Capabilities Negotiation Proposal for Energy-Efficient Ethernet

May 2008, Munich

Aviad Wertheimer & Robert Hays
Intel Corporation

Energy Efficient Ethernet

Contributors & Supporters

- Ozdal Barkan (Aquantia)
- Jim Barnette (Vitesse)
- Hugh Barrass (Cisco)
- Brad Booth (AMCC)
- Mandeep Chadha (Vitesse)
- Joseph Chou (Realtek)
- Dan Dove (ProCurve)
- Bob Grow (Intel)
- Robert Hays (Intel)
- Adam Healey (LSI)
- David Koenen (HP)
- Jeff Lynch (IBM)
- Brian Murray (LSI)
- Wiren Perera (Plato Networks)
- Dimitry Taich (Teranetics)
- Mario Träber (Infineon)
- Aviad Wertheimer (Intel)
- George Zimmerman (Solarflare)

EEE Negotiation Considerations

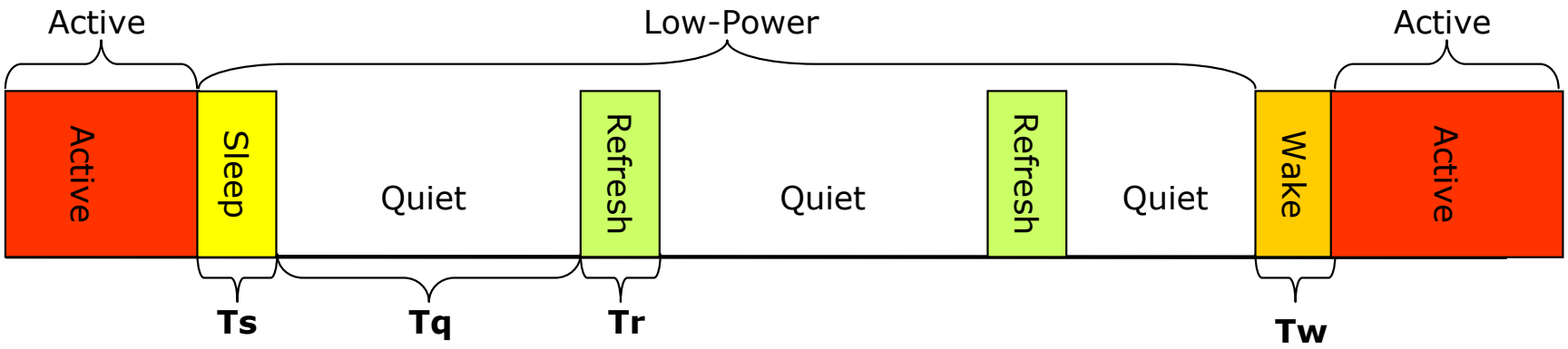
- Negotiation Requirements
  1. Advertise EEE capabilities between link partners
  2. Negotiate the best, common set of PHY parameters
  3. Ability to change Tw, based on system needs, without breaking link
  4. Support for any type of Ethernet system

- Mechanisms Considered:
  - Auto-Negotiation, MAC Control Frames (MCF), Link Layer Data Protocol (LLDP)

- Conclusions:
  - No single mechanism meets all requirements
  - Auto-neg is best for advertising EEE capabilities and PHY parameters because it's supported by the broadest range of applications
  - MCF & LLDP are both capable of Tw re-negotiation
  - LLDP is preferred due to similar usage in other standards (e.g. 802.3at)

EEE Timing Parameters

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.
Wake Time (Tw)	Wait period where no data is transmitted to give the receiving system time to wake up.



