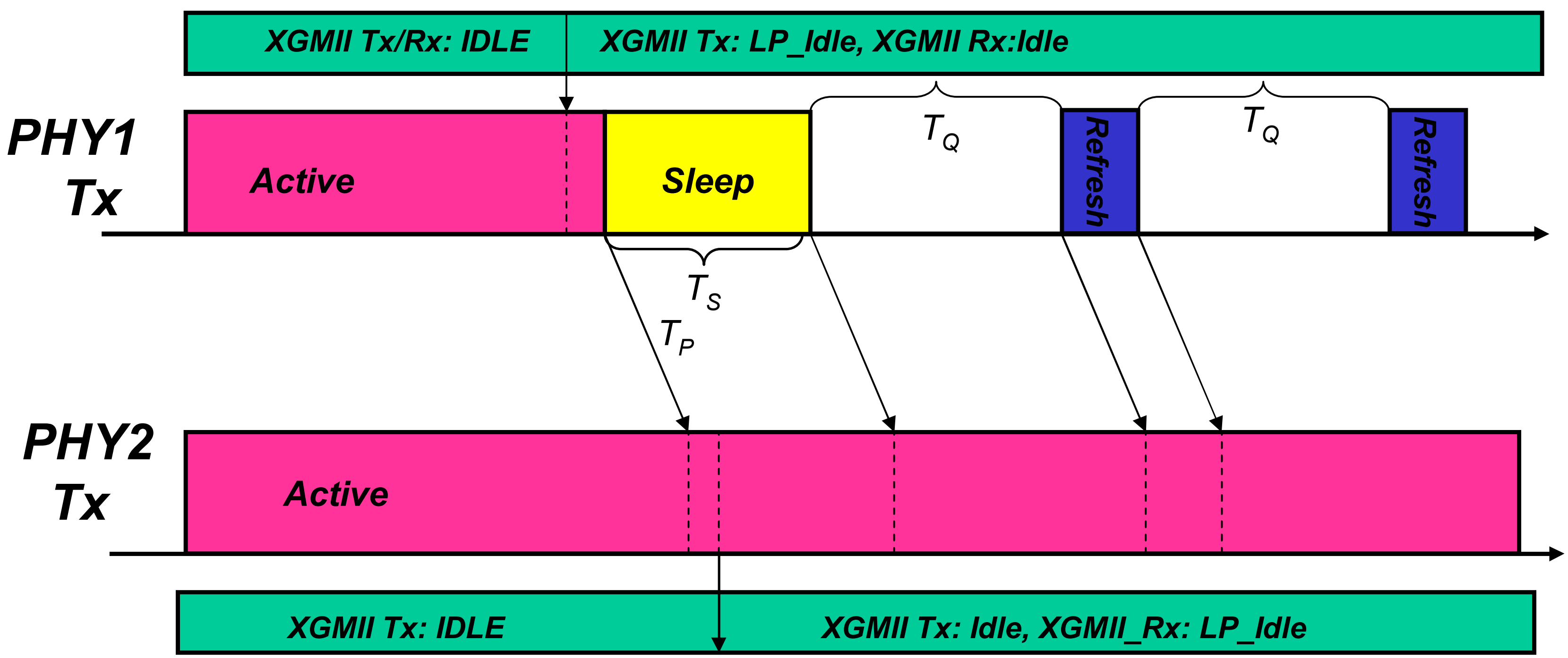
| Time | Description |
|---|---|
| $T_F$ | **Frame Time**: LDPC frame duration – 320nsec |
| $T_P$ | **Propagation delay**: Media delay. |
| $T_S$ | **Sleep Time**: Duration PHY transmits sleep signal before going quiet. |
| $T_Q$ | **Quiet duration**: Duration PHY remains quiet before sending refresh signal |
| $T_R$ | **Refresh duration**: Duration PHYs send refresh signal to enable timing and coefficient update. |
| $T_A$ | **Alert duration:** Duration PHY transmits alert signal transiting to Awake stage |
| $T_{AW}$ | **Awake time:** Duration PHY transmits Awake signal before transiting to Active stage. |
| $T_W$ | **Wake time**: Wait period where no data will be transmitted to allow the receiver to resume normal operation. |

