Energy Efficient Ethernet

IEEE P802.3az
Wait time (Tw) from a system design perspective

David Law
3Com

David_Law@3Com.com

Contributor

- Adam Healey LSI

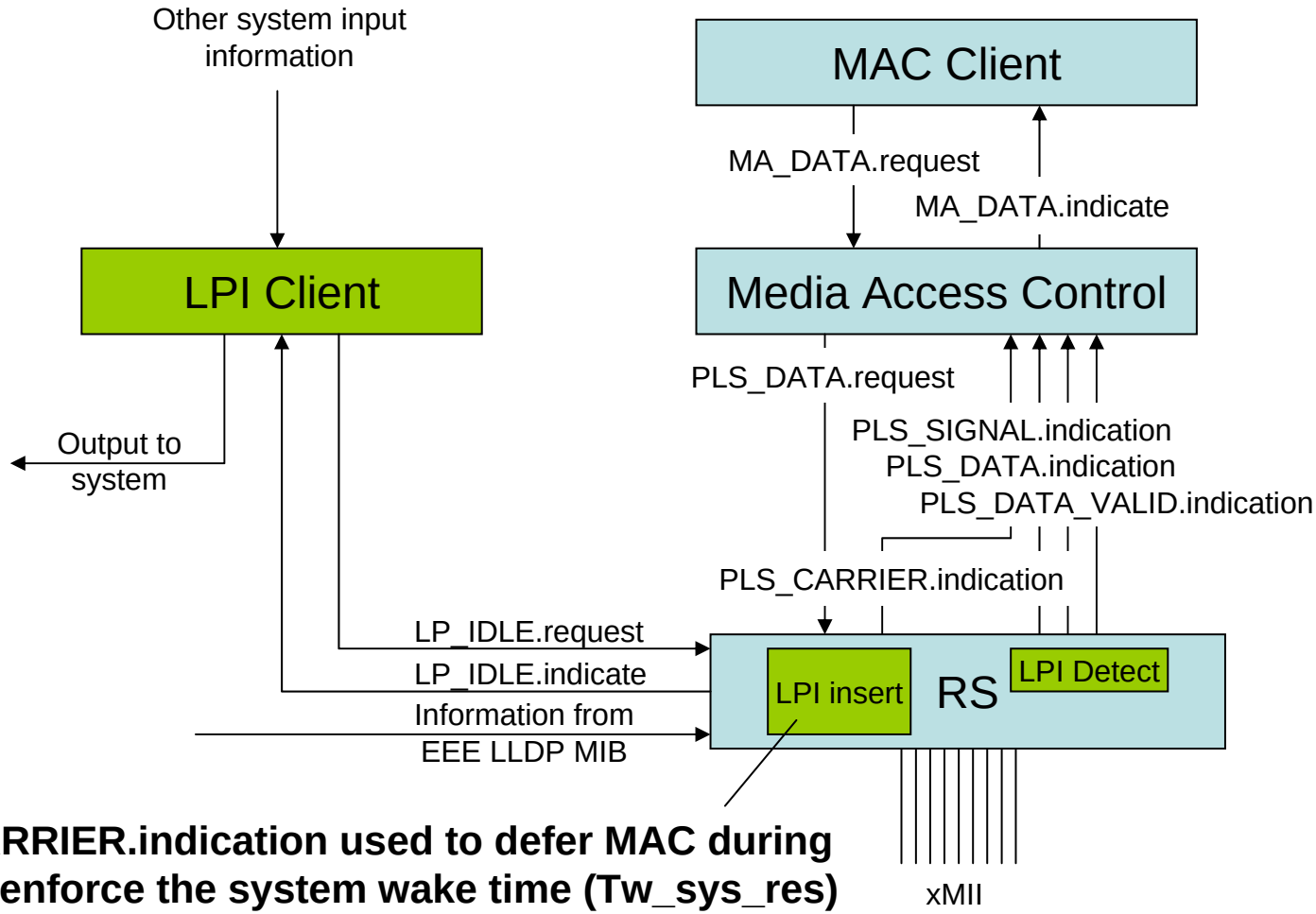
EEE wait time design information

- To design system ‘above’ xMII need
  - The longest time the TX system will ever have to wait between a request to transmit and actually being able to transmit
    - Currently called ‘Transmit Tw’ (see 78.4.2.1)
      - Will use Tw_sys_tx as shorthand in this presentation
  - The shortest time the RX system will ever be given between a request to wake and being able to receive
    - Currently called ‘Receive Tw’ (see 78.4.2.2)
      - Will use Tw_sys_rx as shorthand in this presentation

