



Energy Efficient Ethernet Switching Perspective

Dan Dove

ProCurve Networking by HP



Supporters

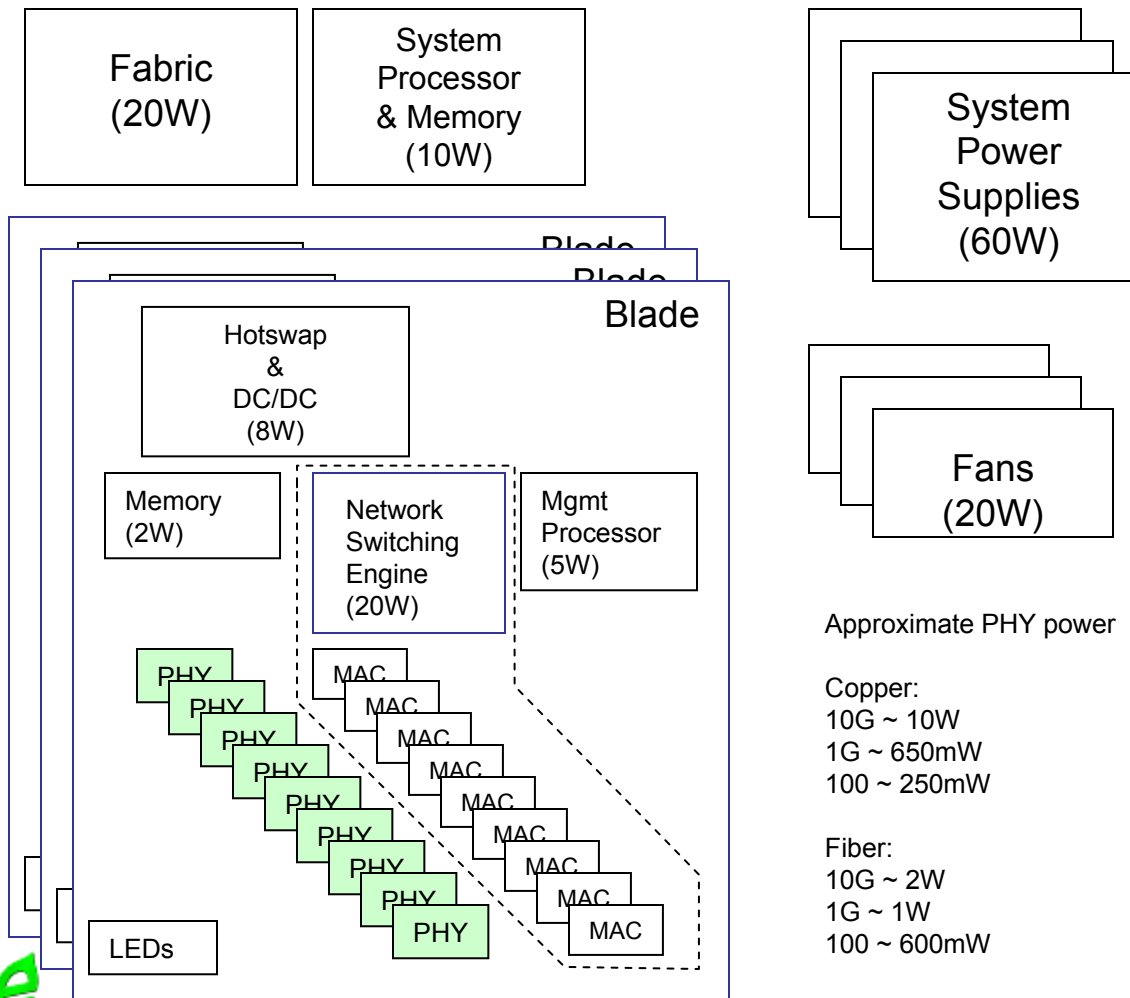
Thanks to the following for review, feedback, and support:

TBD

Switch Perspective

- PHY Power Savings is important
- MAC & Infrastructure Savings also important
- Determine Key control points
- Determine Key control parameters
- Determine Means for communicating control parameters

Switch Infrastructure (example)



PHY Power is a substantial percentage of the overall switch infrastructure power consumption but not sufficient to achieve desired energy savings

10GBASE-T dramatically shifts the percentage of power to PHY but for many customers fiber or 10/100 or 1000BASE-T will be used.