# LPI States description

- Sleep signal:
  - To communicate entering into quiet state, TBD LDPC frames consisting of repeated XGMII control word dedicated to Sleep signaling are transmitted
- Alert signal:
  - Low-frequency pre-defined pattern – to allow simple correlation algorithm to be applied at the receiver for energy detection.
  - Initial simulations indicate that very reliable signal detect circuit is feasible at the presence of the transmitted signal and Echo/Next cancellers switched off
  - Receiver can decide (at PHY implementors' option) to switch on portion of the Echo-Next cancellers to improve detect quality with very minor effect on overall power saving
    - *Saving is dictated by $T_R / T_Q$ ratio rather then power consumption during $T_R$ stage*
- Awake signal:
  - To allow active operation resumption, TBD LDPC frames with XGMII control word dedicated to awake signaling (detailed are TBD) are transmitted

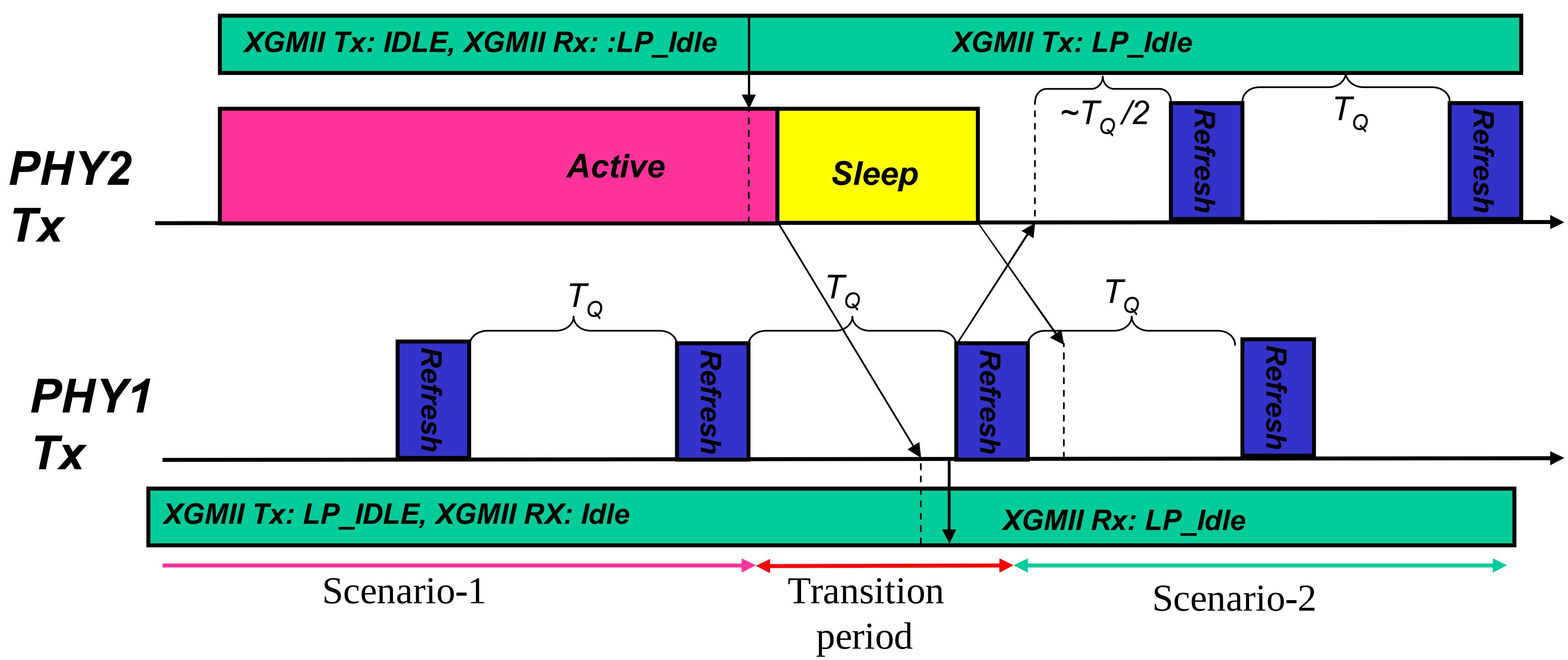
# Scenario-1 (Asymmetrical): PHY1 goes Quiet, PHY 2 remains active

- Transition to Quiet is Agnostic to Master/Slave relationship
- Sleep word (LP_Idle) on PHY1 XGMII interface is triggering transitioning into LPI mode
- PHY1 informs PHY2 about transition into Quiet state by transmitting Sleep signal
- Following transmitting Sleep signal for set # of LDPC frames, PHY1 stops transmitting. Refresh signal is transmitted periodically (each $T_Q$).
- During quiet $T_Q$ periods, PHY1 can also switch off ECN (Echo/Next Cancellers). PHY2 can switch off complete Rx data path functionality and ECN. If PHY2 is slave, PHY2 should be operated with Tx clock frozen during these periods, and update during Refresh periods
- During refresh periods $T_R$, PHY1 adjusts ECN coefficient. PHY2 uses $T_R$ periods to refresh receive filters coefficients and, if slave, update loop timing parameters.



