IEEE 802.3 40G & 100G NGOPTX
July 2012 Plenary Series

LLDP for EEE: Tutorial from 802.3az and potential uses in Next Generation 40Gb/s and 100Gb/s Optical Ethernet

Barrass, Hugh – Cisco
Bennett, Mike – LBNL
Booth, Brad – Dell
Carlson, Steve – HSD
Diab, Wael William – Broadcom
Dove, Dan – APM
Hajduczenia, Marek – ZTE
Law, David – HP
Vetteth, Anoop – Cisco

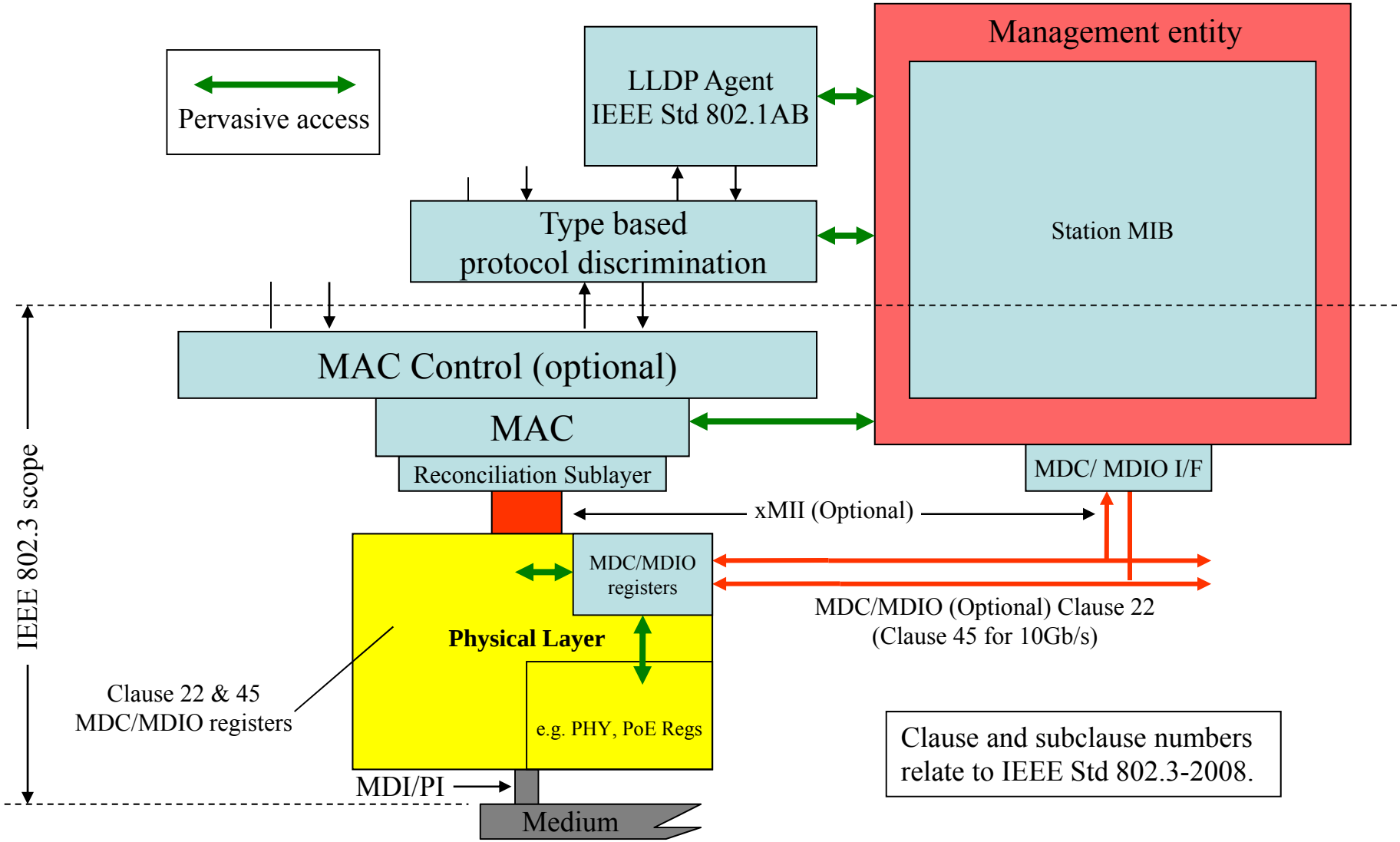
…Other Supporters Welcome…

Background and Overview

- LLDP and layer 2 enhancements are currently defined for IEEE Std 802.3az-2010 and IEEE Std 802.3at-2009 for EEE and PoE
- Purpose of presentation
  - Review how LLDP works
    - What it can and cannot do
  - Present example of how LLDP an layer 2 (DLL) are used in IEEE Std 802.3 (EEE and PoEP)
  - Suggest a mechanism to take advantage of this for EEE capability exchange that would eliminate the need for an auto-negotiation capability