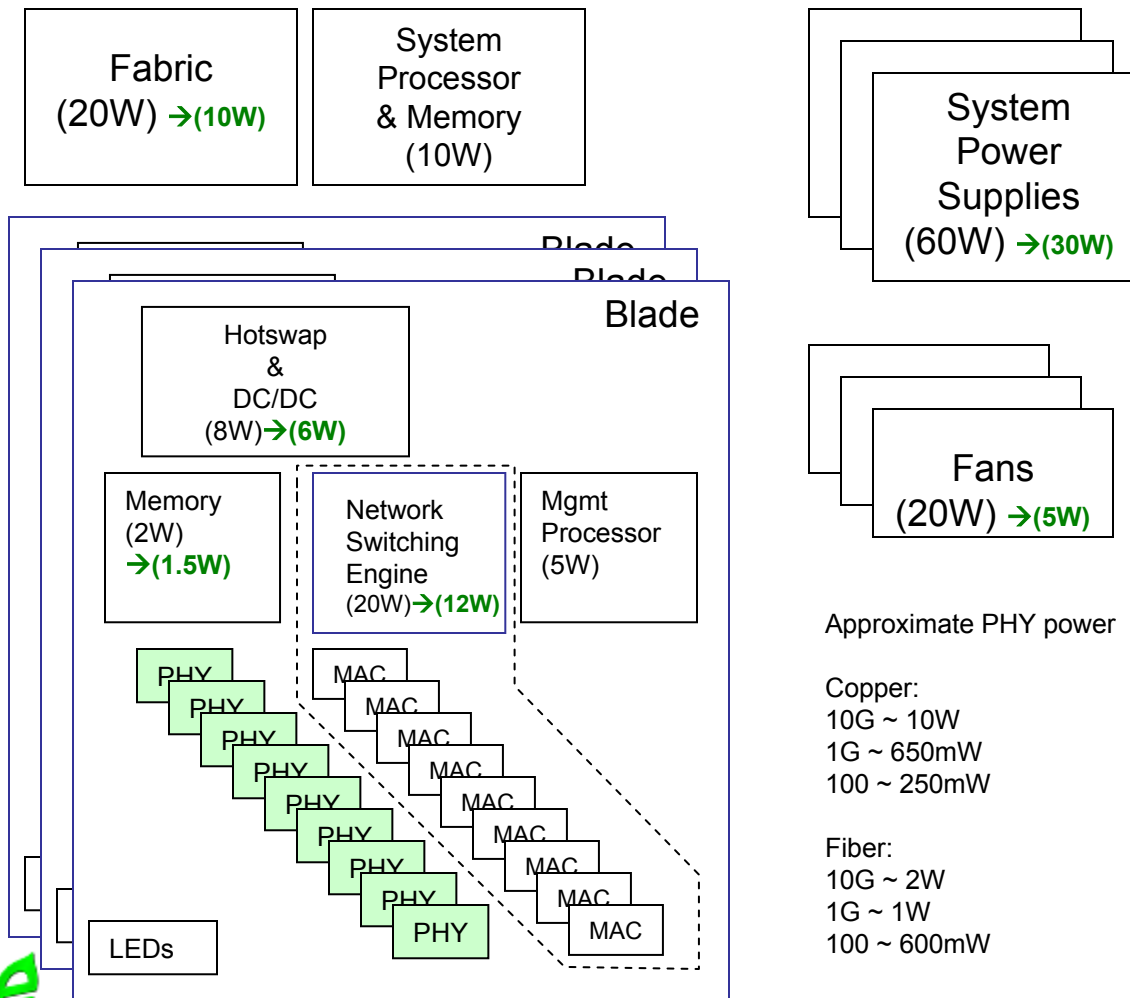
EEE should be useful not just for reducing PHY power, but for reducing system power

Approximate PHY power

Copper:
10G ~ 10W
1G ~ 650mW
100 ~ 250mW

Fiber:
10G ~ 2W
1G ~ 1W
100 ~ 600mW

Switch Infrastructure (example)



Switch MAC, NSE, Memory are a good portion (~3x/port) of energy consumption for most networking link technologies.