EEE wait time during operation

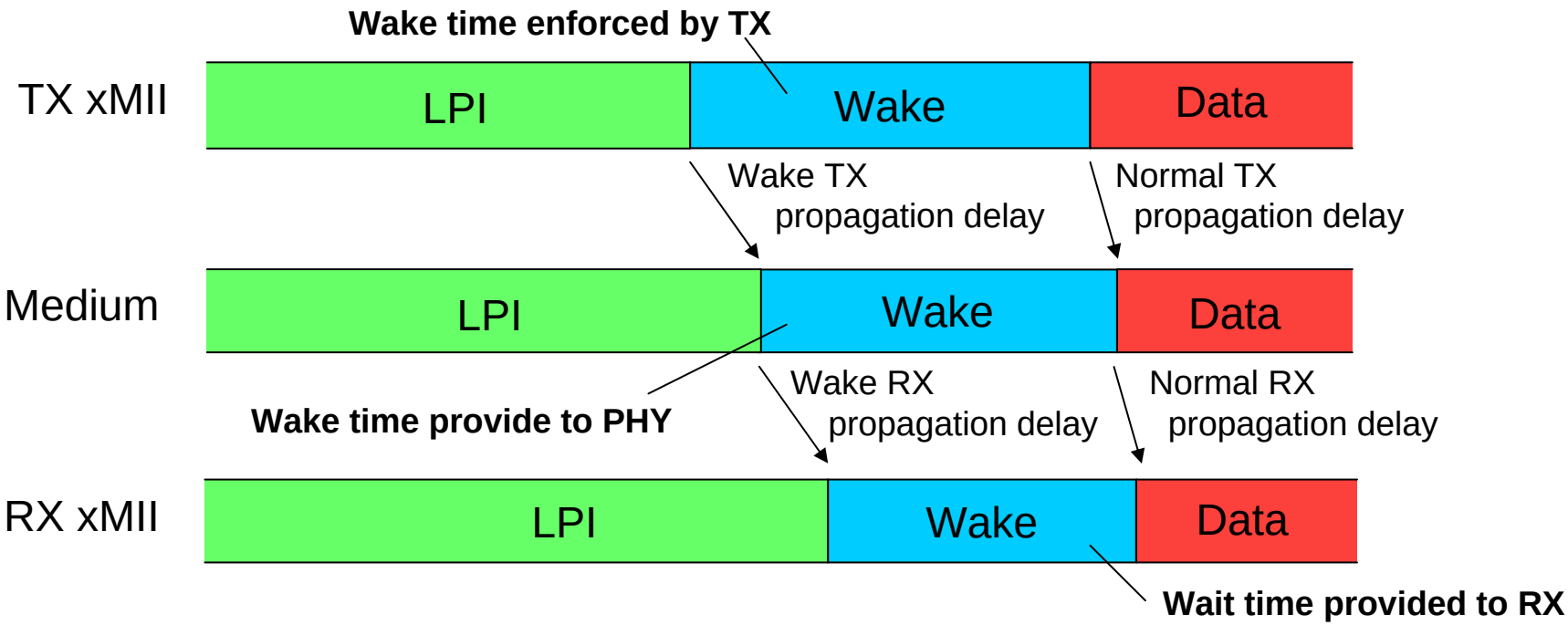
- During operation system ‘above’ xMII needs
  - The shortest time the TX system has to wait between a request to transmit and actually being able to transmit
    - Currently called ‘Resolved Transmit Tw_sys’ (see 78.4.2.3)
      - Will use Tw_sys_res as shorthand in this presentation
  - Value enforced by RS
    - Based on vote at November plenary meeting this value is enforced by using PLS_CARRIER.indication to defer the MAC during LPI and wake time (see 22.7.1)
  - Start-up value based on PHY Type
    - Some PHYs determine value during Auto-Neg (see 55.6.3)
  - May be changed though LLDP exchanges (optional)

