What should AVB do about Energy Efficient Ethernet?

John Nels Fuller

Computer Scientist

jfuller@computer.org

Quick Overview of EEE

- The following six slides are from Mike Bennet, chair of the EEE task group and are used by permission.
- Contact mjbennet@lbl.gov

What is Energy Efficient Ethernet (EEE)?

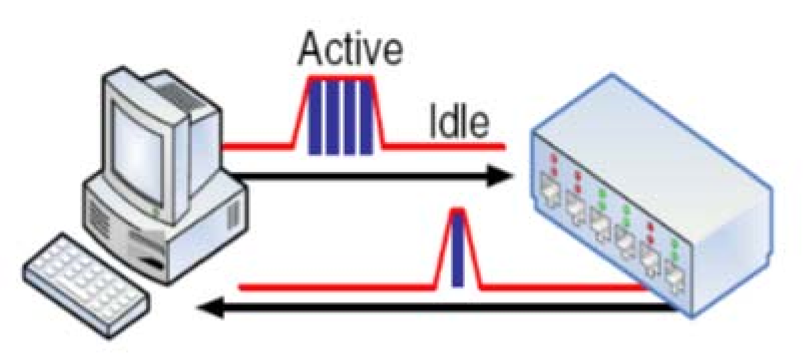
- **Also known as IEEE 802.3az**

- **EEE is a method to facilitate transition to and from Low Power Idle (LPI) mode in response to changes in traffic levels**
    - **In the process of being specified for these copper PHYs**
        - **100BASE-TX (Full Duplex)**
        - **1000BASE-T (Full Duplex)**
        - **10GBASE-T**
        - **10GBASE-KR**
        - **10GBASE-KX4**
        - **1000BASE-KX**

- **Many links have very low utilization most of the time**

What is Low Power Idle (LPI)?

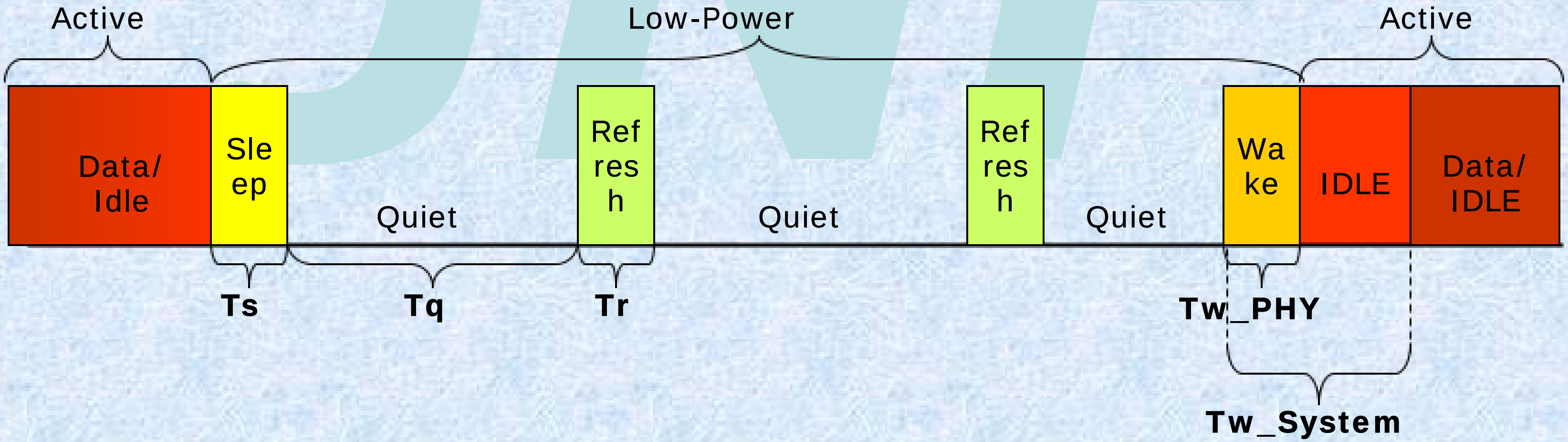
- **LPI is the state of having non-essential PHY circuits turned off when there is no data to send**
  - **Concept: Transmit data as fast as possible, return to Low-Power Idle**
  - **Energy use scales with bandwidth utilization**



What is Low Power Idle?