Powering-down portions of these circuits provides a two-fold benefit

- 1) Reduces energy used
- 2) Provides opportunity to shut-down other infrastructure (DC/DC, Fans, etc)

Reasonable estimates show that **~1.5W- 3W/port** can be reduced in infrastructure

What to power-down and how to do it, is outside the scope of 802.3, but providing means to communicate when to power-down and when to resume operation may be appropriate for 802.3 to address

Control Considerations

Portions of MAC, NSE, Memory may be powered down, but will require a finite time (Twake) to resume operation (resuming clocks, power supplies, etc)

This time is not a constant that can be negotiated at link startup as it may be dependent on aggregate system utilization however a minimum value, required by the PHY, may be negotiated at link startup

A means for negotiating Twake is necessary to balance outbound buffering (of source) against inbound Twake requirements (of receiver and MAC, NSE, Memory)

A means for signaling to the MAC that a packet is coming is required so upstream circuits can be activated

- Quiet vs Wake – A MAC may sleep when the PHY is receiving Quiet and will wake up and be prepared to process data when Wake is received
- Wake may be the currently defined IDLE for each given PHY technology.
- Note: If the MAC is a legacy device, the PHYs can be configured to operate in “legacy mode” which would transmit Quiet, Refresh and Wake based only on auto-negotiated parameters. The MAC/PHY interface would operate as currently defined

Control Considerations (cont)

The value of T_{wake} is a variable that must be communicated to the source of data on a link

- May be negotiated upon link startup via Auto-Negotiation (low-end devices)
- May also be re-negotiated via LLDP (higher end devices)

To prevent chattering, a minimum “on time” should be communicated between link partners so that a port does not resume sleep state only to be awoken shortly after

- A minimum period of normal IDLE signal may be desired after transmission to allow additional packets to be transmitted with minimum IPG once a receiving station has been awoken
- The sum of $T_s + T_w$ could be a minimum IPG or greater, but no less than a minimum IPG.

Low Power IDLE Approach

Operating Modes

- Normal IDLE (Wake) – Active IDLE used by technologies today, indicates “resume operation”
- Low Power IDLE (Quiet) – Mode of operation where TX and RX power are reduced which includes Refresh & Quiescent line signals
- Refresh – Signal transmitted by PHY on occasion to allow destination to refresh its RX parameters, coding is transparent to MAC and appears as Low Power IDLE