System architecture review



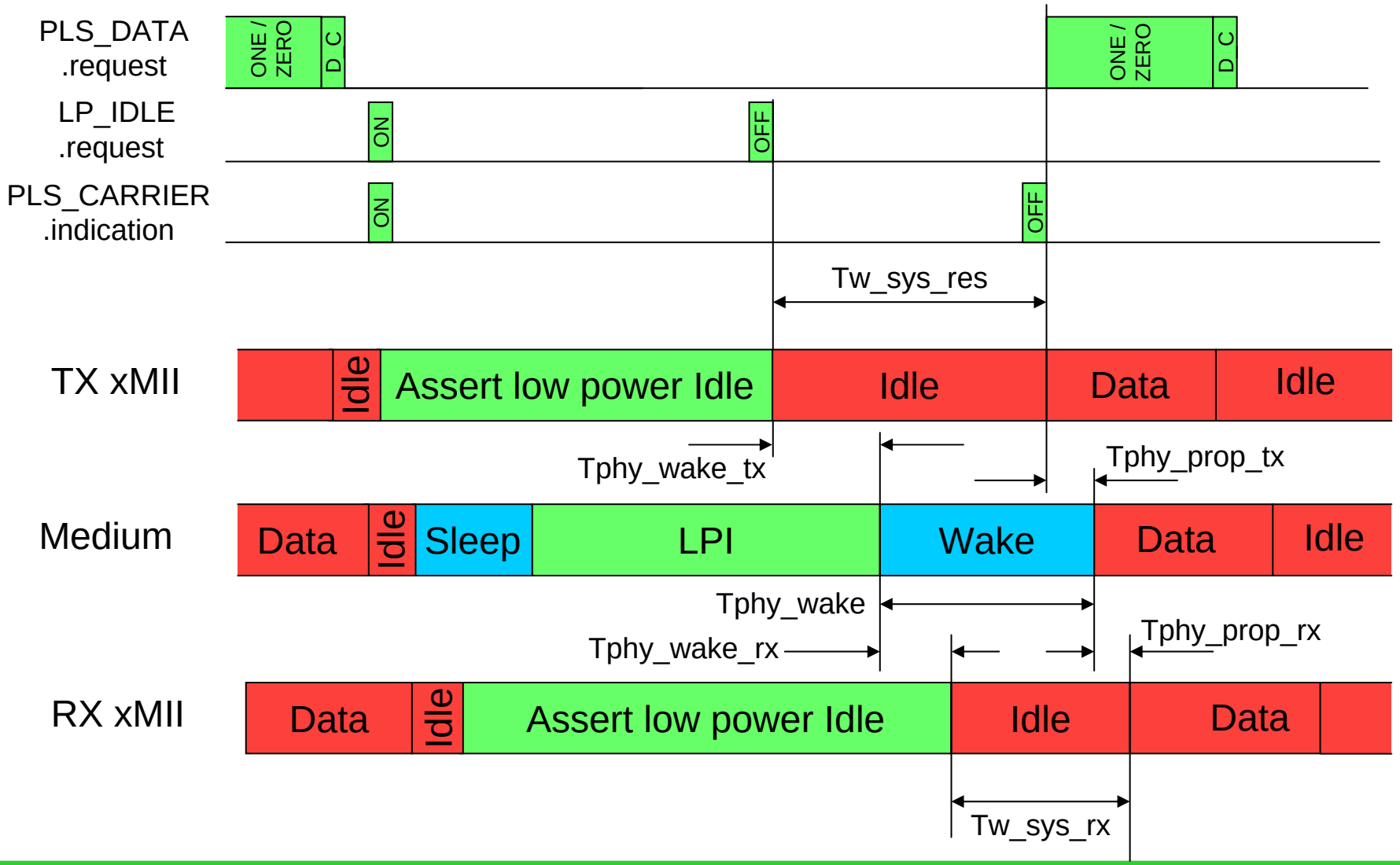
PLS_CARRIER.indication used to defer MAC during LPI and enforce the system wake time (Tw_sys_res)

