3MG Consulting

An Overview of Energy-Efficient Ethernet

Michael Bennett

3MG Consulting

NANOG 61

June 4, 2014

Topics

Network Energy Efficiency

- Framing the Discussion
- Why *Network* Energy Efficiency?

Energy Efficient Ethernet

- Background and Definition
- How it works
  - Trade-offs
- EEE in optical networks

Opportunities for Development

Wrap-up

Disclaimer

Per IEEE-SA Standards Board Operations Manual:

> “At lectures, symposia, seminars, or educational courses, an individual presenting information on IEEE standards shall make it clear that his or her views should be considered the personal views of that individual rather than the formal position, explanation, or interpretation of the IEEE.”

- Any information I present on the topic of IEEE standards are my personal view and is not a formal position, explanation, or interpretation of the IEEE.

Network Energy Efficiency

Network Energy Efficiency is not the lowest hanging fruit

- Picking the low hanging fruit begins with a process
  - Make a Baseline Measurement
  - Analyze results
  - Make changes
- Changes that save lots of energy[1]
  - Optimize air flow, air handling and cooling systems
  - Use most efficient electrical systems, e.g. power supplies
  - Use most efficient ICT equipment

Why network energy efficiency?

- Common thinking
  - If the improvement isn't huge, it isn't worth pursuing
  - If you reduce energy use by a relatively small amount in a lot of places it adds up
  - To put it in perspective, global data center electricity used in 2010 accounted for roughly 1.1% to 1.5% of total electricity use [2]

1. http://hightech.lbl.gov/datacenters-bpg.html
2. http://www.koomey.com/post/8323374335

Network Energy Efficiency

How is network energy efficiency achieved?

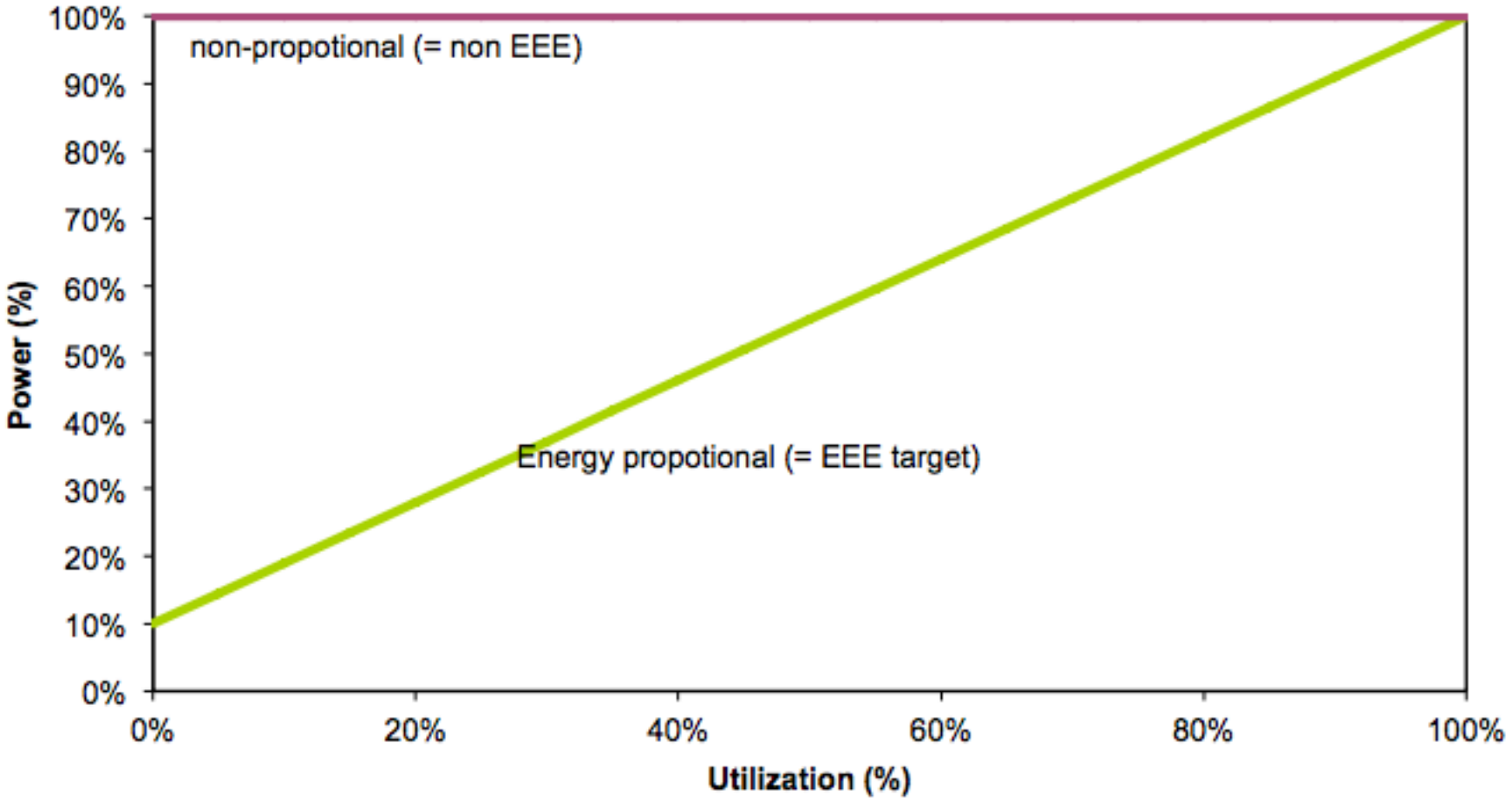
- Make the energy used proportional to useful work
  - When the load is high, power consumption increases
  - When the load is low, power consumption decreases