Auto-Negotiate Key Parameters

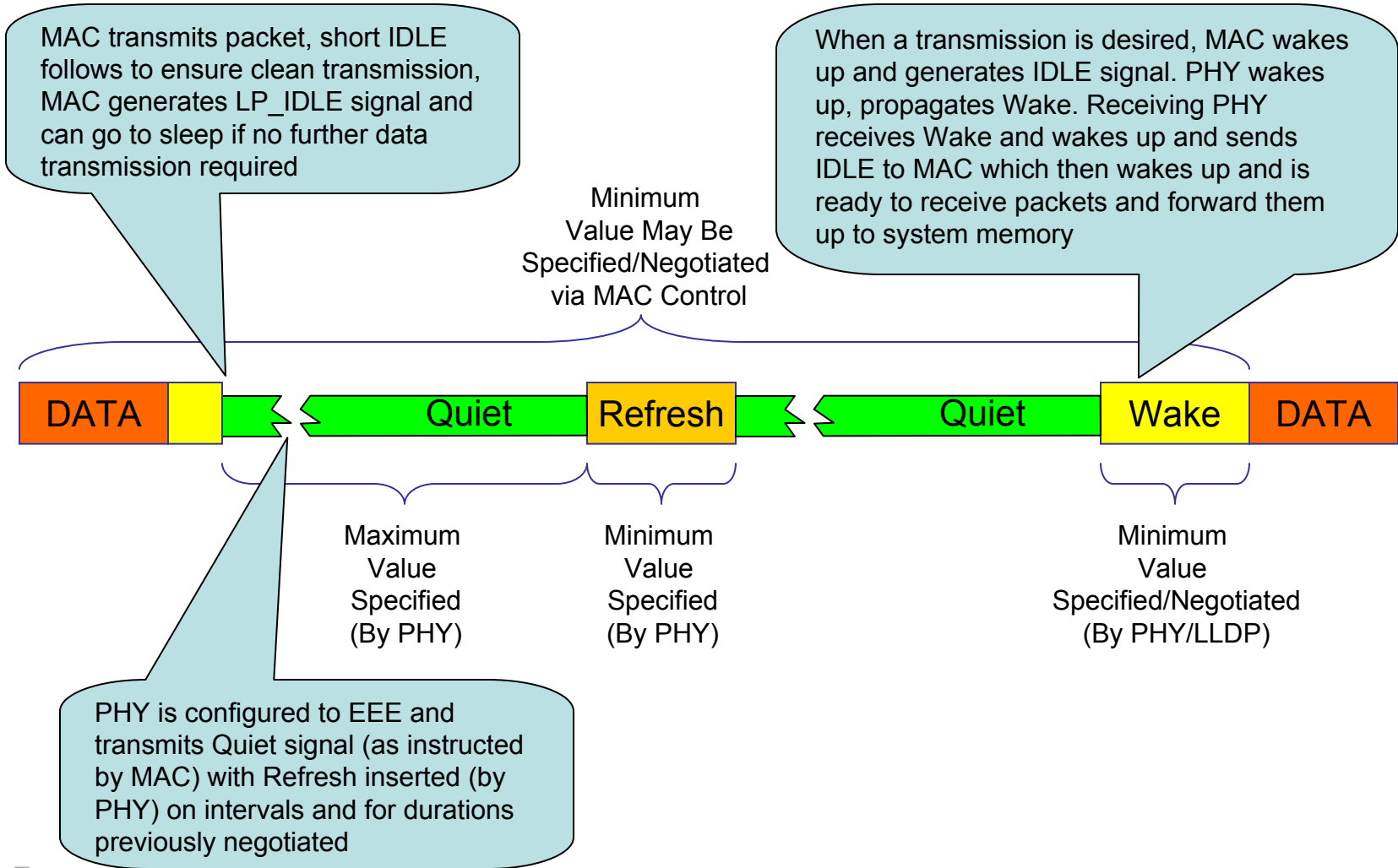
- Support for EEE
- Maximum T_q – Period between refresh signals on line (PHY dependent)
- Minimum T_r – Period of refresh required to bring PHY back to proper state (PHY dependent)
- Minimum T_w – Period of time required by PHY or MAC to regain full functionality (PHY & MAC dependent)

Low Power IDLE Approach (cont)

Communication Methods

- Auto-Negotiation of fundamental operating parameters (Tq_{max} , Tr_{min} , Tw_{min})
- Renegotiation via LLDP if necessary (MAC may require longer Tw than minimum to optimize efficiency)
- Signaling at the PHY layer using Quiet, Refresh and Wake
- Signaling from MAC to PHY to indicate Normal (Data/IDLE), and LP_IDLE states
- Signaling from PHY to MAC to indicate Normal (Data/IDLE), LP_IDLE states (quiet, refresh) received
- LLDP TLV should be defined to negotiate EEE level of functionality
 - Some might argue that LLDP negotiation would be redundant with link Auto-Negotiation, but the two have unique requirements.
 - Link Auto-Negotiation defines minimums and maximums that PHY can tolerate
 - LLDP would define values consistent with Link AN that extend performance. (i.e.: greater than minimums or less than maximums)
 - LLDP would extend efficiency of system based on parameters that vary with time and are not predictable at link initialization

Low Power IDLE Approach



MAC / PHY Signaling

MII

- Normal IDLE (Idle, Wake)
- Low Power IDLE
(Quiet, Refresh)

TX_EN	TX_ER	TXD<3:0>	Indication
0	0	0000 through 1111	Normal inter-frame
0	1	0000	Reserved
0	1	0001	EEE Low Power IDLE
0	1	0010 through 1111	Reserved
1	0	0000 through 1111	Normal data transmission
1	1	0000 through 1111	Transmit error propagation