Wake time propagation - overview

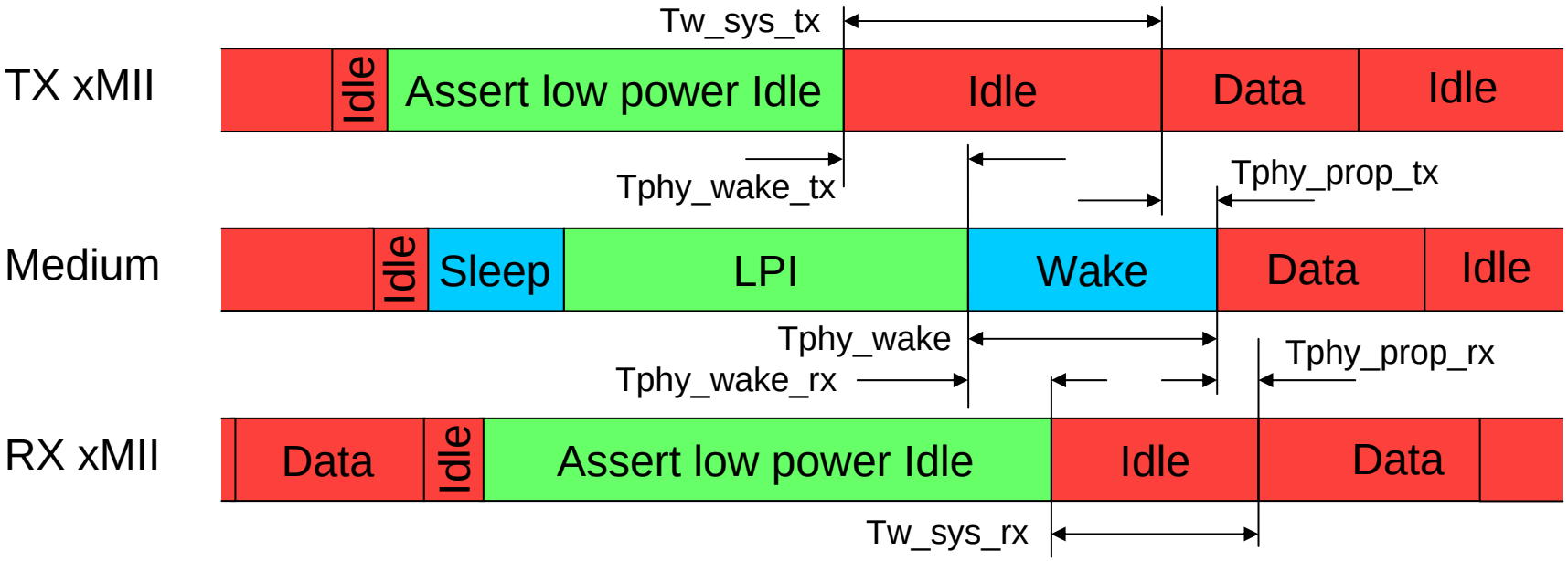


- Mismatch between wake and normal propagation delay
  - If wake propagation delay greater than wait time shrinkage
  - If wake propagation delay less than wait time growth
  - RX and PHY wake times different from TX wake time enforced by RS
- Greater of RX and PHY minimum wake times determines TX wake time
  - RX minimum wake time likely rate dependant
  - PHY minimum wake time PHY type (100BASE-T, 1000BASE-T, etc) dependant

Wake time propagation - detail



Minimum System wake time



Tw_sys_tx (min) = Tw_sys_rx (min) + Tphy_shrink_tx (max) + Tphy_shrink_rx (max)
Tw_phy (min) = Tphy_wake (min) + Tphy_shrink_rx

Tw_sys_res (min) is greater of Tw_sys_tx (min) and Tw_phy (min)

Tphy_shrink_tx (max) = (Tphy_wake_tx(max) - Tphy_prop_tx (min))
Tphy_shrink_rx (max) = (Tphy_wake_rx(max) - Tphy_prop_rx (min))