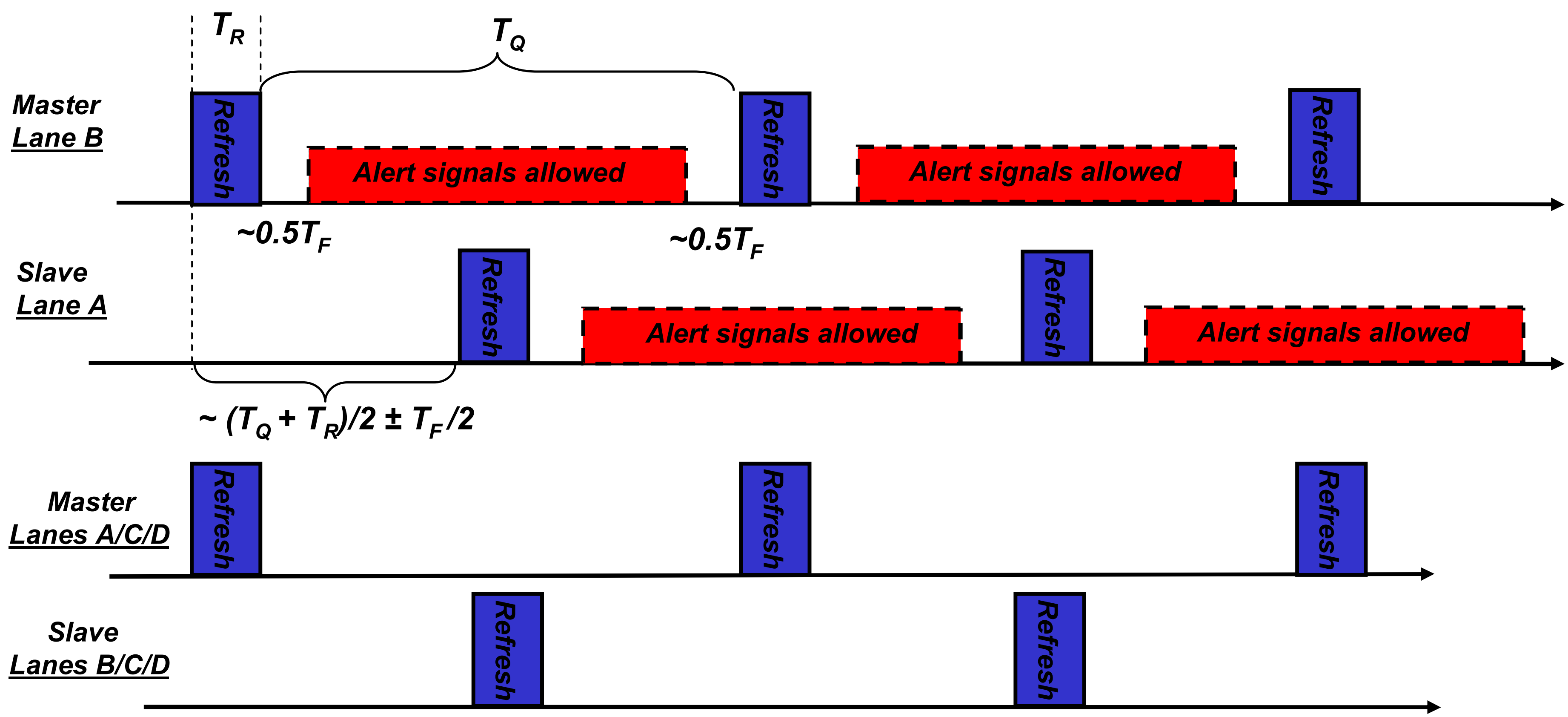
# Scenario-2 (Symmetrical): Second PHY goes quiet

- Transition is agnostic to Master/Slave relationship – start from end state of previous slide
- Sleep word (LP_Idle) onto PHY2s XGMII interface is triggering transitioning into LPI mode
- PHY2 informs PHY1 about transition into Quiet state by transmitting Sleep signal
- Following transmitting Sleep signal, PHY2 stops transmitting and going into full power saving mode. Refresh signal is transmitted periodically (each $T_Q$). First Refresh signal is aligned (with shift of $\sim T_R/2$) comparing to PHY1s Refresh signaling. See next slide for details on the refresh signals transmitting
- Following receiving Sleep signal, PHY1 may switch off receive path and going into full power saving mode. Refresh signal is transmitted periodically (each $T_Q$).