EEE Negotiation Proposal

- Auto-Negotiate EEE support and PHY parameters (required)
  - Ts specified for each PHY as fixed value
  - Tq, Tr combinations advertised
    - Up to two Tq, Tr combinations to be specified for each PHY
      - "Reduced energy" duty cycle – Tq:Tr = n:1
      - "Lowest energy" duty cycle – Tq:Tr > n:1
    - PHYs advertise most energy-efficient combination supported and negotiate to lowest common value to ensure robust link quality
  - Tw default value specified for each PHY
    - Default = Tw_{min} to ensure interoperability and prioritize performance

- Use LLDP to change Tw, based on system capabilities & application requirements (optional)
  - Tw_{min} and Tw_{max} values to be specified as boundary conditions
  - Higher Tw allows systems to enter deeper power-saving states, Lower Tw provides better performance (See Hays_01_0108)

EEE Auto-Neg Extension (1 of 3)

Use Message Next Page (Page 0) with reserved 0x0A message code for EEE Technology Next Page Message code (Annex 28C and Annex 73A).

Message Count#	M 10	M 9	M 8	M 7	M 6	M 5	M 4	M 3	M 2	M 1	M 0	Message Code Description
10	0	0	0	0	0	0	0	1	0	1	0	EEE Technology Message Code. EEE capability 2 unformatted next pages to follow.
11 …	0	0	0	0	0	0	0	1	0	1	1	Reserved for future Auto-Negotiation use

EEE Auto-Neg Extension (2 of 3)

Use Unformatted next page (Page 1) to define IEEE 8023.az support.

Bit	Bit definition
U10	Next page - 1
U9:U6	Reserved, transmit as 0
U5	10GBASE-KR EEE support (0 = no, 1 = yes)
U4	10GBASE-KX4 EEE support (0 = no, 1 = yes)
U3*	Reserved, transmit as 0
U2	10GBASE-T EEE support (0 = no, 1 = yes)
U1	1000BASE-T EEE support (0 = no, 1 = yes)
U0	100BASE-TX EEE support (0 = no, 1 = yes)

*Note: U3 is reserved for possible future 1000BASE-KX EEE support.

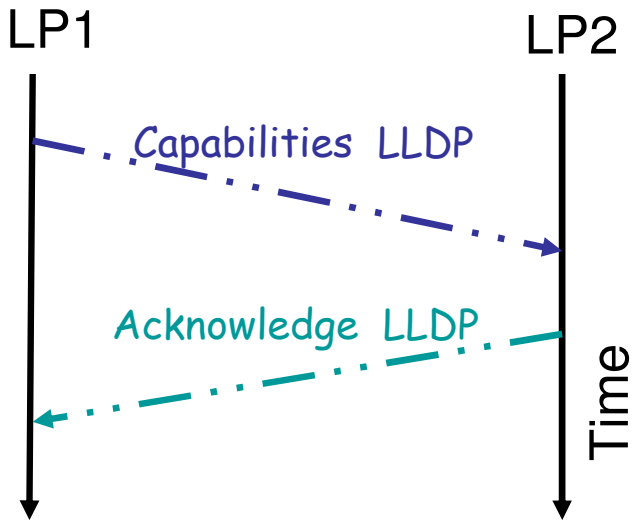
EEE Auto-Neg Extension (3 of 3)

Use Unformatted next page (Page 2) to define PHY energy efficiency level.

Bit	Bit definition
U10	Next page (to enable support of future technologies)
U9:U6	Reserved, transmit as 0
U5	10GBASE-KR PHY energy (0 = Reduced, 1 = Lowest)
U4	10GBASE-KX4 PHY energy (0 = Reduced, 1 = Lowest)
U3*	Reserved, transmit as 0
U2	10GBASE-T PHY energy (0 = Reduced, 1 = Lowest)
U1	1000BASE-T PHY energy (0 = Reduced, 1 = Lowest)
U0	100BASE-TX PHY energy (0 = Reduced, 1 = Lowest)

*Note: U3 is reserved for possible future 1000BASE-KX EEE support.