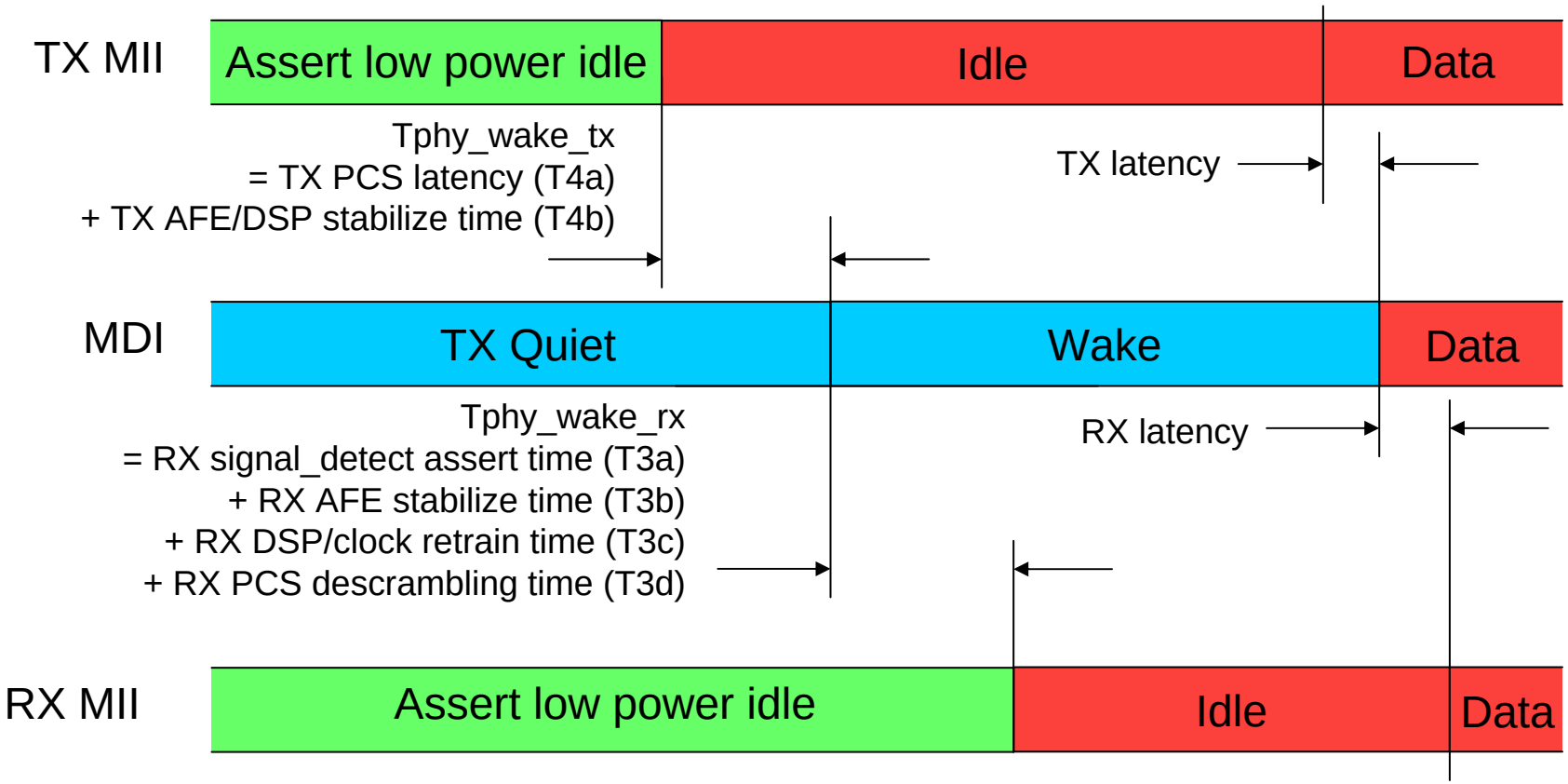
Where:

Tphy_wake_tx: xMII start of wake to MDI start of wake delay
Tphy_prop_tx: xMII to MDI data propagation delay
Tphy_wake_rx: MDI start of wake to xMII start of wake delay
Tphy_prop_rx: MDI to xMII data propagation delay
Tphy_wake: Minimum wake duration required by PHY