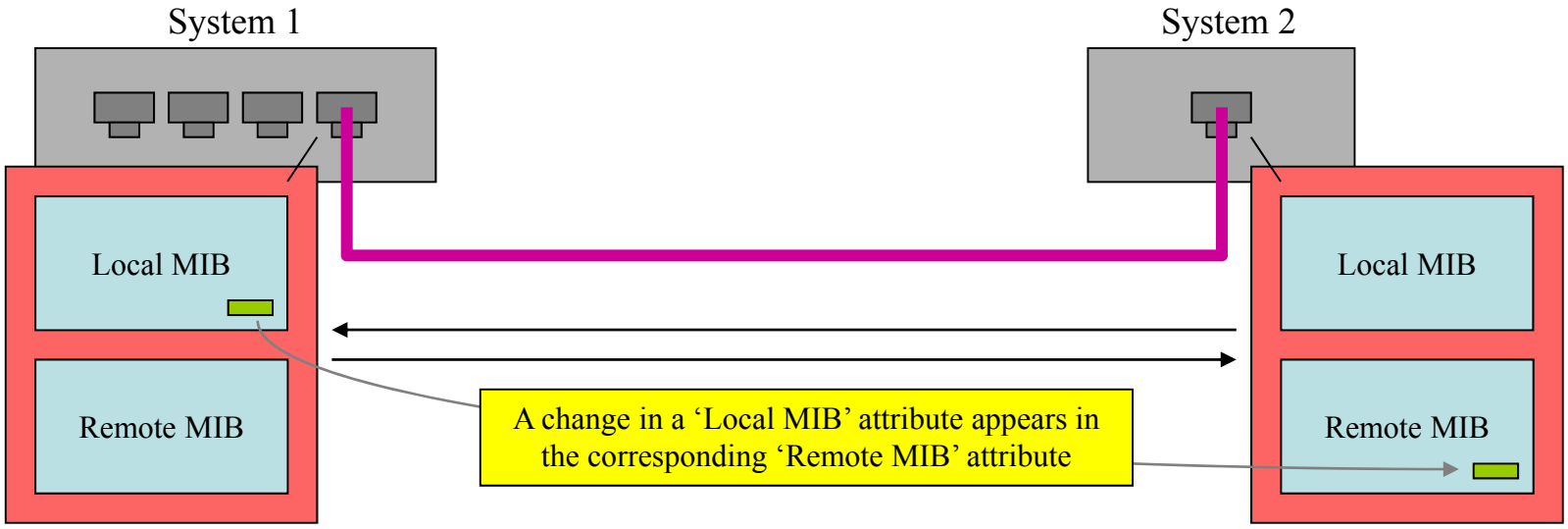
LLDP Overview

Mgmt Access (LLDP vs. MDC/MDIO)



LLDP Overview

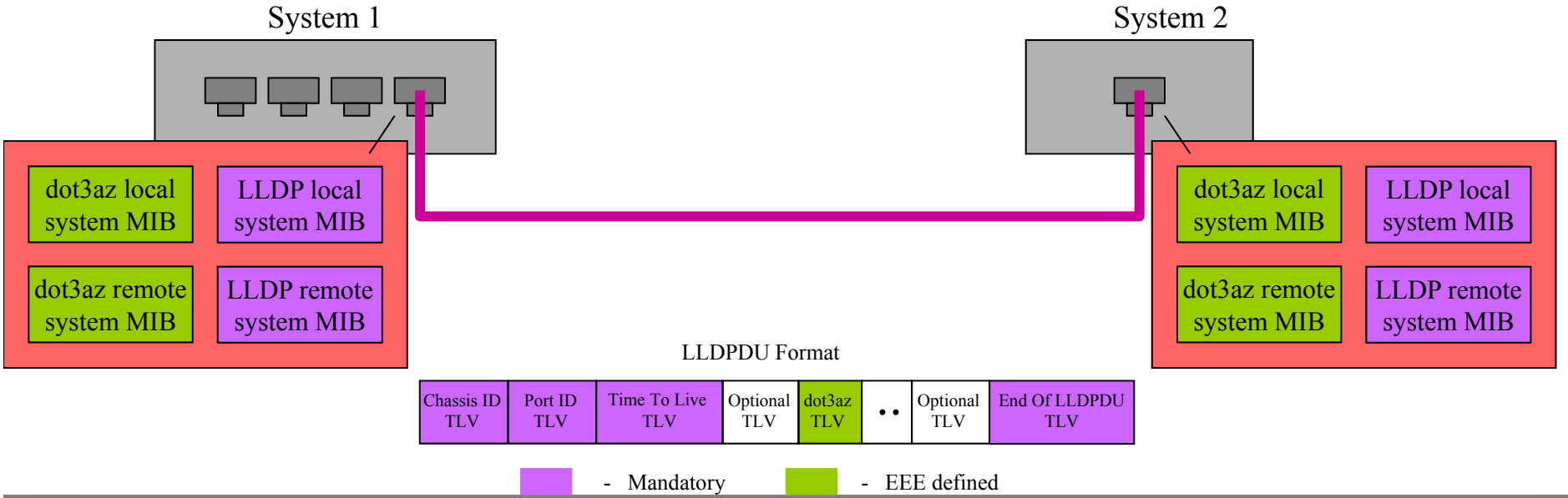
- Operates over a point to point link
- Completely enclosed protocol
  - We define data, it gets transported
    - We don't get to make changes to the protocol
- Data in ‘Local MIB’ transported to ‘Remote MIB’
  - Transported by TLVs (type, length, value)



Source: law_01_0508.pdf

LLDPDU and Associated TLVs

- The LLDP frame consists of an LLDPDU (LLDP Data Unit)
  - LLDPDU is constructed from mandatory TLVs and optional TLVs
  - Mandatory TLVs are chasis ID, Port ID and TTL
  - Optional TLVs can be management TLVs or organizationally specific TLVs
  - Selection of optional TLVs used in the LLDPDU is under network management's control
- TLVs are associated with a station's MIB
  - Mandatory basic LLDP MIB: Associated with basic TLVs
  - Optional LLDP MIB extensions: Associated with optional TLVs
  - IEEE P802.3az needs to define an LLDP MIB extension and associated TLV
- **Consequence: LLDPDU contains more than EEE's TLV and exchange may be triggered by other TLV changes**