# Steady-State Transmit Diagram for Scenario 2

- Refresh signals are skewed in time with approximately 50% duty cycle
- Alert signaling:
  - **Separated in time from Refresh signal by ~ $T_F/2$**
  - **Loop-timing Master PHY will transmit on logical lane B, Slave transmits on logical lane A**
  - **Red boxes are to show when Alert signal can be transmitted.**

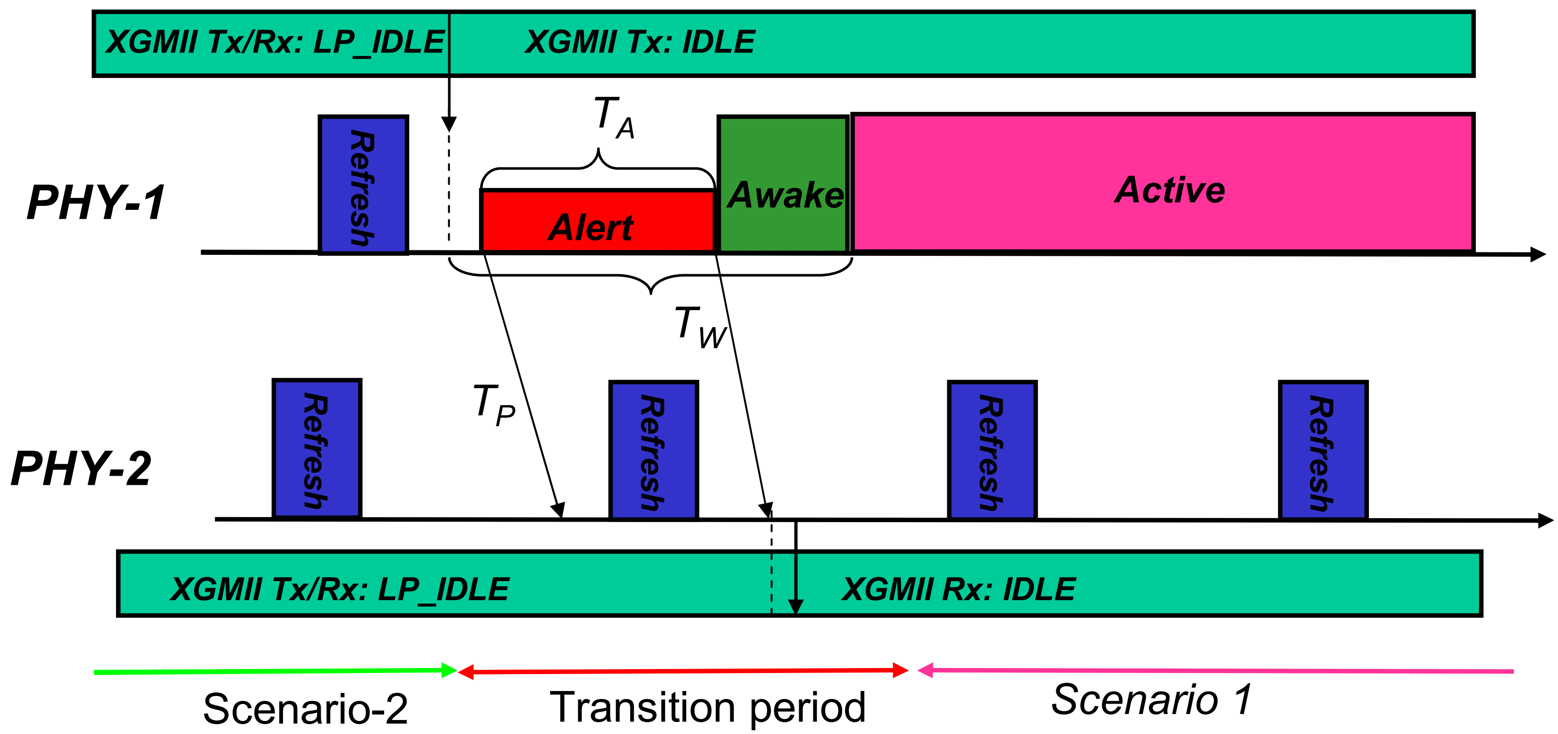


# *Quiet State parameters*

- $T_R$ and $T_Q$ *proposed values analysis:*
  - Shifting in time (by approximately $(T_Q + T_R)/2$ ) is to prevents situation when Received frames arrived in the middle of the Transmit frame
    - *Allows switching off Echo/Next cancellers and saving additional power during refresh state*
  - $(T_Q + T_R)/T_R = 25$:
    - *allows relievable timing tracking operation under extreme timing offset wandering conditions – see grimwood_01_0308.pdf*
    - Enjoys > 90% of possible power saving on LPI mode - *see taich_01_0308.pdf*
    - Higher $(T_Q + T_R)/T_R$ *are likely to be feasible but TBD at this stage*
- Alert periods are *to allow* alert signal transmission to start
  - Separated in time from Refresh signal transmission periods
  - Transmitted on *logical* lane B for Master and *logical* Lane A for Slave
    - *Lanes identification is done at AN stage*
    - *No collisions possible*
    - *Additional power saving for receive side sensing mechanism*

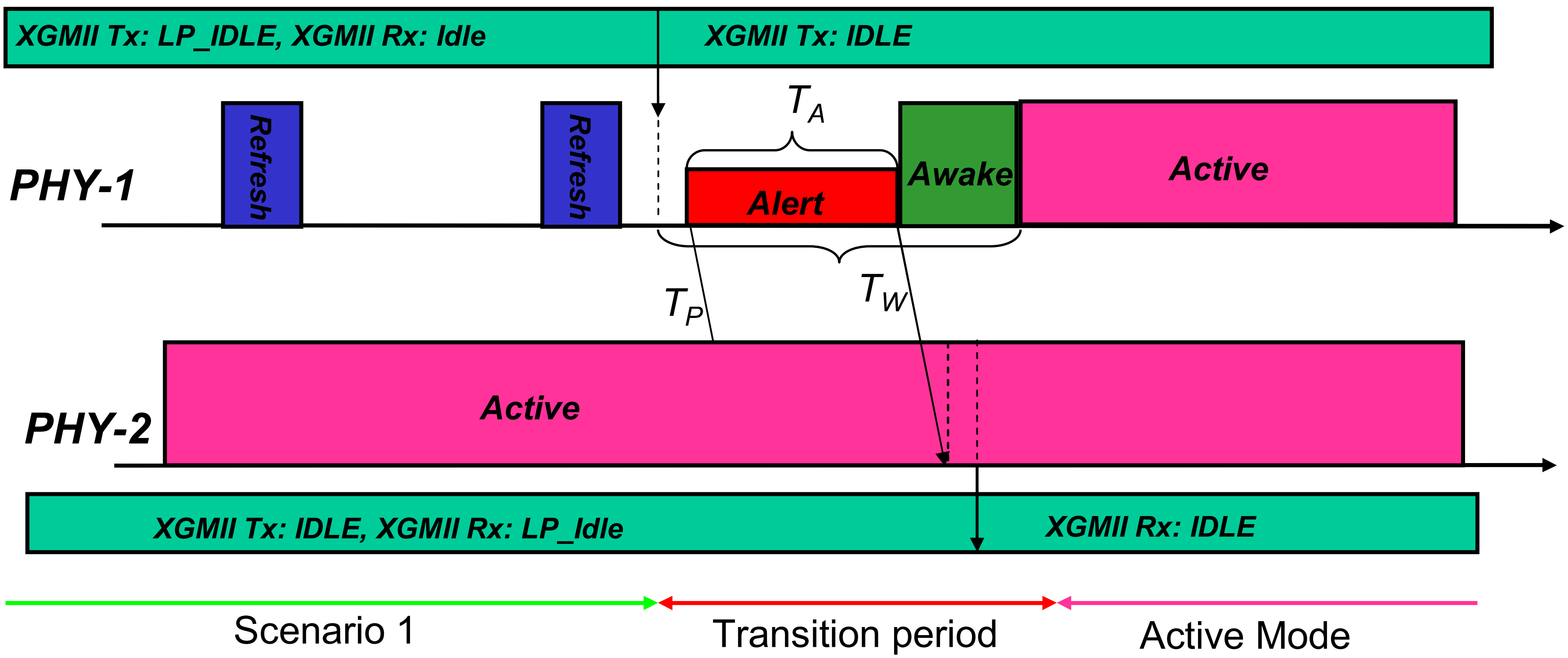
# Fully Quiet to Asymmetrical State transition Transmit Diagram