Requirements for interoperability

- For the system
  - Specify default minimum wake to be enforced by system TX
    - Tw_sys_res (min)
      - Derived from greater of PHY or default system RX minimum wake time
  - Specify default minimum wake time provided to system RX
    - Tw_sys_rx (min)
      - In some cases may be greater than that provided by the PHY
- For each PHY type
  - Separate normative PHY TX and RX shrinkage values required
    - A single value would not ensure interoperability as differing allocation between TX and RX by different vendors
  - Specify maximum wake shrinkage on TX
    - Tphy_shrink_tx (max)
  - Specify maximum wake shrinkage on RX
    - Tphy_shrink_rx (max)
  - Specify the minimum wake time required by PHY
    - Tw_phy (min)

Example:100BASE-T



For 100BASE-T PHY Tphy_wake (min) = Tphy_wake_rx

Reference for worst-case: Timing parameters of LPI 100BASE-T
Joesph Chou, Realtek
http://www.ieee802.org/3/az/public/jul08/chou_02_0708.pdf

100BASE-T calculations

Tphy_wake_tx = TX PCS latency (T4a) + TX AFE/DSP stabilize time (T4b)
= 0.1 + 5 = 5.1 us (max)

Tphy_shrink_tx = Tphy_wake_tx - Tphy_prop_tx
= 5.1 us - TX latency (assume 0.1us based on T4a)
= 5 us (max)

Tphy_wake_rx = RX signal_detect assert time (T3a)
+ RX AFE stabilize time (T3b)
+ RX DSP/clock retrain time (T3c)
+ RX PCS descrambling time (T3d)
= 5 + 5+ 5 + 0.5 = 15.5us (max)

Tphy_shrink_rx = Tphy_wake_rx - Tphy_prop_rx
= 15.5 us - TX latency (assume 0.5us based on T3d)
= 15 us (max)

Tw_sys_tx = Tw_sys_rx + Tphy_shrink_tx + Tphy_shrink_rx
= Tw_sys_rx + 5 + 15
= Tw_sys_rx + 20us

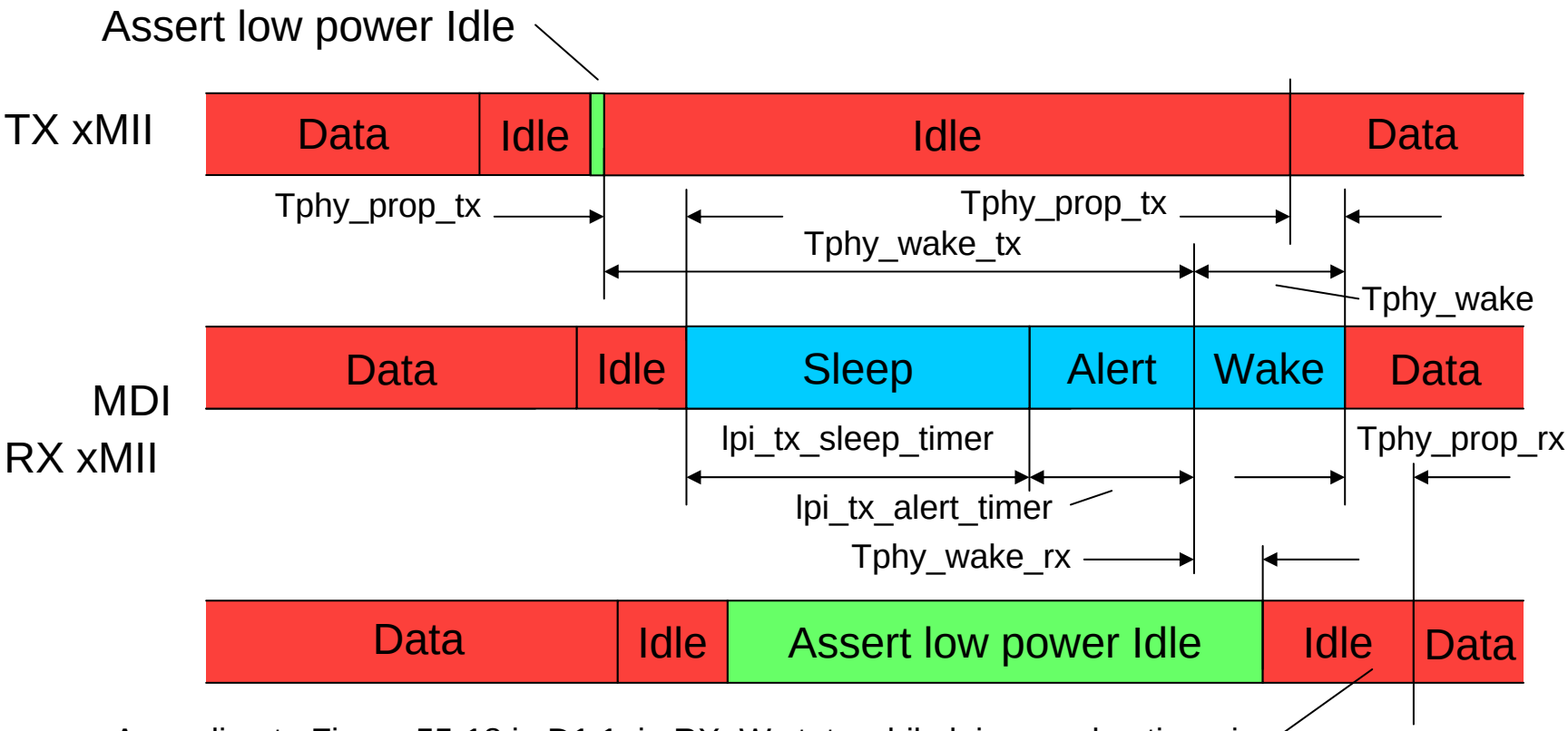
Tw_phy = Tphy_wake (min) + Tphy_shrink_tx
= 15.5 + 5 = 20.5us

