Energy Efficient Ethernet
Switching Perspective

Dan Dove

Dove Networking Solutions

for

ProCurve Networking by HP

Supporters

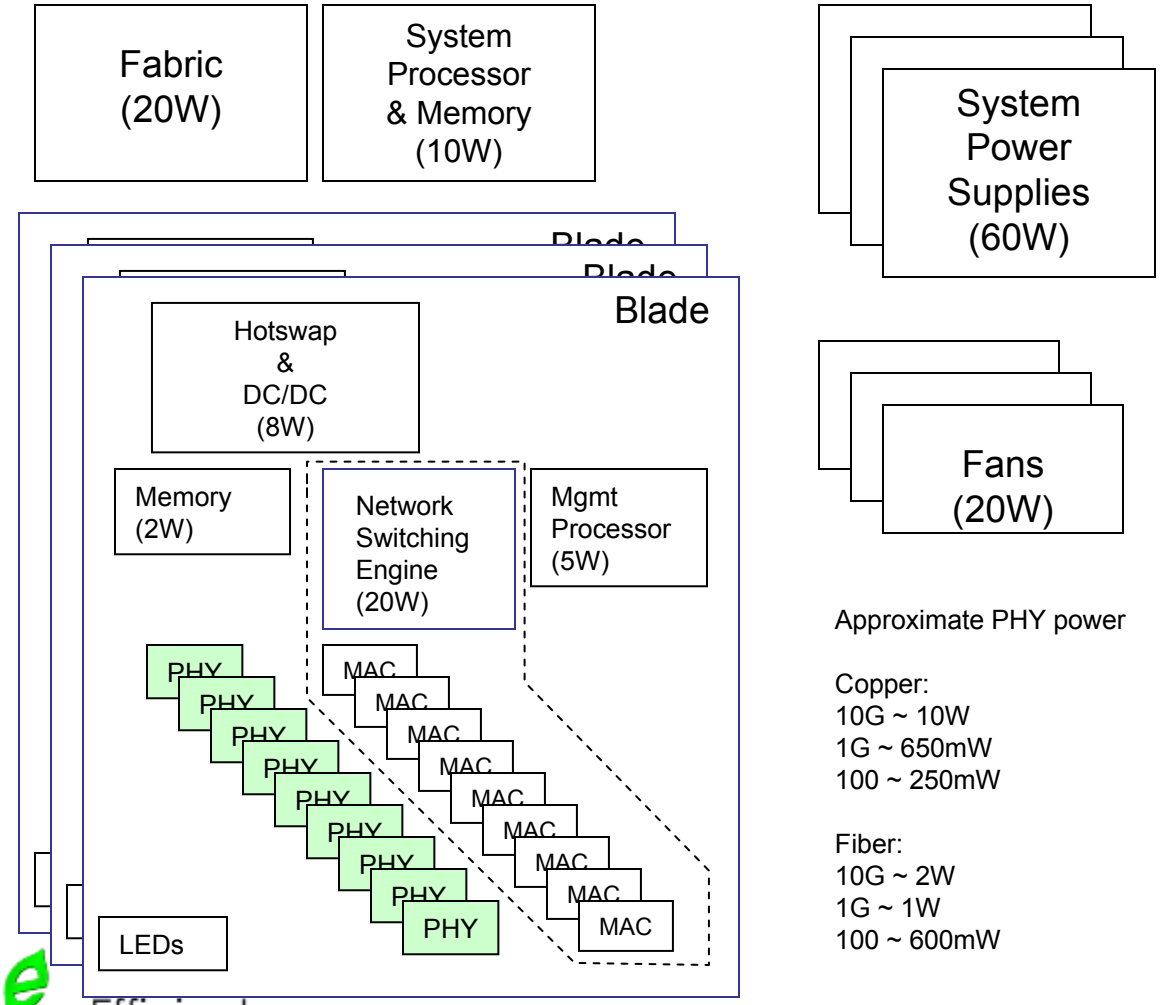
Thanks to the following for review, feedback, and support:

David Law	3Com
Brad Booth	AMCC
David Koenen	Hewlett Packard
Rob Hays	Intel
Brian Murray	LSI
Adam Healey	LSI
Joseph Chou	Realtek
Gavin Parnaby	Solarflare
George Zimmerman	Solarflare
Sanjay Kasturia	Teranetics

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)



Approximate PHY power

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

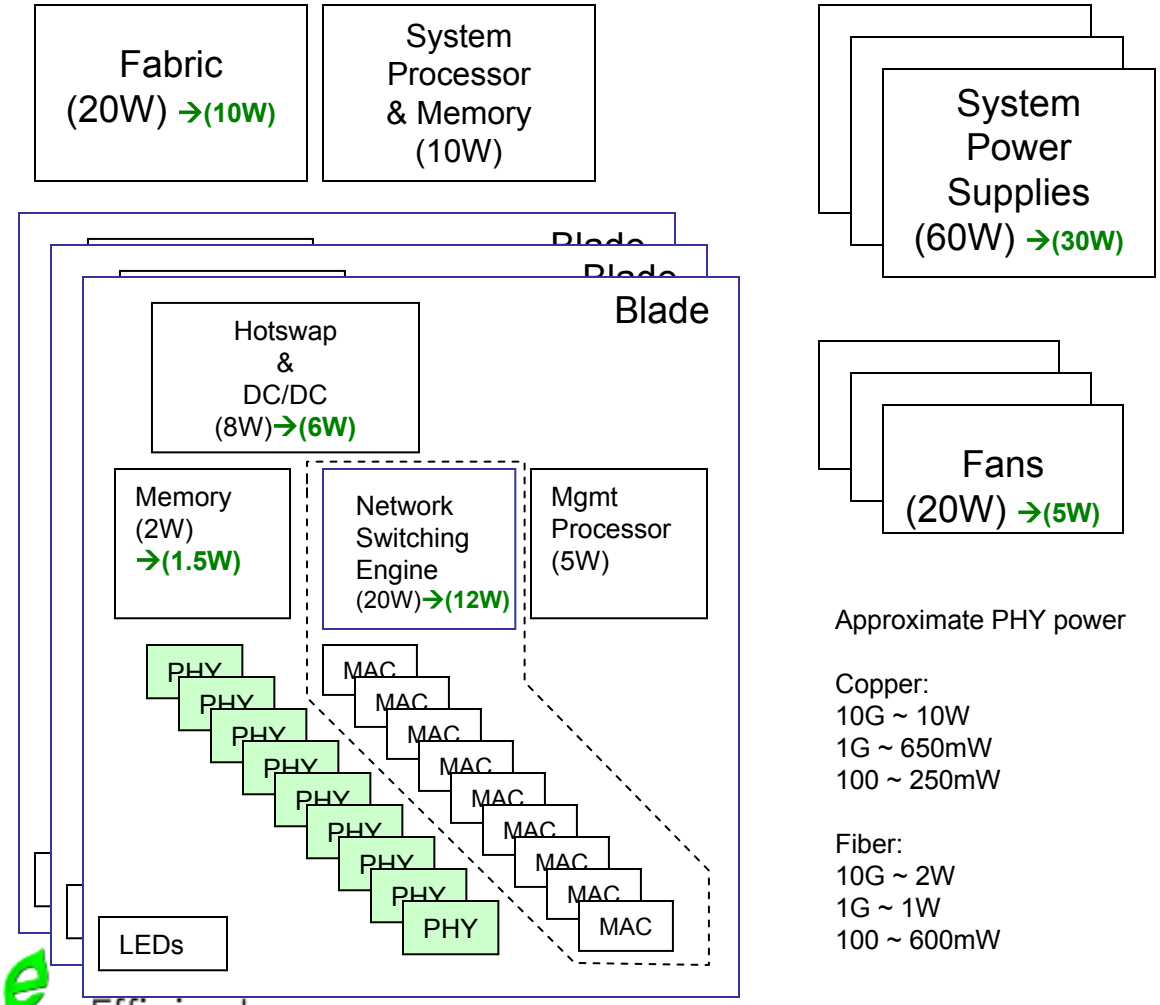
Fiber:
10G ~ 2W
1G ~ 1W
100 ~ 600mW

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 1000BASE-T will be used.