RX_DV	RX_ER	RXD<3:0>	Indication
0	0	0000 through 1111	Normal inter-frame
0	1	0000	Normal inter-frame
0	1	0001	EEE Low Power IDLE
0	1	0010 through 1111	Reserved
0	1	1110	False Carrier indication
0	1	1111	Reserved
1	0	0000 through 1111	Normal data reception
1	1	0000 through 1111	Data reception with errors

MAC / PHY Signaling

GMII

- Normal IDLE (Idle, Wake)
- Low Power IDLE (Quiet, Refresh)

TX_EN	TX_ER	TXD<7:0>	Description	PLS_DATA.request parameter
0	0	00 through FF	Normal inter-frame	TRANSMIT_COMPLETE
0	1	00	Reserved	—
0	1	01	EEE Low Power IDLE	No applicable parameter
0	1	02 through 0E	Reserved	—
0	1	0F	Carrier Extend	EXTEND (eight bits)

RX_DV	RX_ER	RXD<7:0>	Description	PLS_DATA.indication parameter
0	0	00 through FF	Normal inter-frame	No applicable parameter
0	1	00	Normal inter-frame	No applicable parameter
0	1	01	EEE Low Power IDLE	No applicable parameter
0	1	02 through 0D	Reserved	—

MAC / PHY Signaling

SERDES (Clause 36)

- Normal IDLE (Idle, Wake)
- Low Power IDLE
(Quiet, Refresh)

Table 36-3—Defined ordered_sets

Code	Ordered_Set	Number of Code-Groups	Encoding
/C/	Configuration		Alternating /C1/ and /C2/
/C1/	Configuration 1	4	/K28.5/D21.5/Config_Reg ^a
/C2/	Configuration 2	4	/K28.5/D2.2/Config_Reg ^a
/I/	IDLE		Correcting /I1/, Preserving /I2/
/I1/	IDLE 1	2	/K28.5/D5.6/
/I2/	IDLE 2	2	/K28.5/D16.2/
/LPI/	Low Power Idle		Correcting /LI1/, Preserving /LI2/
/LI1/	Low Power Idle 1	2	/K28.5/D6.5/
/LI2/	Low Power Idle 2	2	/K28.5/D26.4/
	Encapsulation		
/R/	Carrier_Extend	1	/K23.7/
/S/	Start_of_Packet	1	/K27.7/
/T/	End_of_Packet	1	/K29.7/
/V/	Error_Propagation	1	/K30.7/

^aTwo data code-groups representing the Config_Reg value.

- Use alternative ordered sets to communicate Low Power Idle state

MAC / PHY Signaling

XGMII

- Normal IDLE (Idle, Wake)
- Low Power IDLE (Quiet, Refresh)

TXC	TXD	Description	PLS_DATA.request parameter
0	00 through FF	Normal data transmission	ZERO, ONE (eight bits)
1	00 through 05	Reserved	—
1	06	EEE Low Power IDLE	No applicable parameter
1	07	Idle	No applicable parameter (Normal inter-frame)

RXC	RXD	Description	PLS_DATA.indication parameter
0	00 through FF	Normal data transmission	ZERO, ONE (eight bits)
1	00 through 05	Reserved	—
1	06	EEE Low Power IDLE	No applicable parameter
1	07	Idle	No applicable parameter (Normal inter-frame)

MAC / PHY Signaling

XAUI (Clause 36,37)

- Normal IDLE (Idle, Wake)
- Low Power IDLE
(Quiet, Refresh)

XGMII TXC	XGMII TXD	PCS code-group	Description
0	00 through FF	Dxx.y	Normal data transmission
1	07	K28.0 or K28.3 or K28.5	Idle in I
1	07	K28.5	Idle in T
1	06	K28.0 or K28.3 or K28.5 ⁽¹⁾	Low Power Idle
1	9C	K28.4	Sequence
1	FB	K27.7	Start
1	FD	K29.7	Terminate
1	FE	K30.7	Error
1	Other value in Table 36-2	See Table 36-2	Reserved XGMII character
1	Any other value	K30.7	Invalid XGMII character