- Agnostic to Master/Slave role
- Sleep word onto PHY-1 XGMII interface is triggering transitioning into Active mode
- PHY-1 initiates transition by transmitting Alert signal during $T_A$ following by transmitting Awake signal on all 4 lanes. PHY-1 is transiting into Active state after $T_W$.
- PHY-2 should be able to detect Alert signal during Refresh signal transmission. Successful Awake signal receiving and decoding translates into wake word onto PHY-2 XGMII interface
- PHY-2 is assumed to be ready to receive data after $T_W$



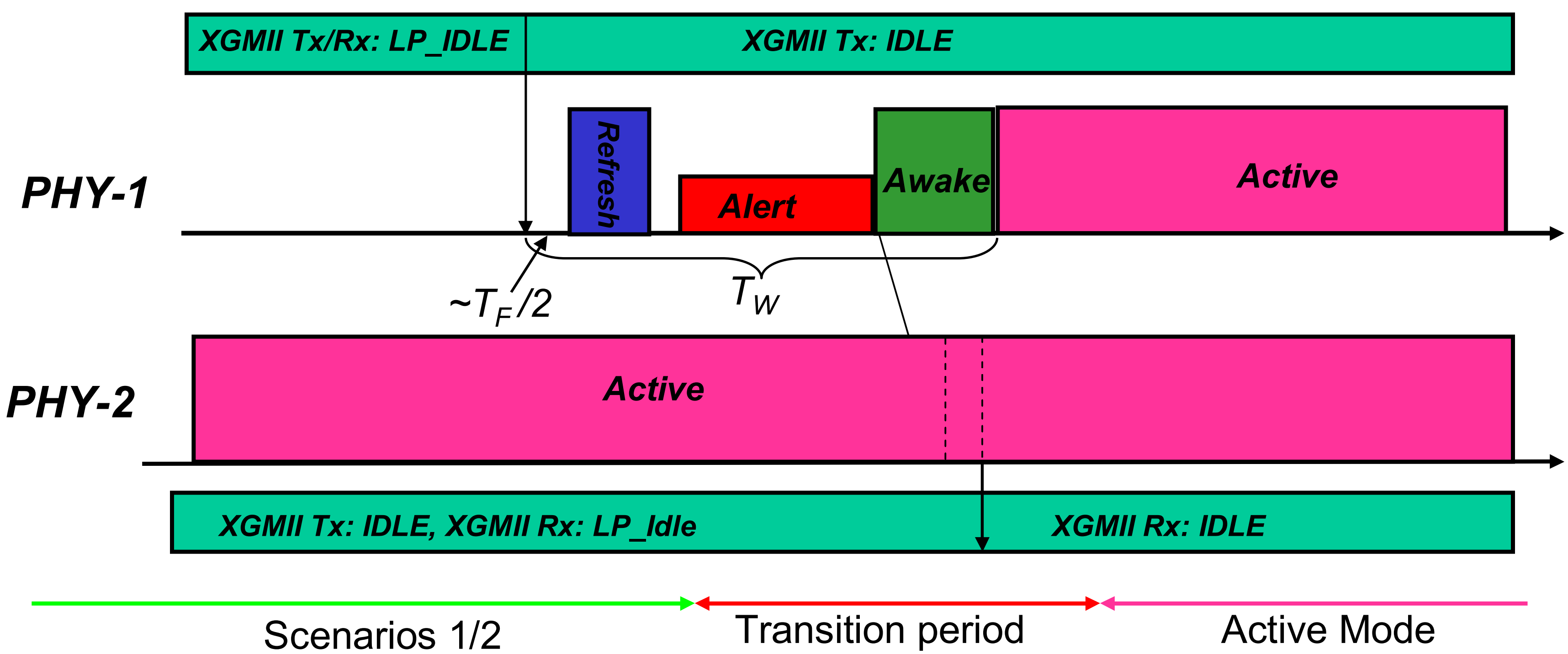
# Asymmetrical to Both Active State transition Transmit Diagram

- Agnostic to Master/Slave role
- PHY-1 initiates transition by transmitting Alert signal during $T_A$ following by transmitting Awake signal on all 4 lanes.
- PHY-1 is transiting into Active state $T_W$ period after wake word is received on XGMII interface.
- PHY-2 detect Alert signals and activates Rx data path functionality. Successful Awake signal receiving and decoding translates into wake word onto PHY-2 XGMII interface
- PHY-2 is assumed to be ready to receive data after $T_W$