LLDP Twake Negotiation



Capabilities LLDP (CLLDP):
LP1 EEE Tw value change request

Acknowledge LLDP (ALLDP):
LP2 EEE Tw value change confirmation

- LLDP-based control protocol:
  1. LP1 requests Tw change to value X
  2. LP2 acknowledges request with new Tw value of X or Y, which ever is lower
  3. EEE may resume with new Tw value

- Tw values:
  - Tw_{min} (default value) to be specified as a function of PHY resume capabilities
  - Tw_{max} to be specified to bound design and test requirements
  - Tw (actual value) negotiated to LCD as dictated by system capabilities (e.g. buffer depths) and performance priority

EEE LLDP Frame

Bytes	Content	Value	Description
6	MAC DA	01-80-C2-00-00-0E	LLDP_Multicast address
6	MAC SA		MAC address of sending station or port
2	Ethertype	88-CC	LLDP Ethertype
9	Chassis ID TLV		Mandatory TLV (see 802.1AB)
9	Port ID TLV		Mandatory TLV (see 802.1AB)
4	Time To Live TLV		Mandatory TLV (see 802.1AB)
2	TLV type/Length	127/7	TLV Type and Length
3	OUI	00-12-0F	802.3 OUI
1	Subtype	TBD	Energy Efficient Ethernet 802.3 subtype
2	Twake		Wake Time prior to move into Active state
1	Acknowledge		Acknowledge
2	End Of LLDPDU TLV	00-00	Mandatory TLV (see 802.1AB)
16	Padding + CRC		

EEE LLDP Twake and Acknowledge

Twake Field

Bit	Content	Value	Description
15:0	Twake		Wake Time, in microseconds, when no data is transmitted to give the receiving system time to wake up from Low-Power state. Valid range of values is TBD

Acknowledge Field

Bit	Content	Value	Description
7:1	Reserved	0x0	Reserved
0	Ack		Acknowledge - Indicates the response to the last change in requested 802.3az parameters change. 0 = Initial CLLDP request frame 1 = Acknowledge frame

Opens

- Auto-neg state machine modifications
- EEE capabilities in MII extension registers
- 802.1AB spec changes for EEE
- Ts values
- Tq, Tr combinations
- Tw_{min} and Tw_{max} values

Thank You!

- Questions?