100BASE-T wait time summary

PHY	Tw_sys_res default (min)	Tw_phy (min)	Tphy_shrink_tx (max)	Tphy_shrink_rx (max)	Tw_sys_rx default (min)
100BASE-T	Greater of 20.5us or TDB + 20us	20.5us	5us	15us	TBD

- Default value
  - Tw_sys_res greater of Tw_phy or (Tw_sys_rx + Tphy_shrink_tx + Tphy_shrink_rx)
    - All values from table
- LLDP rules
  - Tw_sys_res > Tw_phy
    - If not PHY will not have sufficient wake time
  - Tw_sys_res lesser of Tw_sys_tx or (Tw_sys_rx + Tphy_shrink_tx + Tphy_shrink_rx)
    - Tw_sys_rx value is from LLDP
    - Need to take account of wake shrinkage

Worst case10GBASE-T



According to Figure 55-18 in D1.1, in RX_W state while lpi_rx_wake_timer is running /LI/ is sent on XGMII. Assume this should be /I/ to wake the MAC.

Reference for worst-case: EEE P802.3az/D1.0 Clause 55 PHY Wake Time
Mike Grimwood, Broadcom
http://www.ieee802.org/3/az/public/nov08/grimwood_03_1108.pdf

10GBASE-T calculation

Tphy_wake_tx = Tphy_prop_tx + lpi_tx_sleep_timer + lpi_tx_alert_timer
= Tphy_prop_tx + 9 LDPC frames + 4 LDPC frames
= Tphy_prop_tx + 7.36us + 12.8us
= Tphy_prop_tx + 20.16 us

Tphy_shrink_tx = Tphy_wake_tx - Tphy_prop_tx
= Tphy_prop_tx + 20.16 us - Tphy_prop_tx
= 20.16us

Tphy_wake_rx = Tphy_prop_rx

Tphy_shrink_rx = Tphy_wake_rx - Tphy_prop_rx
= 0us

Tw_phy = Tphy_wake + Tphy_shrink_tx
= 2.88us + 20.16us (lpi_tx_wake_timer 9 LDPC frames max)
= 23.04us

Note – Minimum wait time provided to RX system by PHY is 1 LDPC frame as lpi_tx_wake_timer is 1 LDPC frame minimum