# Wake-up time estimation

- If alert is initiated during Quiet time, typical wake-up time can be estimated as $\sim T_A + T_{AW} + T_F/2$
  - *For the case $T_A = 4xT_F$ and $T_{AW} = 2xT_F \rightarrow T_W \approx 2.1\mu sec$* - **less then PHY latency!**
- Corner case happens when alert should be initiated less then $T_F/2$ seconds before Refresh signal – see sketch below. In this case, worst-case wake-up time can be estimated as $T_F + T_R + T_A + T_{AW} + T_F/2$
  - *For the case $T_A = 4xT_F$ & $T_{AW} = 2xT_F \rightarrow T_W \approx 3.7\mu sec$* - **similar to PHY + $T_P$ latency!**



# Power Saving Estimation

- *Scenario-2: Both PHY's are quiet:*
  - ***In taich_0103.pdf and zimmerman_0103.pdf, power saving for the case of** $(T_Q + T_R)/T_R = 25$ **was estimated at the order of 92% of possible power saving (> 80% of absolute power saving).***
  - ***Alert sensing mechanism can be implemented with negligible complexity comparing to the rest of the active circuit***
- *Scenario-1: PHY-1 is receiving only, PHY-1 is transmitting only:*
  - ***PHY-1 has full Tx data path and ECN switched off, estimated power saving is ~40% of nominal power***
  - ***PHY-2 has full Rx data path and ECN switched off, estimated power saving is ~70% of nominal power***
  - ***Overall <u>asymmetrical operational mode provides additional ~55% power saving</u> on the PHY level for the scenarios 1 – which is most likely to dominate traffic profile!***

# Low Power Idle Parameters proposal

| State | Description |
|---|---|
| *Sleep* | Sequence of the LDPC frames consisting of repeated XGMII control word dedicated to Sleep signal. |
| *Refresh* | Sequence of the LDPC frames to allow DSP coefficients refresh and timing parameters adjustment for both Link Partners |
| *Alert* | Low-frequency signal with pre-determined structure that can be fast and reliably detected by Link Partner |
| *Awake* | Sequence of the LDPC frames consisting of repeated XGMII control word dedicated to Awake signal. |

| Time | Description |
|---|---|
| $T_P$ | Less then 550nsec |
| $T_S$ | TBD; 6x$T_F$ should be investigated |
| $T_Q$ | 100 x $T_F$; other values TBD |
| $T_R$ | 4x$T_F$; other values TBD |
| $T_A$ | TBD; 4x$T_F$ should be investigated |
| $T_{AW}$ | TBD; 2x$T_F$ should be investigated |
| $T_W$ | 4 microseconds; other values TBD |

# Conclusions

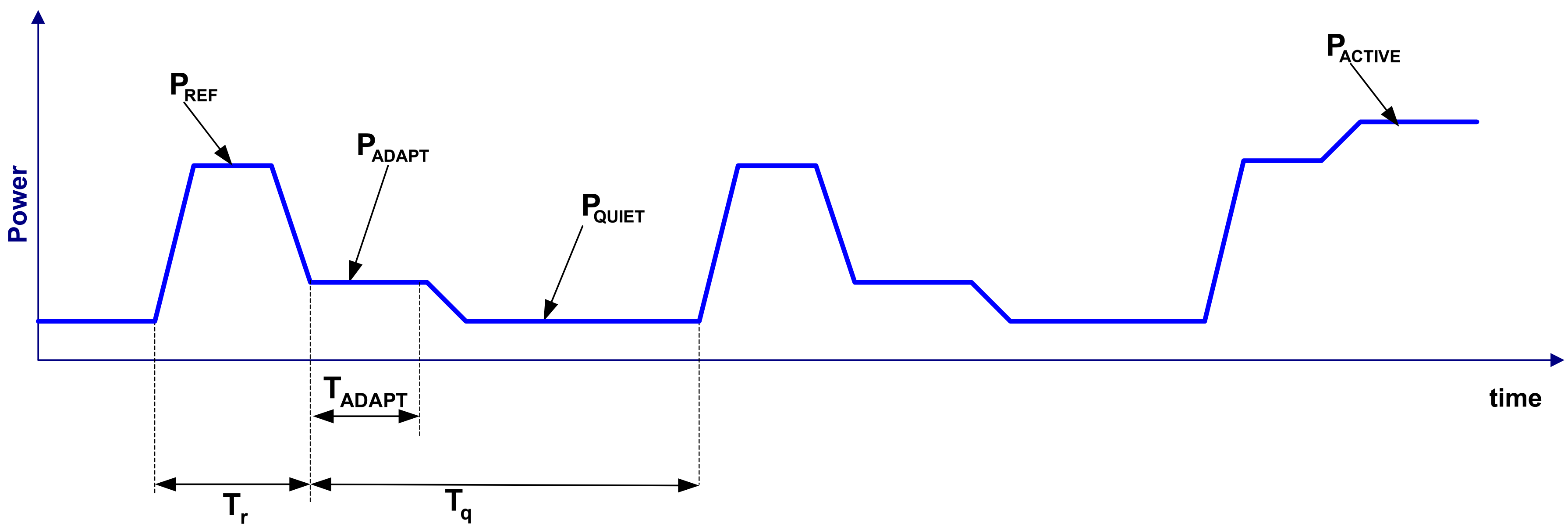
- **Overall approach is similar to 1000BASE-T LPI approach**
- **LPI method is well-suited for 10GBASE-T**
  - Wake-up time is reduced by introducing separated wake and refresh mechanisms
  - Timing tracking concern was addressed by selecting appropriate $T_Q$, $T_R$ pair
    - Allows more then 90% of max power saving
  - Longer $T_Q$ likely to be feasible – subject to further investigation
  - Backward compatible communication mechanism (using new control XGMII words) is proposed.
- **Less then 4microseconds worst-case Wake-up time is feasible**
  - Actual $T_W$ to be negotiated during AN stage
  - PHY decides on sense approach based on internal architecture