Back-Up

Auto-negotiation - Transmitter , Fig 40C-2

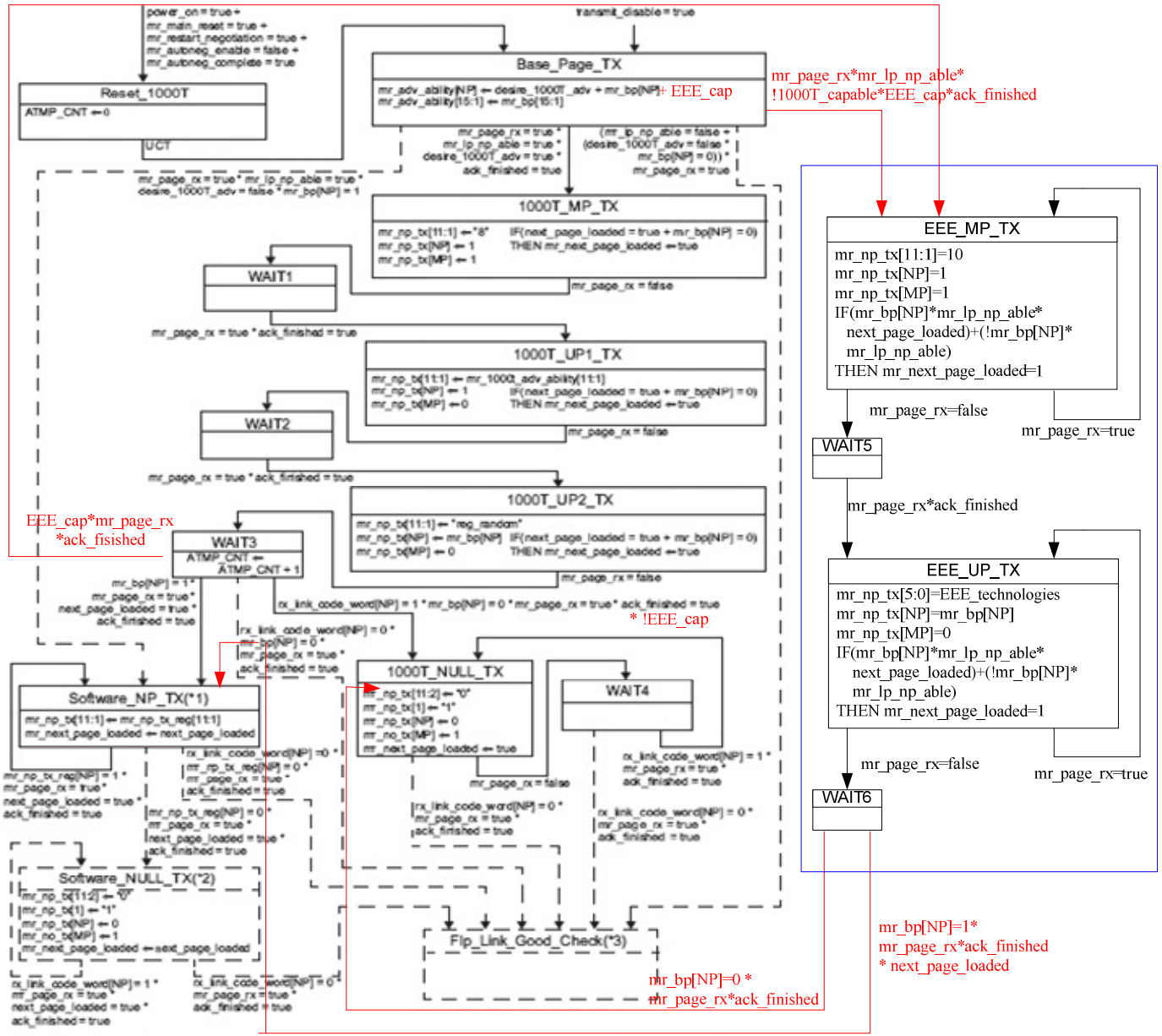


Figure 40C–2—Auto-Negotiation Transmit state diagram add-on for 1000BASE-T

Auto-negotiation – Receiver, Fig 40C-3

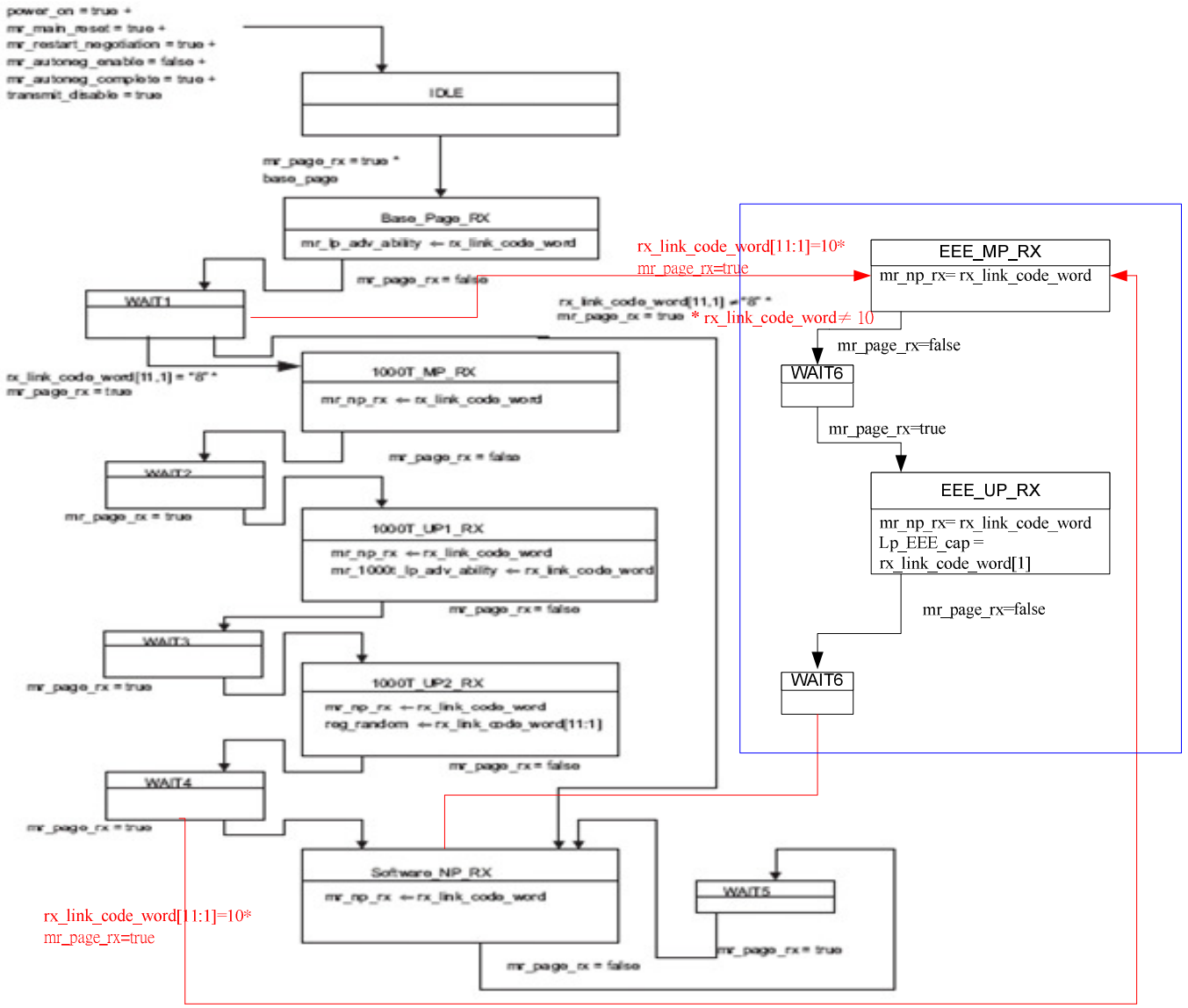
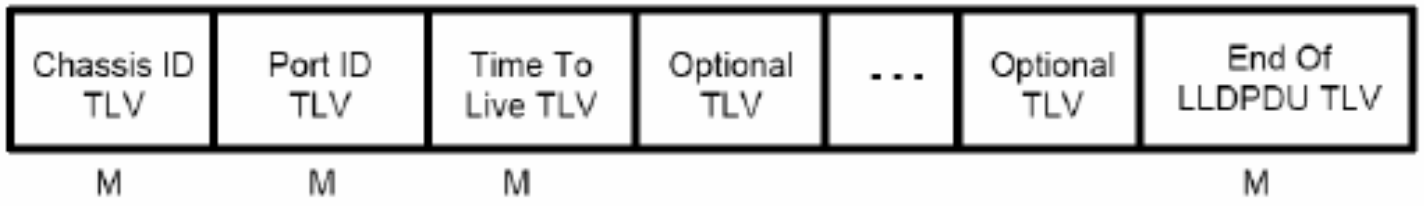


Figure 40C–3—Auto-Negotiation Receive state diagram add-on for 1000BASE-T

Energy Efficient Ethernet

LLDPDU and TLV format (802.1AB)

LLDPDU Format



M - mandatory TLV - required for all LLDPDUs

TLV Format

LLDP IEEE 802.3 TLV Subtypes (802.1AB)

Table G-1—IEEE 802.3 Organizationally Specific TLVs

IEEE 802.3 subtype	TLV name	Subclause reference
0	reserved	—
1	MAC/PHY Configuration/Status	G.2
2	Power Via MDI	G.3
3	Link Aggregation	G.4
4	Maximum Frame Size	G.5
5–255	reserved	—