To Summarize…

- What LLDP *can* do
  - **Transport** parameters defined in TLVs across a link
  - **Keep** a copy of the remote and local value of a parameter
  - **Automatically** initiate an update upon a change in the local variable's value and/or notify the local agent of a remote change
  - **Support** SM (See guidance from 802.1 in following section)
  - Offers benefits of a **packet based protocol**: CRC protection, so there is no need to worry about this
- What LLDP *cannot* do
  - **Force** specific number of LLDPUs to go out for a single change
  - **Rely** on a fixed rate of exchange
    - There are mechanisms to ensure "quick" updates but there are system interactions that may not be under the control of EEE
  - **Assume** that a LLDPU received or transmitted is due to a specific change in a specific TLV
    - Changes elsewhere in a station's MIB or MIB extension can trigger an LLDPDU exchange. Delay timers to consolidate TLV changes into fewer LLDPUs exist but do not eliminate the issue and their use may be constrained by other system requirements

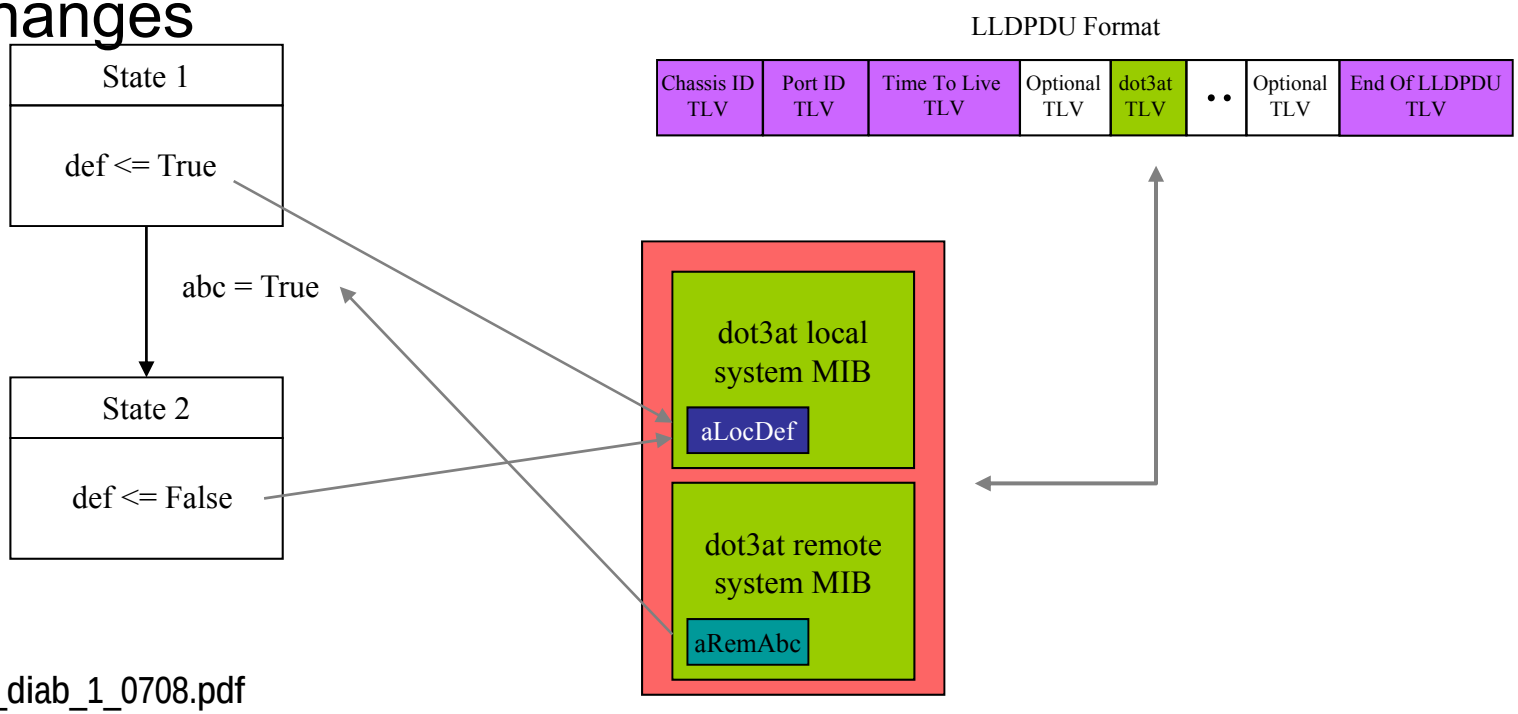
LLDP and Layer 2 Enhancements (DLL) in 802.3

Layer 2 Enhancements

- Officially called “Data Link Layer” or DLL Capabilities
- Several Components
  - (a) Transport mechanism (b) State machine behavior (c) MIB and management (d) Potential additional features
- Transport mechanism (a) above is LLDP. As noted prior, LLDP is stateless, hence desired functionality for EEE and PoEP required additional work beyond the transport mechanisms which gives rise to components (b) and (c)
- Result is a mechanism that employs state machines and allows for negotiation, management and additional features
- State machines define desired behavior and build on top of the transport mechanism

Review: LLDP and State diagrams

- Can't map directly to TLV contents
  - Map through objects in dot3at local and remote MIB
  - Define MIB attribute to variable mapping
  - Allows .3 layers to take action based on variable changes