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

Switch Infrastructure (example)



Approximate PHY power

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

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

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 (N_IDLE) 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 may be negotiated at link startup

A means for negotiating N_IDLE is necessary to balance outbound buffering (of source) against inbound N_IDLE requirements (of receiver)

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

- LP_IDLE vs N_IDLE – A MAC may sleep when receiving LP_IDLE and will wake up and be prepared to process data when N_IDLE is received
- Note: If the MAC is a legacy device, the PHY can be configured to operate in "legacy mode" which would transmit LP_IDLE and N_IDLE based only on auto-negotiated parameters. The MAC/PHY interface would operate as currently defined

Control Considerations (cont)

The value of N_IDLE is a variable that must be communicated to the source of data on a link

- May be negotiated upon link startup
- May also be re-negotiated via LLDP or other means

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 N_IDLE may be desired after transmission to allow additional packets to be transmitted with minimum IPG once a receiving station has been awoken

Switches would be able to optimize buffer usage if guaranteed “off time” was provided

- MAC Control may be used to set minimum on time, guarantee minimum IPG
- In other words, Rate Control might help energy efficiency

Low Power IDLE Approach

Operating Modes

- Normal IDLE (N_IDLE) – Active IDLE used by technologies today, indicates “resume operation”
- Low Power IDLE (LP_IDLE) – Mode of operation where TX and RX power are reduced
- Refresh IDLE (R_IDLE) – Transmitted by PHY on occasion to refresh allow destination to refresh its RX parameters

Auto-Negotiate Key Parameters

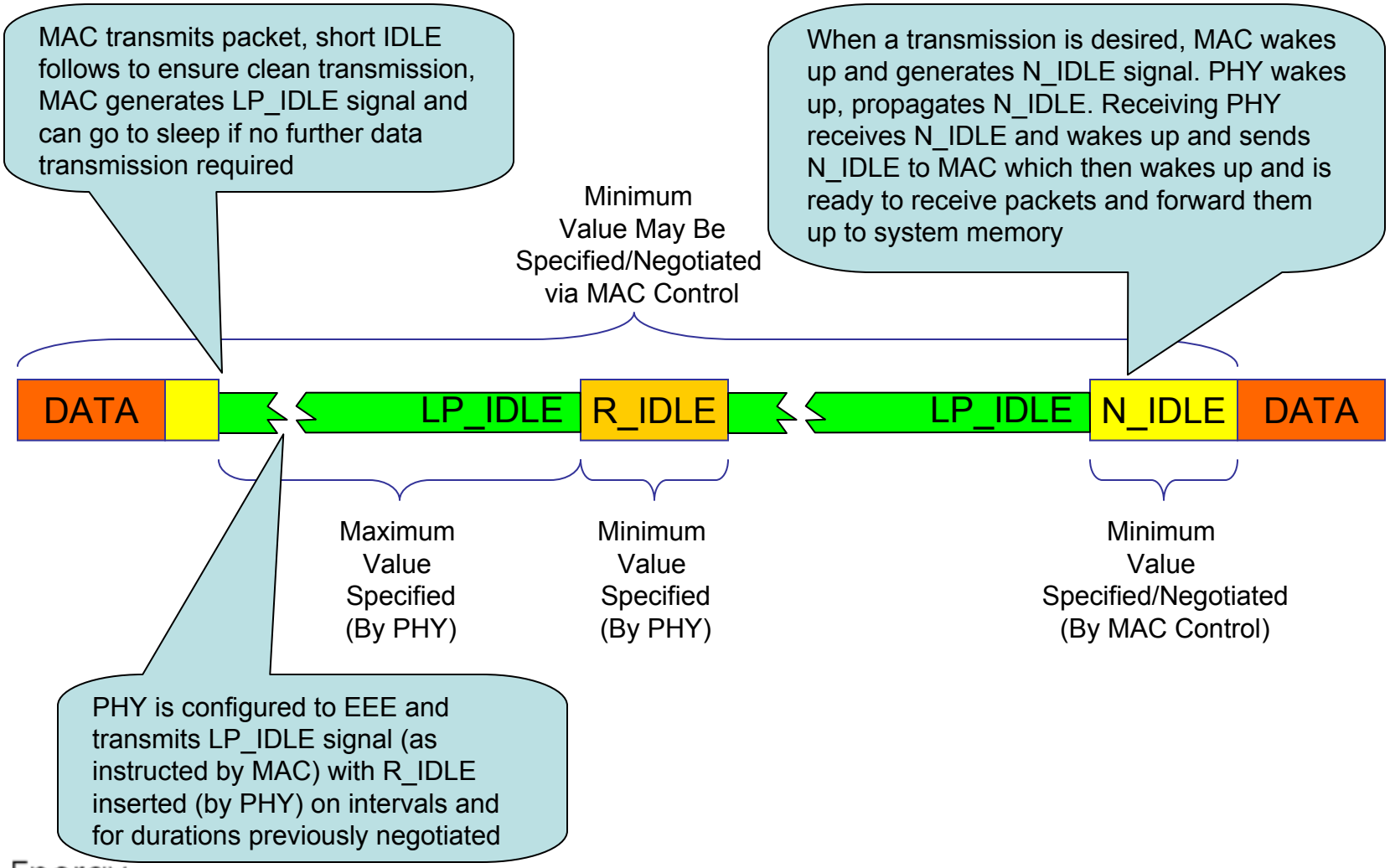
- Support for EEE
- Maximum LP_IDLE (PHY dependent)
- Minimum R_IDLE (PHY dependent)
- Minimum N_IDLE time (MAC dependent)