- **A closer look**

Term	Description
Sleep Time (Ts)	**Duration PHY sends Sleep symbols before going Quiet.**
Quiet Duration (Tq)	**Duration PHY remains Quiet before it must wake for Refresh period.**
Refresh Duration (Tr)	**Duration PHY sends Refresh symbols for timing recovery and coefficient synchronization.**
PHY Wake Time (Tw_PHY)	**Duration PHY takes to resume to Active state after decision to Wake.**
System Wake Time (Tw_System)	**Wait period where no data is transmitted to give the receiving system time to wake up.**



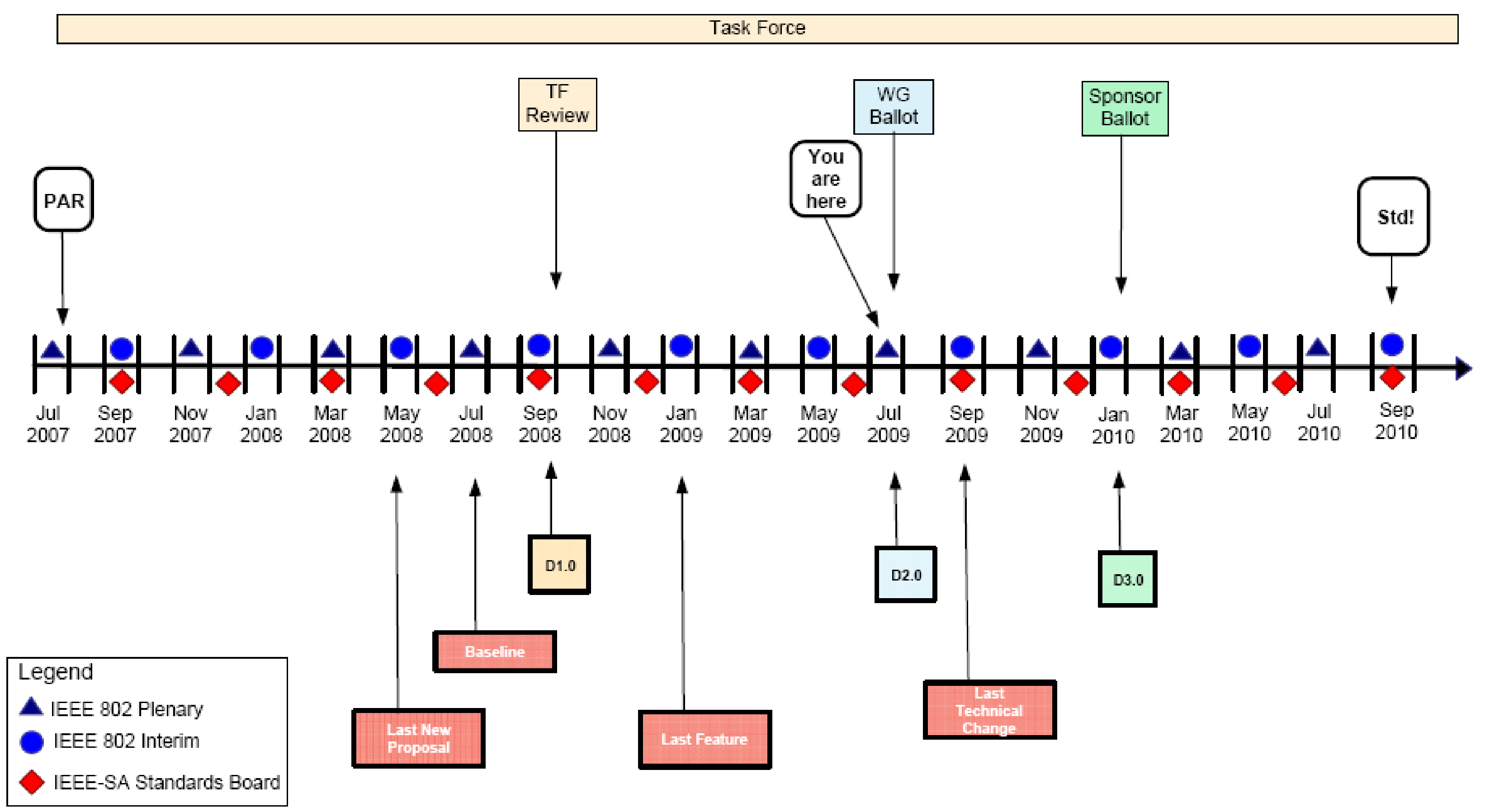
Optimizing Energy Efficiency

- **Energy Efficiency can be optimized by using link-partner communications after the link is established**
  - **Use Link Layer Discovery Protocol (LLDP) to change wake times.**
  - **The longer the wake time, the longer the delay till frames can pass, i.e. latency increases**
  - **Trade-off between energy savings and latency**

- **There are opportunities to save energy in the system in addition to PHY energy savings**

State of the standard

- **Hoping to go to 802.3 Working Group Ballot at the end of the week**
- **If we stay on track, should be finished Sept. 2010**



Task Force

PAR

TF Review

You are here

WG Ballot

Sponsor Ballot

Std!