# Back-up Slides

# Power Consumption Estimation - 1

- $P_{ACTIVE}$ – full 10G Operating Power
- $P_{REF}$ - power during Refresh signal; $P_{REF} < P_{ACTIVE}$
- $P_{ADAPT}$ – additional power required to complete coefficients update after ceasing transmitting and receiving Refresh signal; $P_{ADAPT} << P_{REF}$
- $P_{QUIET}$ – Minimal Energy Mode (Tx/Rx are off, leakage + timing circuit + MDIO/etc); $P_{QUIET} << P_{REF}$, $P_{QUIET} << P_{ACTIVE}$

- The target is minimizing <u>power consumption</u> over $T_q + T_r$ period; because $P_{QUIET} << P_{REF}$ and $P_{ADAPT} << P_{REF}$ the key is achieving $T_q >> T_r$



# Power Consumption Estimation - 2

- $P_{REF} = P_{ACTIVE} - P_{LDPC} - P_{ENX} \sim 70\%$ of $P_{ACTIVE}$
  - Assuming all interfaces buses are ON
  - $P_{ENX}$ (Power consumed by Echo and Next Cancellers) cannot be completely neglected as cancellers should be switched on periodically for training purpose
  - Slightly higher than in "parnaby_01_0108"
- $P_{ADAPT}$:
  - Majority of the data path circuits can be switched off
  - Since channel is very stable coefficients update process can be spread among big number of refresh cycles thus further reducing power consumption per refresh cycle
  - Timing circuit might require frequent update but usually consumes very little power
  - $P_{ADAPT}$ can be estimated as ~20% of $P_{ACTIVE}$ (conservative estimation)
  - $T_{ADAPT}$ for most real-life scenarios is limited to 1 LDPC frame
- **Power Consumption over non-quiet period can be calculated as**

  $P_{REF} \times M + P_{ADAPT} \times M + \Delta$, where
  - $\Delta$ is overhead of power consumption due to on/off switching;
  - *M is number of the LDPC frames in one Refresh signal ($M = T_r / 0.32\mu sec$)*
  - According to our estimation, duration of the overhead period (combined for power on and power off) is less then 1 LDPC period, thus associated power consumption can be approximated by $P_{REF} \times M/2$
- $P_{QUIET}$ ~10% of $P_{ACTIVE}$
  - Slightly lower than in "parnaby_01_0108"

# Power Consumption Estimation - 3

- Thus worst-case power consumption over $T_q + T_r$ period can be approximated as

$$P_{REF} \times M + P_{REF} \times M/2 + P_{ADAPT} \times 1 + P_{QUIET} \times (N-M-2),$$

$$\text{where } N = (T_r + T_q)/0.32\mu sec$$

  and should be compared to $P_{ACTIVE}$ x N; Obviously max power saving asymptotically converges to $P_{QUIET}/P_{ACTIVE}$ (90%)
- For M << N, overall power consumption is dominated by $P_{QUIET}$
  - N/M = 10 allows better then 80% of possible power saving (72% of absolute power saving)
  - N/M = 20 allows better then 90% of possible power saving (81% of absolute power saving)
  - N/M = 100 allows better then 98%(!) of possible power saving (88% of absolute power saving); not much energy left to go after…