How much of the energy saved can come from the network?

- Network equipment consumes about 5% of data center electricity[1]
  - Roughly $225M in 2006, U.S. only
- Why would this matter?
  - Maybe it would be enough to achieve an incentive
  - EPA recommends electric utilities offer incentives to “facilitate Data Center energy efficiency requirements”[2]
    - New York State Energy Research and Development Authority offers up to $10M to fund energy efficiency intiatives[3]

1. http://www.energystar.gov/ia/partners/prod_development/downloads/EPA_Datacenter_Report_Congress_Final1.pdf?0e1b-1681
2. http://www.energystar.gov/index.cfm?c=prod_development.data_center_efficiency_info
3. http://www.nyserda.ny.gov/Commercial-and-Industrial/Sectors/Data-Centers.aspx

Network Energy Efficiency



1. http://www.ieee802.org/3/400GSG/public/13_07/diab_400_01b_0713.pdf

Why Network Energy Efficiency?

Even in high transaction-rate networks, utilization is not 100% 24 hours/day, 365 days/year[1]

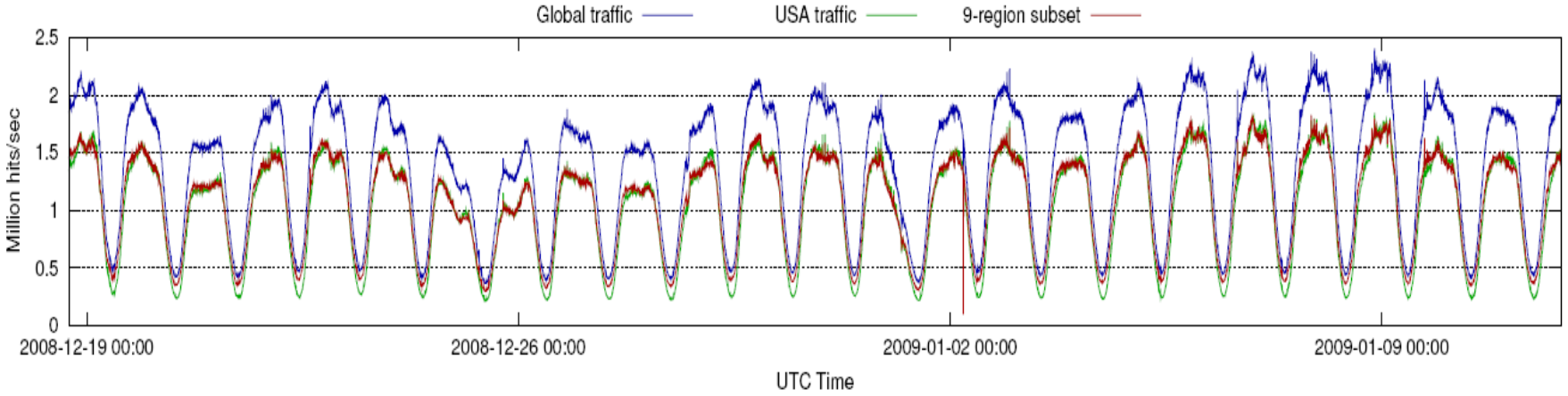


Figure 14: Traffic in the Akamai data set. We see a peak hit rate of over 2 million hits per second. Of this, about 1.25 million hits come from the US. The traffic in this data set comes from roughly half of the servers Akamai runs. In comparison, in total, Akamai sees around 275 billion hits/day.

1. Cutting the Electric Bill for Internet-Scale Systems, Qureshi et. al, SIGCOMM '09 Proceedings of the ACM SIGCOMM 2009 conference on Data communication, ISBN: 978-1-60558-594-9

Why Network Energy Efficiency?

Energy cost is still a significant operational expense in data centers [1]

Company	Servers	Electricity	Cost
eBay	16K	$\sim 0.6 \times 10^5$ MWh	~$3.7M
Akamai	40K	$\sim 1.7 \times 10^5$ MWh	~$10M
Rackspace	50K	$\sim 2 \times 10^5$ MWh	~$12M
Microsoft	>200K	$> 6 \times 10^5$ MWh	>$36M
Google	>500K	$> 6.3 \times 10^5$ MWh	>$38M
USA (2006)	10.9M	610×10^5 MWh	$4.5B
MIT campus		2.7×10^5 MWh	$62M

1. Cutting the Electric Bill for Internet-Scale Systems, Qureshi et. al, SIGCOMM '09 Proceedings of the ACM SIGCOMM 2009 conference on Data communication, ISBN: 978-1-60558-594-9. Estimated annual electricity costs for large companies (servers and infrastructure) @ $60/MWh (6 cents / KWh)

Why Network Energy Efficiency?

Savings for the IEEE 802.3az PHY alone should be around 90% and energy reduced by up to 70% for the NIC when in Low Power Idle mode[1].

- much greater savings possible in systems using LLDP
  - See dove_02_05_08.pdf (slide 5)

1. P. Reviriego, K. Christensen, J. Rabanillo, and J. A. Maestro, 'An Initial Evaluation of Energy Efficient Ethernet' in IEEE communications letters, VOL. 15, NO. 5, May 2011

Energy Efficient Ethernet (EEE)