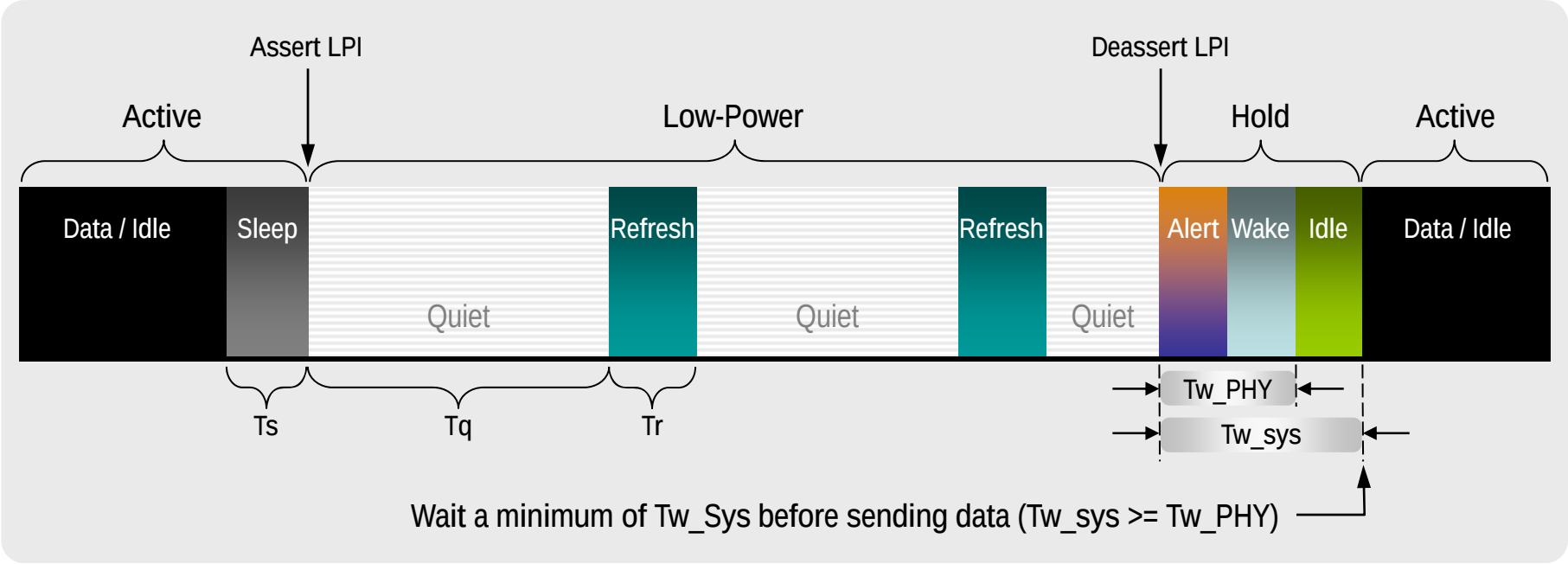
Source: joint_diab_1_0708.pdf

802.1's Guidance for use of a SM w/LLDP

- During the development of IEEE Std 802.3at, discussions with 802.1 via joint sessions on the topic occurred
- Feedback from those discussions
  - No fundamental problem to do State Machine
  - Preferably don't do ACK/NACKs, if you do, you need serial numbers
  - Don't make it too chatty
    - LLDP may be running other protocols
    - Minimize the number of frames transmitted

EEE's Basic Functionality with Respect to LLDP

EEE Low Power Idle Overview



- Low Power Idle (LPI) – PHY powers down during idle periods
- Tw_sys is negotiated via the use of LLDP in IEEE Std 802.3az-2010

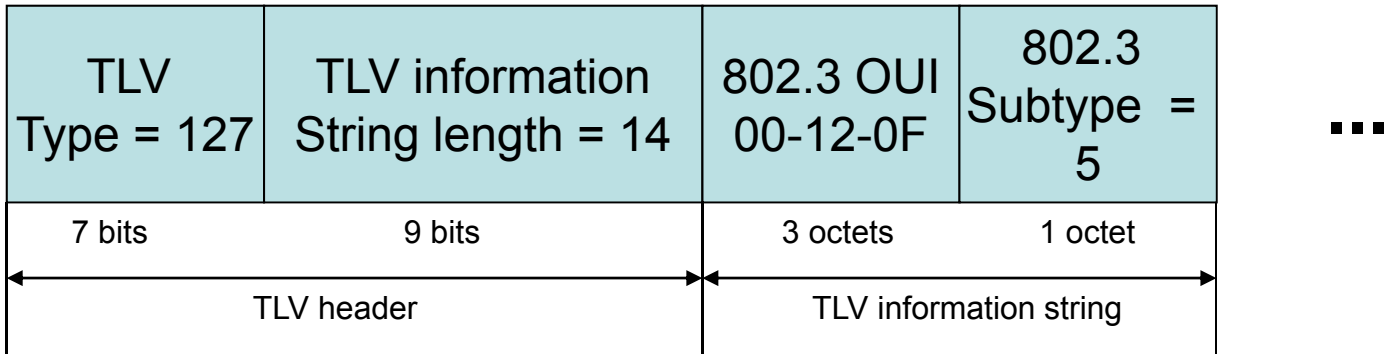
Functionality w.r.t LLDP

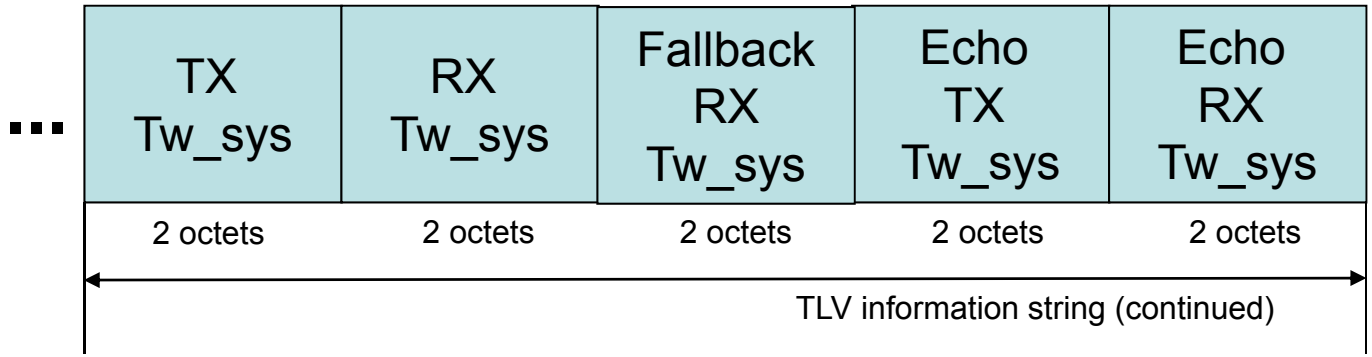
- Use of LLDP functionality
  - Allows the link partners to negotiate for how long to hold-off after wake prior to sending data
    - This can be done in each direction of the link
  - This can be used by RX to turn off more circuitry when it goes to sleep as it has additional time from when the PHY is woken up
  - Additional features like Fallback states
- 2 basic requirements:
  - Initial capability exchange of Tw_sys
    - Upon initialization exchange the Tw_sys parameters to allow for resolution
  - Dynamic negotiation of Tw_sys
    - At any time during operation, allow either link partner in either path to initiate a change its Tw_sys to allow for dynamic savings / performance optimization
- Capability exchange of LLDP ability for speeds <10G
  - In EEE, mandatory for ≥10G speeds. Optional for lower speeds