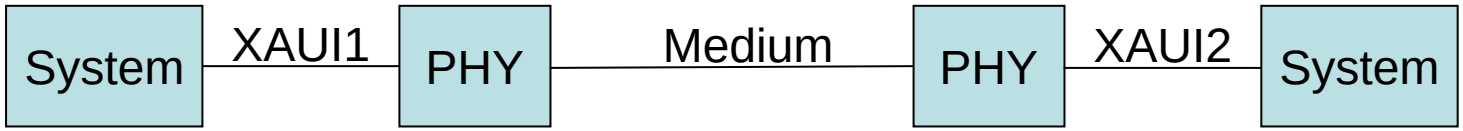
10GBASE-T wait time summary

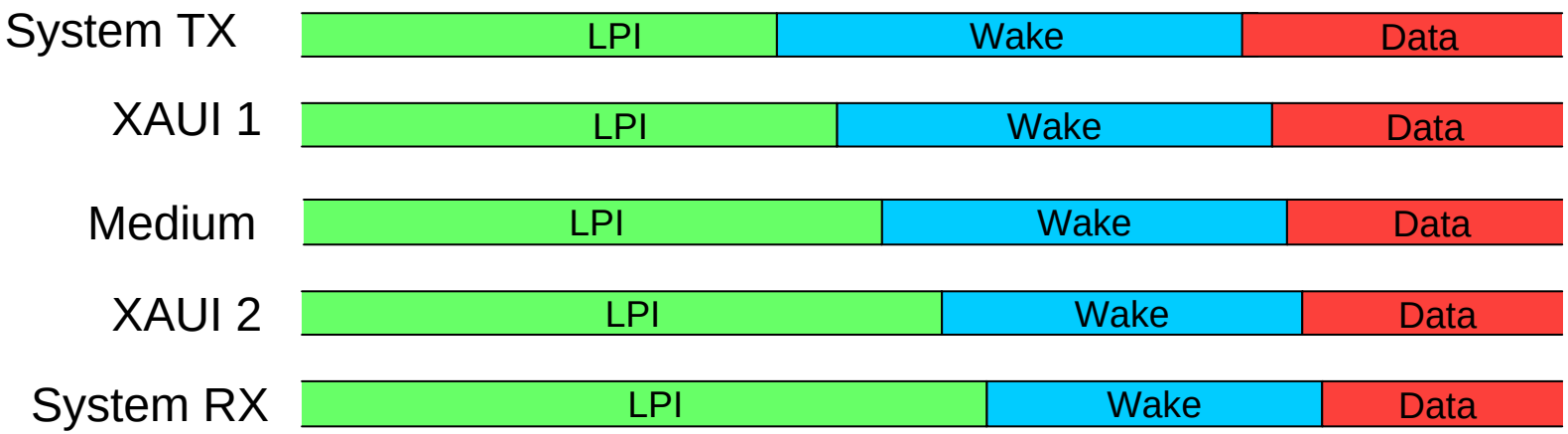
PHY	Tw_sys_res (min)	Tw_phy (min)	Tphy_shrink_tx (max)	Tphy_shrink_rx (max)	Tw_sys_rx (min)
10GBASE-T	Greater of 23.04us or TDB + 20.16us	23.04us	20.16us	0us	TBD

- Large TX wait shrinkage but no RX wait shrinkage
  - Worst case occurs when the PHY hasn't entered Quiet
- Could the RS be used to compensate for this
  - Enforce a minimum LPI assert time in RS for 10GBASE-T
    - Once this is over the required shrinkage time is likely smaller
    - No need to compensate for large TX wait shrinkage once in Quiet
    - Other 10Gb/s PHYs can set minimum time to zero

Possible interface EEE modes

- Should we make provision in shrinkage time for interfaces ?
  - If provision isn't made now in may prevent future optimal EEE support
  - Examples of interfaces that could provide EEE support
    - XAUI, SGMII (although not specified in IEEE Std 802.3)





Conclusions

- Need to calculate values as described for all PHYs
  - Tw_sys_res default (min), Tw_phy (min), Tphy_shrink_tx (max), Tphy_shrink_rx (max), Tw_sys_rx default (min)
    - Seem to be missing minimum system wake times
  - Decide if margin is provided for interface EEE modes in future
- Suggest forming an adhoc to complete this work
- Suggest the following renaming for clarity
  - Transmit Tw -> Tw_sys_tx
  - Receive Tw -> Tw_sys_rx
  - Resolved Transmit Tw_sys -> Tw_sys
- Need to adjust LLDP values to accommodate shrinkage
- 10GBASE-T
  - While lpi_rx_wake_timer is running send /I/ to wake the MAC
  - Consider minimum LPI enforcement in RS to optimise wake times