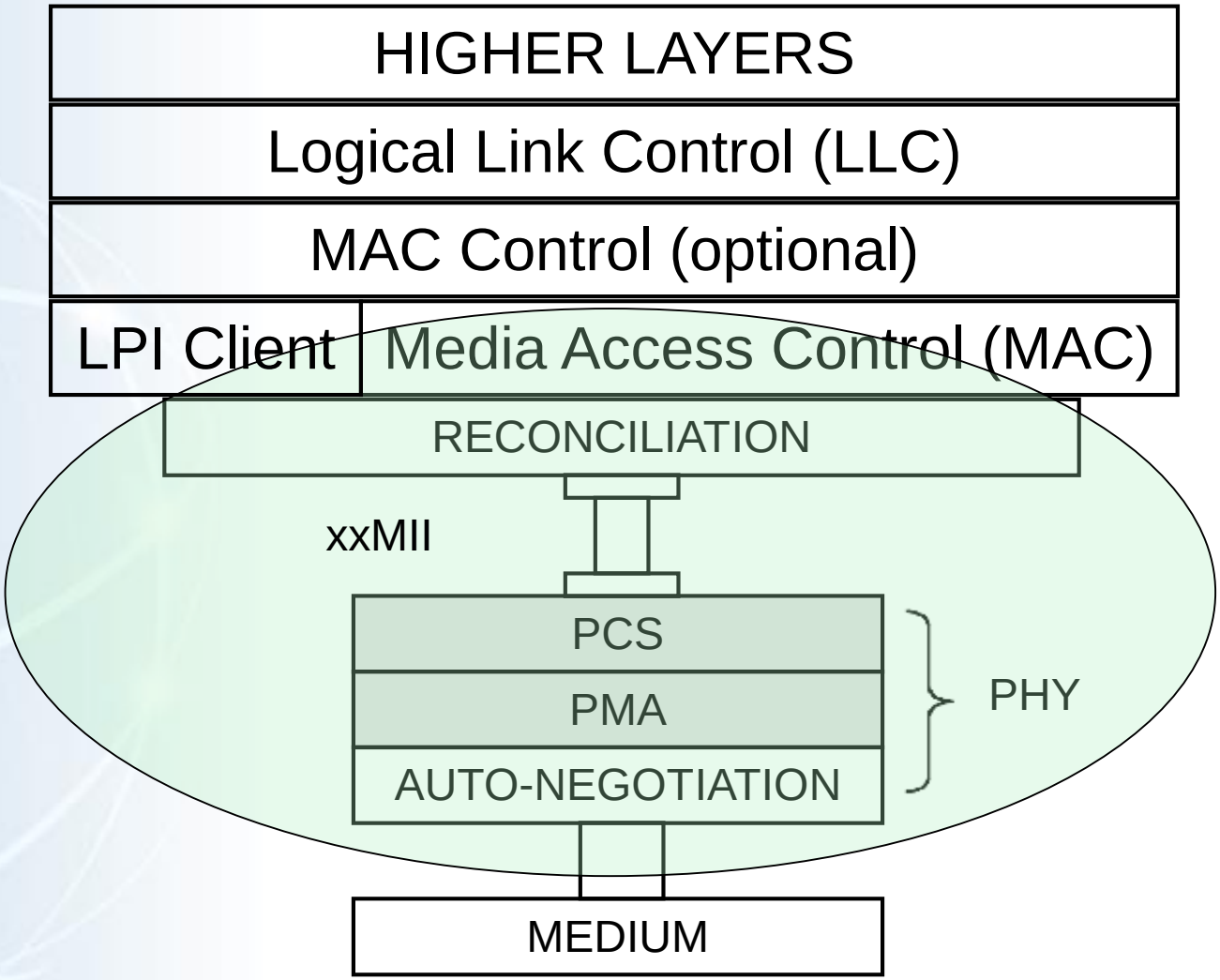
Background

- Energy Efficient Ethernet is a method scale energy used by an Ethernet device with link utilization
- Specified in Std 802.3az-2012 Energy-efficient Ethernet amendment
  - now part of IEEE Std 802.3-2012™
- The premise for EEE is that Ethernet links have idle time providing opportunity to save energy
- Specified for copper interfaces
  - “BASE-T’s
  - Backplane (except 40G)
    - 40G, 100G backplane, next gen optics and 400G in progress
- The method is called Low Power Idle (LPI)
  - For optical Ethernet > 10G, new concept of “fast wake”