802.3az's Layer 2: Structure and Operation

- Structure
  - RX and TX on a link maintain their own state and a copy of the link partner's state
    - More accurately, a local and a remote (mirrored) MIB. This is done by the use of LLDP
  - Each link partner instantiates an RX and a TX machine
  - A change of state either locally or remotely triggers an action
    - Action governed by a state machine
    - State machines not symmetric for RX and TX
- As an example, a RX can request more time from its TX link partner
  - RX's SW via management updates a local variable in its MIB
  - This triggers a transition in the RX's state machine and an LLDP frame
  - LLDP frame receipt @TX triggers an update to the mirrored MIB
    - Causes a transition in the TX's state machine to consider request
  - TX can chose to update the wake time allotted to RX through similar process

Energy Efficient Ethernet TLV





The general state change procedure for transmitter is shown in Figure 78–5.

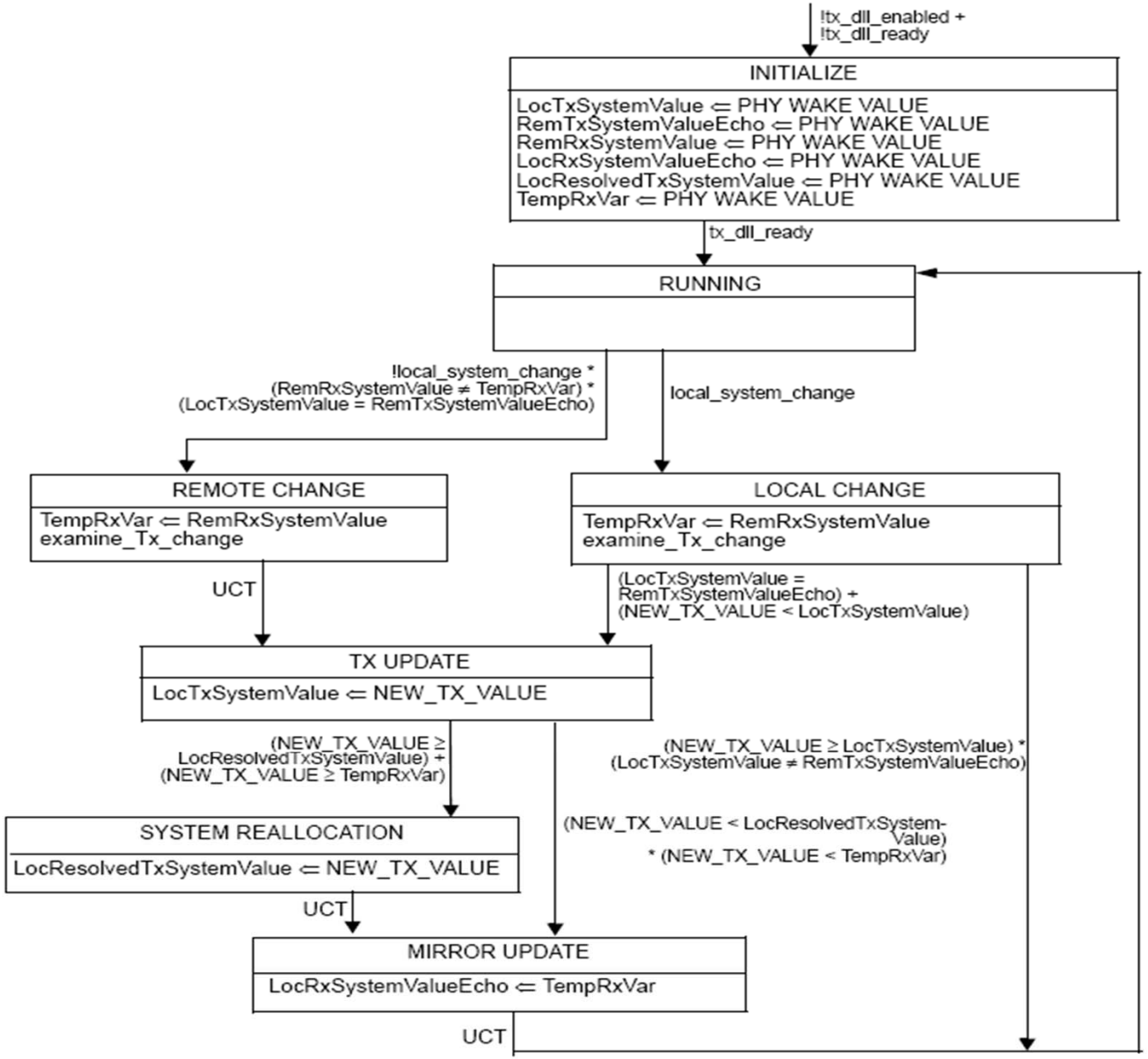


Figure 78–5—EEE DLL Transmitter State Diagram

The general state change procedure for receiver is shown in Figure 78–6.