Jul 2007
Sep 2007
Nov 2007
Jan 2008
Mar 2008
May 2008
Jul 2008
Sep 2008
Nov 2008
Jan 2009
Mar 2009
May 2009
Jul 2009
Sep 2009
Nov 2009
Jan 2010
Mar 2010
May 2010
Jul 2010
Sep 2010

D1.0

D2.0

D3.0

Baseline

Last New Proposal

Last Feature

Last Technical Change

Legend
▲ IEEE 802 Plenary
● IEEE 802 Interim
◆ IEEE-SA Standards Board

Final thoughts …

- **The 802.3az Task Force estimated 75% of PHY power savings possible using Low Power Idle**
  - **Assuming 100% adoption in the US alone that translates to roughly $300M to $470M per year in savings**
    - **Does not include cooling or additional system power savings**
- **Energy Star is planning to reference IEEE 802.3az**
  - **As soon as it is reasonable to do so**
- **More work to do?**
  - **Energy Efficient Ethernet is not specified for optical PHYs and some copper PHYs**
  - **Should there be a higher layer power management specification?**

What does AVB need to do?

- EEE expects 802.1 to define the LPI client
  - When to assert / deassert LPI
- EEE expects 802.1 to define the LLDP negotiation of additional wait time using their TLV
  - More of device can be in low power with longer wait time
- We need to describe when transmission selection is done in relation to LPI
  - Avoid committing to a best effort frame while waiting for LPI to be exited.
- 802.1BA appears to be the proper place to address these issues

Transmission Selection

- PLS_CARRIER.indication(CARRIER_ON) indicates transmitter is not ready (during LPI and for Tw_sys after deasserting LPI)
- Transmission Selection algorithm must not select a frame for transmit while CARRIER_ON is indicated.
  - Avoids adding Tw_sys to transmit time of a selected frame before a higher priority frame can be transmitted.
  - Still must determine if one or more frames are ready for transmission as an input to LPI Client (described later).

LLDP Negotiation

- Probably shouldn't fully specify this as there are too many implementation choices. Should just define constraints:
  - If idle_slope(s) for port are non-zero then value of transmit Tw_sys must be less than transmission time of maximum length frame at Fast Ethernet speed. Could specify lower limits at higher speeds, but probably don't need to do so.
  - If idle_slope(s) for port are zero then no restriction imposed by AVB
  - When idle_slope(s) go(es) from zero to non-zero and Tw_sys does not already meet the above constraint then must renegotiate before asserting LPI.

LPI Client

- Uses LP_IDLE.request and LP_IDLE.indication service primitives.
- LP_IDLE.request used in transmit direction, parameter is either ASSERT or DEASSERT
- LP_IDLE.indication used in receive direction, parameter is either ASSERT or DEASSERT indicating the state of LPI received from link partner

LPI Client – Receive Direction

- When LP_IDLE.indication(ASSERT) is received:
  - Depending on negotiated value of Tw_sys, additional components may be powered down and/or upper layers may be passed the indication so that they may power down components
- When LP_IDLE.indication(DEASSERT) is received:
  - Any powered down components should be powered up and/or upper layers may be passed the indication so that they may power up components

LPI Client – Transmit Direction

- Each 802.1BA profile needs to specify when to use LP_IDLE.request(ASSERT) and LP_IDLE.request(DEASSERT)
- For Residential profile, propose:
  - ASSERT when transmission selection finds no frame ready to transmit
  - DEASSERT when transmission selection finds at lease one frame ready to transmit

References

- 8023az-D1-5.pdf
  - On 802.3 EEE website, get password from 802.3 chair
- Ethernet AVB Technology Assessment Report
  - This document was generated for Lawrence Berkeley National Laboratory and is posted to the 802.1 website with permission: avb-fuller-ethernet-technology-assessment-0709-v01.pdf