Energy Efficient Ethernet (EEE)

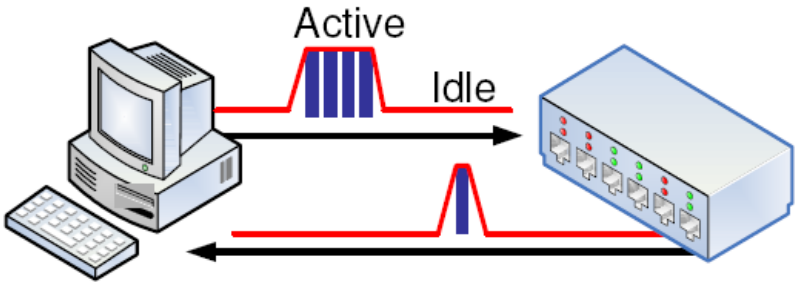
- How does EEE achieve energy-efficiency?
  - For electrical energy:
  - $E_t = [P_{active} * T_{active}] + [P_{idle} * T_{idle}]$
    - E_t is the amount of energy consumed over a period of time
    - P is power and T is time
    - Active is when data is transferred (useful work)
    - Idle is when signals are sent, but no data
- Higher speeds during data transmission use less energy
    - Smaller bit times
    - Good to go to higher speeds from an energy perspective