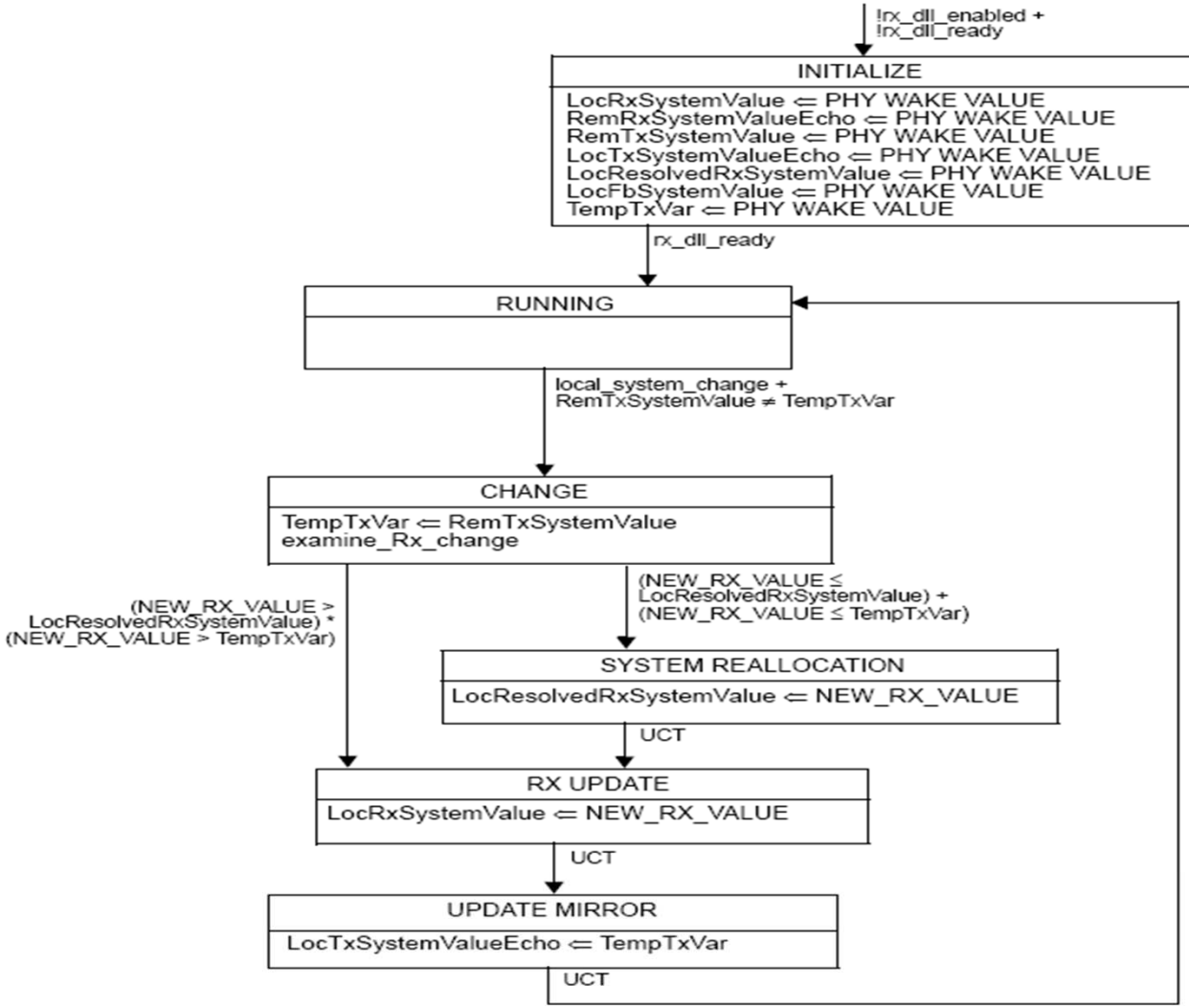


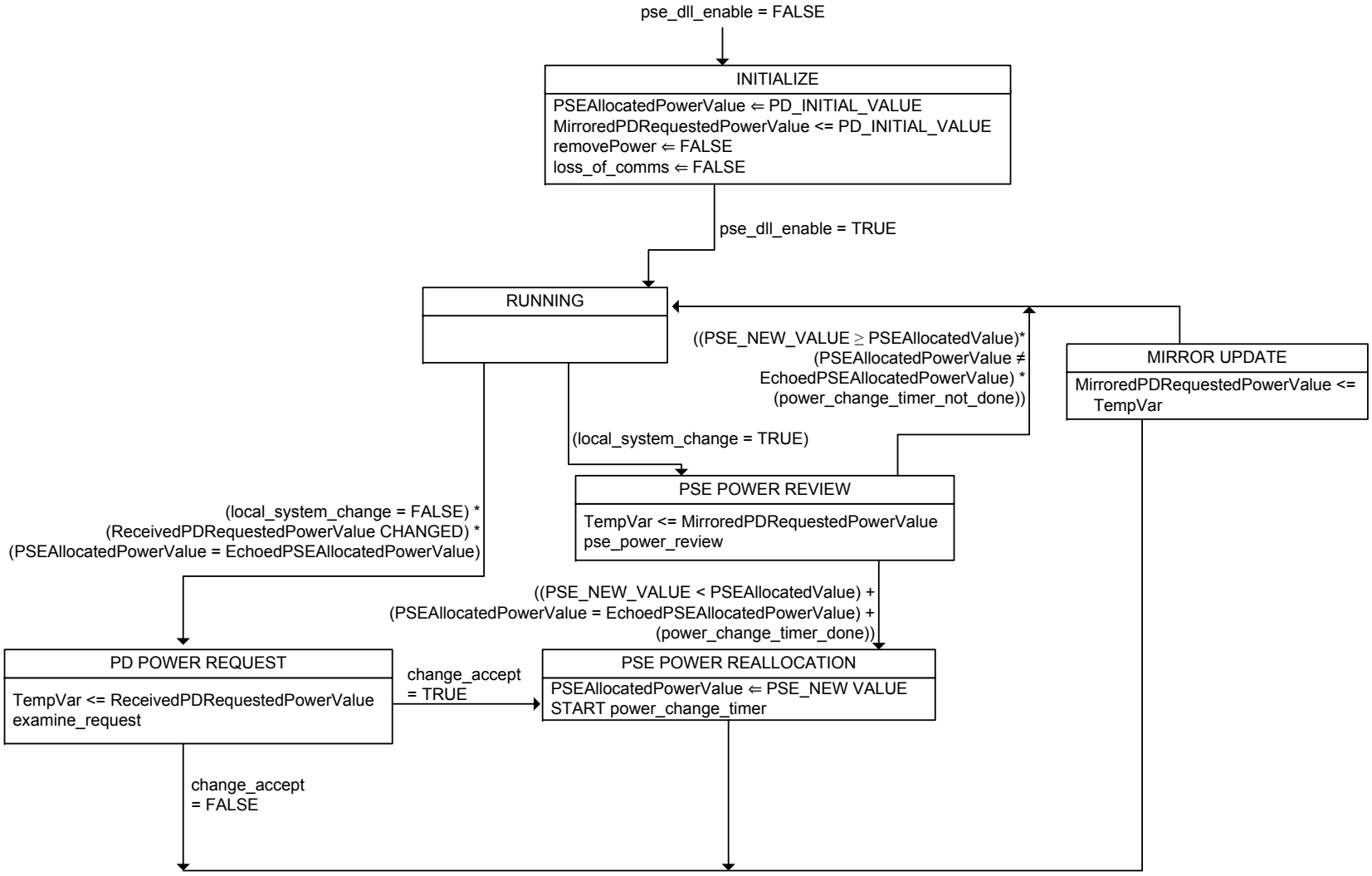
Figure 78–6—EEE DLL Receiver State Diagram