Low Power IDLE Approach (cont)

Communication Methods

- Auto-Negotiation of fundamental operating parameters (LP_IDLE_max, R_IDLE_min, N_IDLE_min)
- Renegotiation via LLDP if necessary (MAC may require longer N_IDLE than minimum to optimize efficiency)
- Signaling at the PHY layer using LP_IDLE, R_IDLE and N_IDLE
- Signaling from MAC to PHY to indicate LP_IDLE and N_IDLE states
- Signaling from PHY to MAC to indicate LP_IDLE and N_IDLE received
- MAC Control Frames to establish minimum IPG (If further optimization of energy mgmt required)
    - Rate Limiting would allow switches to shut down more circuitry and possibly not restore power/function to circuits if they knew that traffic would not return to line rate.
- We may want to consider/research use of LLDP to negotiate EEE level of functionality

Low Power IDLE Approach



MAC / PHY Signaling

XGMII

- Normal IDLE (N_IDLE)
- Low Power IDLE (LP_IDLE)

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	LP_IDLE	EEE Low Power IDLE
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	LP_IDLE	EEE Low Power IDLE
1	07	Idle	No applicable parameter (Normal inter-frame)

MAC / PHY Signaling

GMII

- Normal IDLE (N_IDLE)
- Low Power IDLE (LP_IDLE)

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	LP_IDLE	EEE Low Power IDLE
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	LP_IDLE	EEE Low Power IDLE
0	1	02 through 0D	Reserved	—

MAC / PHY Signaling

MII

- Normal IDLE (N_IDLE)
- Low Power IDLE (LP_IDLE)

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

Conclusions

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

Auto-Negotiate Key Parameters

- EEE Capable
- LP_IDLE_max
- R_IDLE_min
- N_IDLE_min (defined at AN, may be changed via LLDP or MCF)

Optimize MAC power savings by ensuring minimum IPG

- MAC Control may be used to ensure turn-off times of infrastructure components
- Enforcing data rates can be used to establish power savings as a priority

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

More work necessary to define extensions to extended interfaces like XAUI, TBI, R