Energy Efficient Ethernet (EEE)

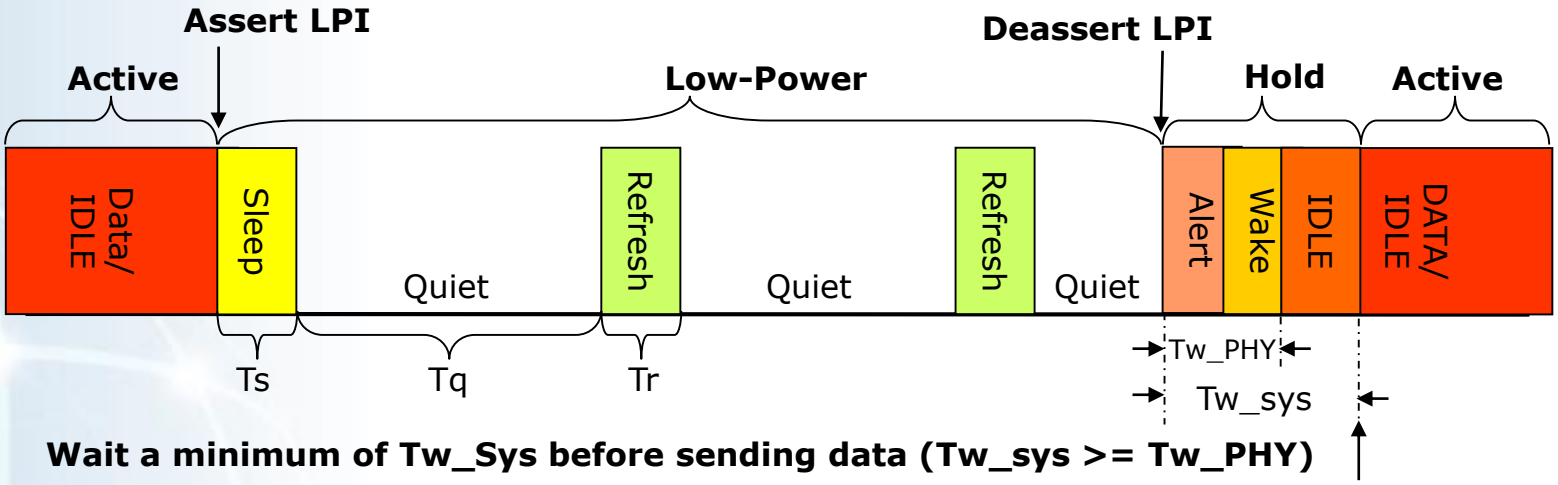


How it Works

- Concept: Transmit data as fast as possible, return to Low-Power Idle
- Saves energy by cycling between Active and Low Power Idle
  - –Power reduced by turning off unused circuits during LPI
  - –Energy use scales with bandwidth utilization



How it Works



LPI – PHY non-essential circuits shut down during idle periods

During power-down, maintain coefficients and sync to allow rapid return to Active state

Wake times (Tw_PHY) for Twisted-Pair PHYs:

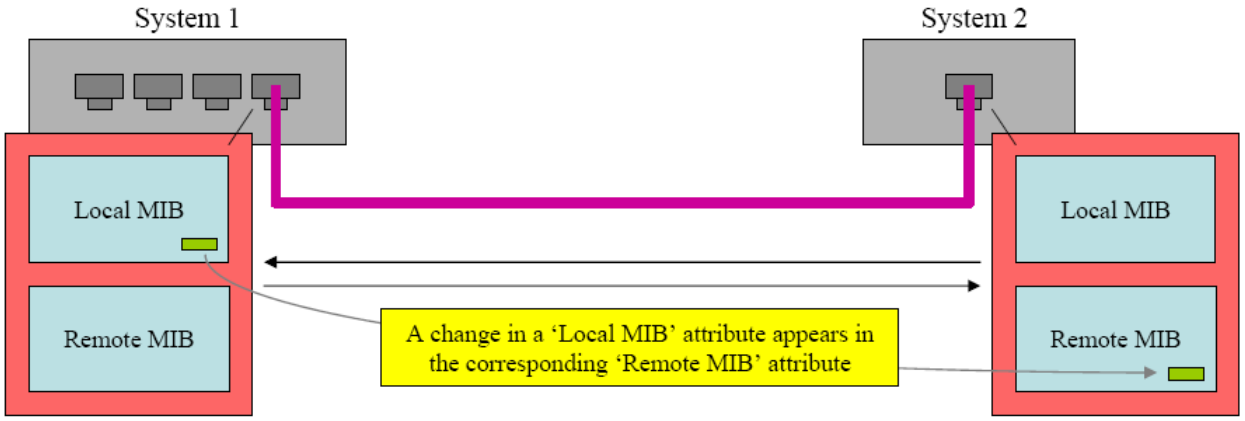
- 100BASE-TX: <= 30 usec
- 1000BASE-T: <= 16.5 usec
- 10GBASE-T: <= ~8 usec (2 modes)

How it Works

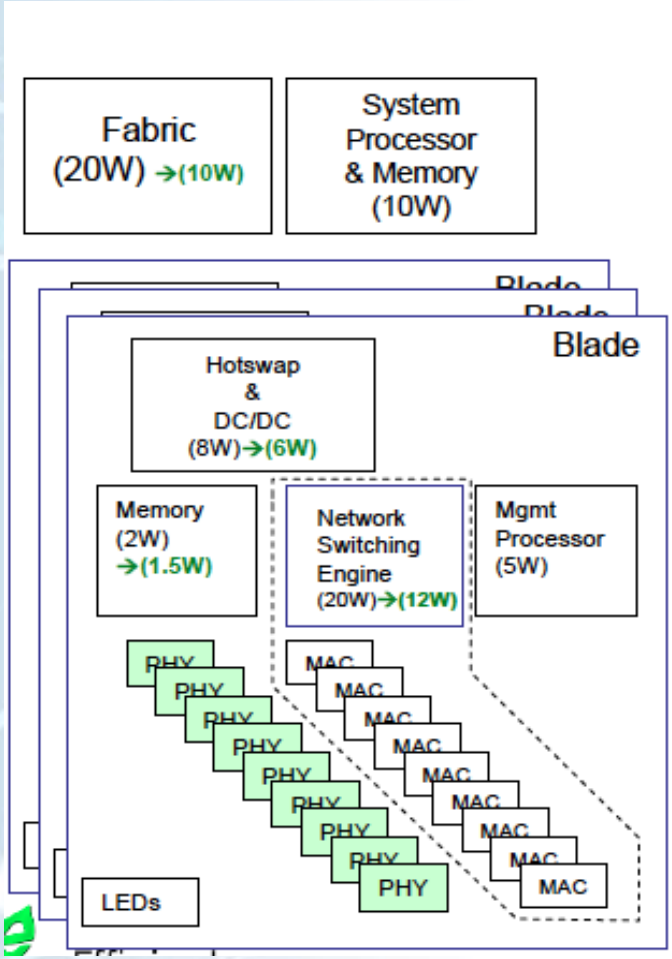
Uses auto-negotiation to notify link partner of EEE capabilities

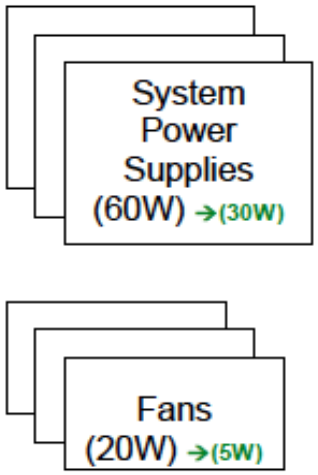
Uses Link Layer Discovery Protocol (LLDP) to notify link partner of parameter changes

- E.g. control policy
  - User can choose energy savings over performance or vice versa



Why Network Energy Efficiency?