PoEP's Basic Functionality with Respect to LLDP

PoEP's L2: Motivation and SM

- PoEP wanted to use LLDP for a dynamic power allocation between the PD and PSE
  - In an end-span configuration PSE and PD are link partners
  - PD may request new power allocation. PSE responds to PD's request and/or initiates re-allocation
- Similarities between PoEP's L2 and EEE
  - Desire to use LLDP as it is a widely deployed protocol
  - Value of parameters exchanged set by "upper layers"
  - Startup and dynamic requirements on solution
  - Decision on allocation requires assurance that information being acted on is most recent
    - Avoid deadlocks
  - Prior to changing behavior, PSE/PD needs to be confident other side is ready – random power change may have drastic effects!
  - One side has to enforce the decision (PSE for PoEP, TX for EEE)
- Differences between PoEP's L2 and EEE
  - Behavior for PSE and PD unidirectional. EEE has duplex
    - EEE instantiates a TX and an RX for each station

Example 802.3at PSE SM



EEE's LLDP for 40G & 100G NGOPTX

Work Already Done in 802.3az

- Can leverage – point to – IEEE 802.3az for all the LLDP behavior
  - This should all be the same and can leverage work already there
- Pointers to existing work
  - IEEE 802.3 Organizationally Specific TLVs now reside in Clause 77 with EEE TLV defined there.
    - Moved from 802.1AV to 802.3, as part of IEEE Std 802.3bc and amended by 802.3at and 802.3az (79.3.5 and 79.4.2)
  - IEEE 802.3az defined
    - LLDP TLV selection management variables
    - LLDP MIB extensions
    - MIB/TLV cross reference table
    - MIB to state diagram cross reference table
    - State diagram using MIB derived variables
  - IEEE 802.3az defined
    - Management related in Clause 30 (mainly in 30.2.5 and 30.12) and 45 for capability
    - DLL in Clause 78.4

Using DLL / LLDP for Capability Exchange

Capability Exchange without Auto-negotiation

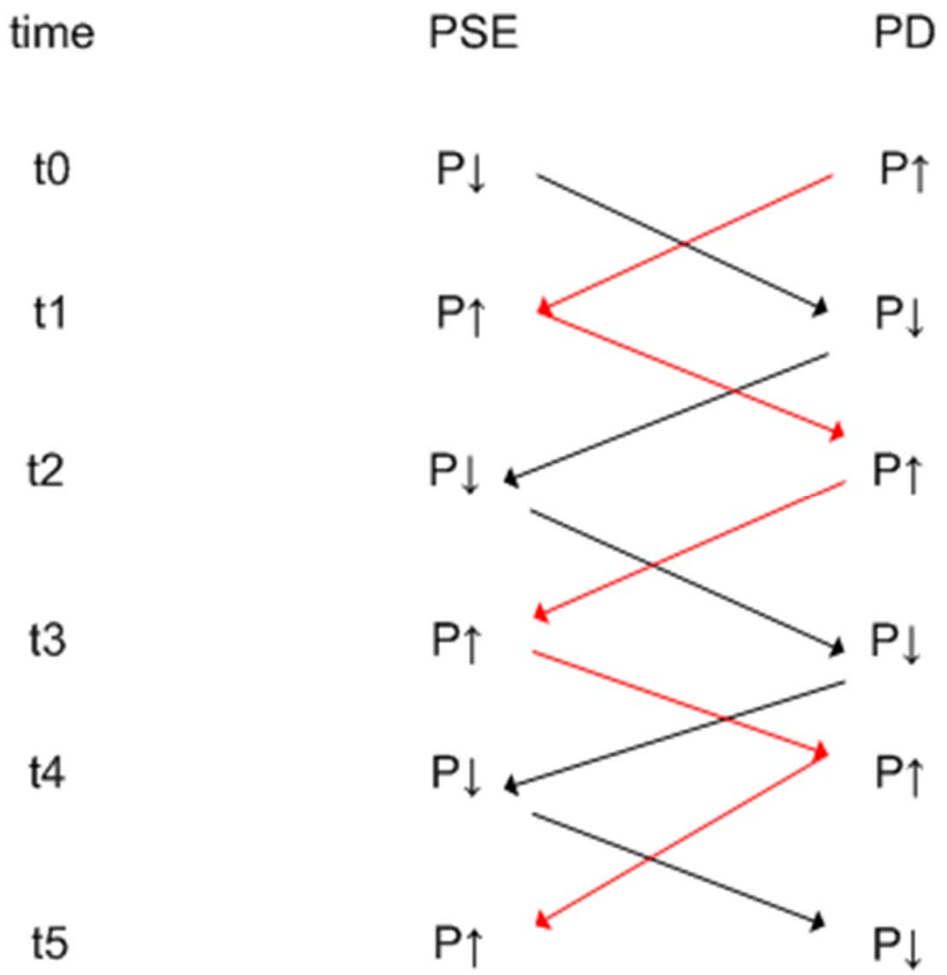
- DLL has been optional on a subset of the interfaces for both PoEP and EEE
  - In EEE, DLL was optional for some EEE interfaces
  - Similarly in PoEP, DLL was optional for Type 1 end-point PSEs
- For that and other reasons, separate mechanisms were necessary for capability exchange. However, the SM are designed to allow for a DLL capability exchange
- Above in conjunction with EEE DLL mandatory for EEE interfaces ≥ 10G, can be extended to allow DLL capability exchange to function as EEE capability exchange
  - Support required when EEE option is implemented
    - EEE consensus: ≥10G systems implement LLDP anyway for other protocols
  - Eliminates the need for auto-negotiation
  - Timing parameters for starting EEE operation separate from starting DLL operation
    - DLL can take time to start-up due to SW subsystems
    - Once capability is exchanged EEE operation commences in a fashion similar to Tw_sys re-negotiation, which is much faster