Backup Slides

Important Timing Parameters

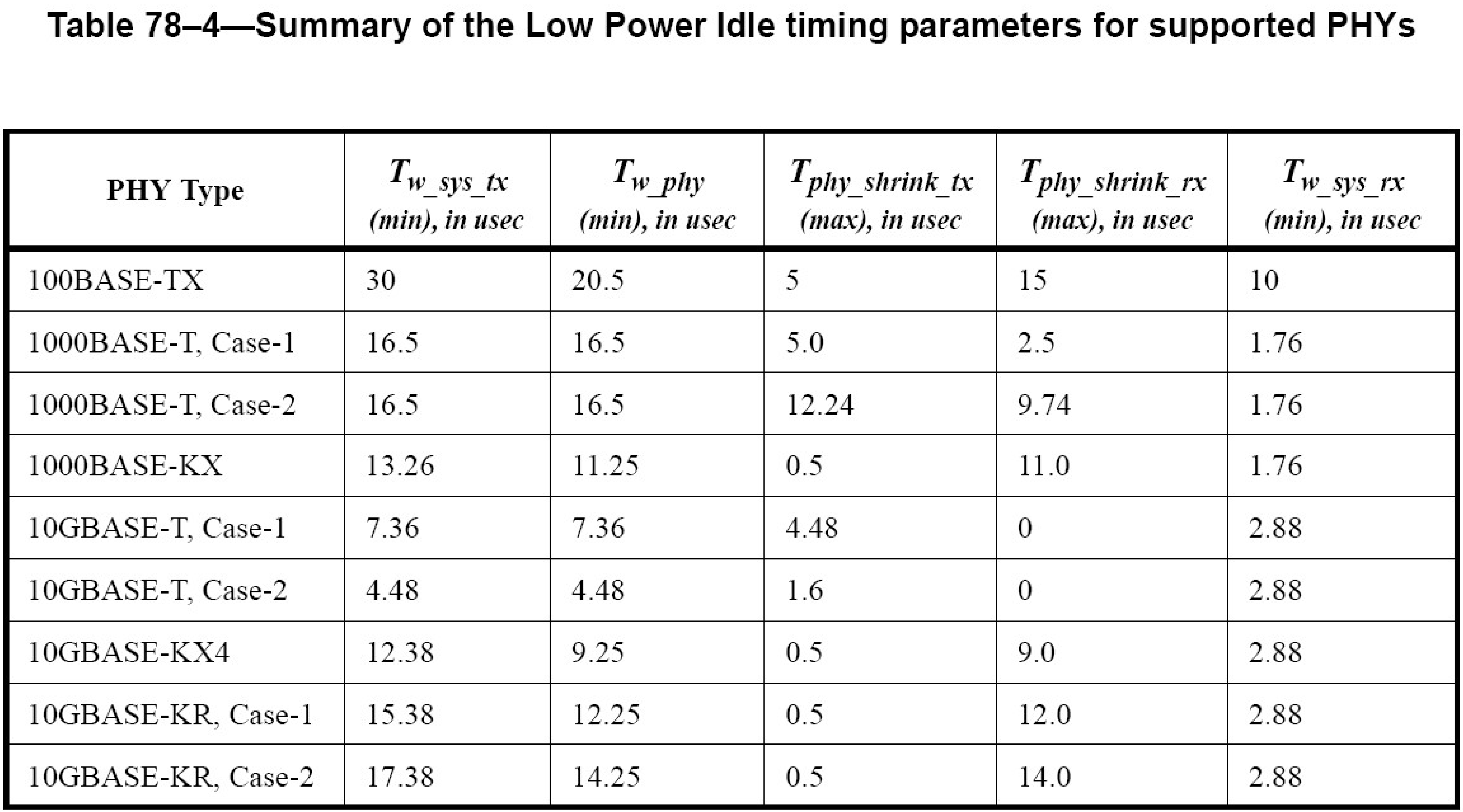


Table 78–4—Summary of the Low Power Idle timing parameters for supported PHYs

PHY Type	$T_{w\_sys\_tx}$ (min), in usec	T_{w_phy} (min), in usec	$T_{phy\_shrink\_tx}$ (max), in usec	$T_{phy\_shrink\_rx}$ (max), in usec	$T_{w\_sys\_rx}$ (min), in usec
100BASE-TX	30	20.5	5	15	10
1000BASE-T, Case-1	16.5	16.5	5.0	2.5	1.76
1000BASE-T, Case-2	16.5	16.5	12.24	9.74	1.76
1000BASE-KX	13.26	11.25	0.5	11.0	1.76
10GBASE-T, Case-1	7.36	7.36	4.48	0	2.88
10GBASE-T, Case-2	4.48	4.48	1.6	0	2.88
10GBASE-KX4	12.38	9.25	0.5	9.0	2.88
10GBASE-KR, Case-1	15.38	12.25	0.5	12.0	2.88
10GBASE-KR, Case-2	17.38	14.25	0.5	14.0	2.88