NOTE—Values in TXD column are in hexadecimal.
(1) Insertion of /D20.5/ per rules defined below

- Insertion of /D20.5/ to delineate Low Power Idle is being transmitted

MAC / PHY Signaling

XAUI (Clause 36,37)

- Normal IDLE (Idle, Wake)
- Low Power IDLE

(Quiet, Refresh)

A sequence of $\|I\|$ ordered_sets consists of one or more consecutively transmitted $\|K\|$, $\|R\|$ or $\|A\|$ ordered_sets, as defined in Table 48–4. Rules for $\|I\|$ ordered_set sequencing shall be as follows:

- $\|I\|$ sequencing starts with the first column following a $\|T\|$.
- The first $\|I\|$ following $\|T\|$ alternates between $\|A\|$ or $\|K\|$ except if an $\|A\|$ is to be sent and less than r [see item d)] columns have been sent since the last $\|A\|$, a $\|K\|$ is sent instead.
- $\|R\|$ is chosen as the second $\|I\|$ following $\|T\|$.
- Each $\|A\|$ is sent after r non- $\|A\|$ columns where r is a randomly distributed number between 16 and 31, inclusive. The corresponding minimum spacing of 16 non- $\|A\|$ columns between two $\|A\|$ columns provides a theoretical 85-bit deskew capability.
- When not sending an $\|A\|$, either $\|K\|$ or $\|R\|$ is sent with a random uniform distribution between the two. **Insertion of /D20.5/ to communicate “Low Power Idle” will not alter the distribution.**
- Whenever sync_status=OK, all $\|I\|$ received during idle are translated to XGMII Idle control characters for transmission over the XGMII. All other $\|I\|$ received during idle are mapped directly to XGMII data or control characters on a lane by lane basis. **with the exception of /D20.5/ (Low Power Idle) being detected in any row which will result in all columns reporting LP_IDLE.**

MAC / PHY Signaling

XAUI (Clause 36,37)

- Normal IDLE (Idle, Wake)
- Low Power IDLE

(Quiet, Refresh)

	PCS																	
LANE 0	K	R	S	Dp	D	D	D	---	D	D	D	D	A	D	R	K	K	R
LANE 1	K	R	Dp	Dp	D	D	D	---	D	D	D	T	A	R	R	D	K	D
LANE 2	K	R	Dp	Dp	D	D	D	---	D	D	D	K	A	R	D	K	K	R
LANE 3	K	R	Dp	Ds	D	D	D	---	D	D	D	K	A	R	R	K	D	R

- Insertion of /D20.5/ to into one of four R or K symbols (per column) randomly to notify the other end of the link that it is receiving Quiet
- Maintains randomness of /A/K/R/ symbols
- PCS detects normal Idle when continuous A/K/Rs are received

MAC / PHY Signaling

BASE-R (Clause 49)

- Normal IDLE (Idle, Wake)
- Low Power IDLE

(Quiet, Refresh)

Table 49–1—Control codes

Control Character	Notation	XGMII Control Code	10GBASE-R Control Code	10GBASE-R O Code	8B/10B Code ^a
idle	/I/	0x07	0x00		K28.0 or K28.3 or K28.5
LP_IDLE	/LPI/	0x06	0x07		K28.0 or K28.3 or K28.5 ⁽¹⁾
start	/S/	0xfb	Encoded by block type field		K27.7

- Replace idle control codes with 0x07 to signify PCS should transmit “Quiet”
- PCS receives 0x07 control codes and replaces them with XGMII Control Code 0x06 to communicate that Quiet is being received.

Conclusions

System Energy Savings beyond the PHY can be gained via EEE specification

Auto-Negotiate Key Parameters

- EEE Capable
- Tr_min (minimum PHY requirements to remain functional)
- Tq_max (maximum PHY requirement to remain functional)
- Twake_min (defined at AN, may be changed via LLDP)

Simple Extensions to MAC/PHY interface allow rapid signaling to MAC when power-up is necessary

Defined extensions for SERDES, XAUI, BASE-R when used for MAC/PHY interface