Proposed Next Steps

- Detailed baseline proposal with text and pointers can be provided when group moves to TF and is entertaining baselines

BACKUP

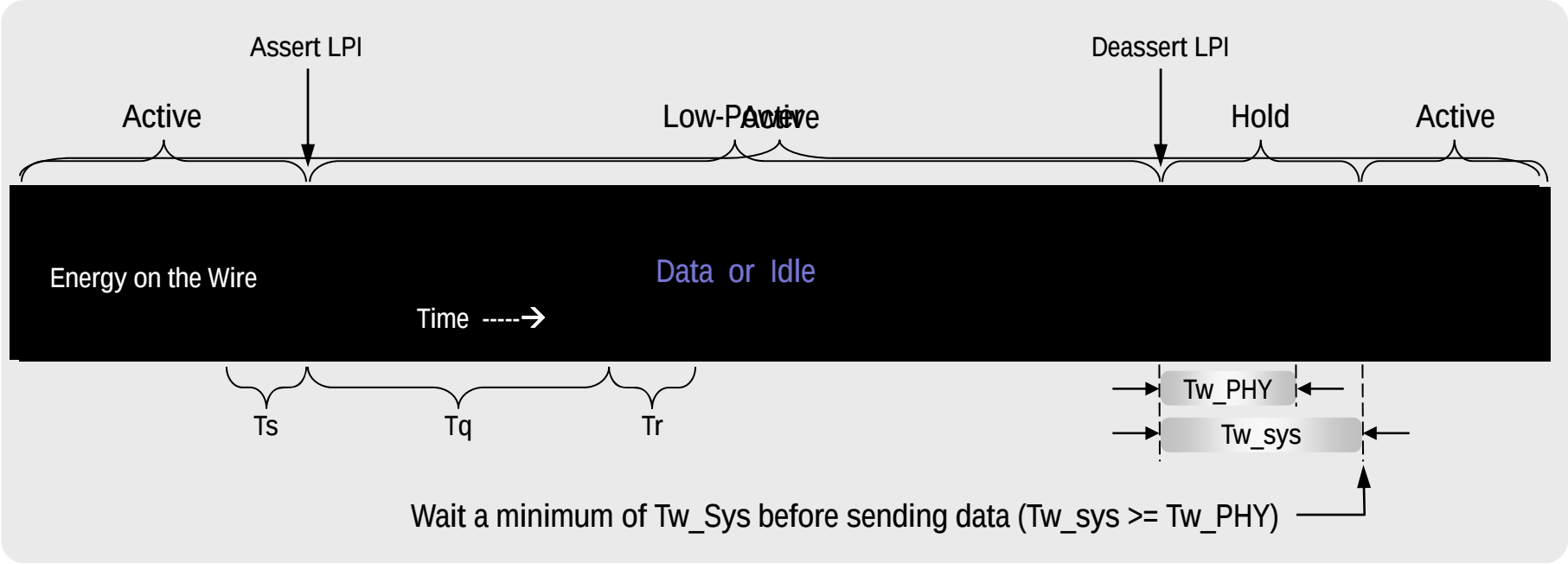
Keeping Track of The Value Set



Keeping Track of The Value Set

- Value advertised by the local partner in some part may depend on the value being advertised by the remote
- Since LLDP's agents and review process may not be real time, a review process may be operating on stale information and/or out of synch information. This can cause unwanted positive feedback
- To ensure this does not happen
  - The Mirrored value is the received value corresponding to which the power review is conducted
  - If a PSE receives a PDU where the echoed value does not match the Allocated Power Value, it ignores the PDU
  - If a PD receives a PDU where echoed value does not match Requested Power Value, it continues to treat the PDU as valid
- EEE: Control policy may react to a change in buffering on the remote side, then theoretically this could occur
  - Put in a similar mechanism into the SM or chose to ignore if its not a practical concern

EEE Low Power Idle Overview



- Low Power Idle (LPI) – PHY powers down during idle periods
- Tw_sys is negotiated via the use of LLDP in IEEE Std 802.3az-2010