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

http://www.ieee802.org/3/az/public/may08/dove_02_05_08.pdf

Trade-offs

Latency

- The trade-off for energy-savings is latency
  - As energy use decreases, latency increases
- Currently there is a “latency arms race”
  - Narrow market segment (financial)
  - Lots of $ driving it
  - EEE likely not a feature for use with ultra-low latency networks
- How much does latency increase with EEE?
  - Single digit microseconds for 10G
  - Hundreds of nanoseconds for 40G and 100G

EEE for Optical Networks

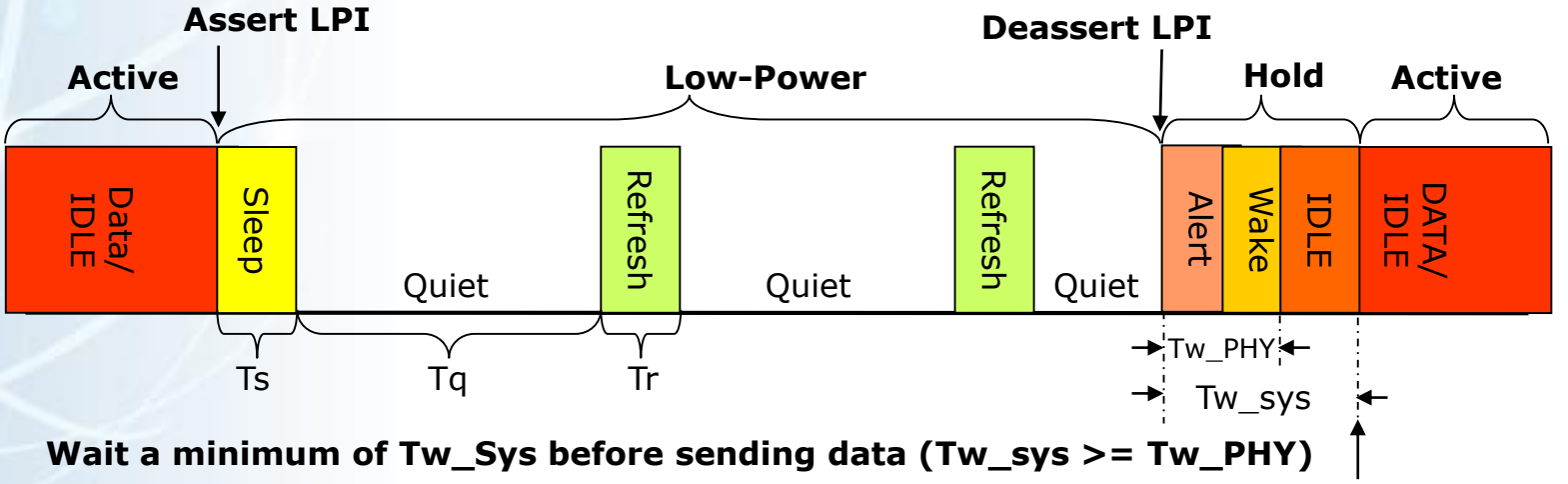
- EEE is currently being specified for 40G and 100G backplanes and optical Ethernet
  - Cycling between Active and Low Power Idle may be problematic for optical transmitters
  - Introduces the concept of “fast wake” and “deep sleep”

- Current proposal for EEE
  - Signal LPI without shutting off the transmitter
    - Enables Network Equipment Manufacturers to use the LPI signal to power off system components during idle periods
    - Need to be careful when used with OTN networks
  - Not as many parameters
    - No sleep time, quiet time, or refresh time since the transmitter is not powered off

EEE for Optics

Possibly This:

Instead of this:



EEE for Optical Networks

- Potential energy savings depends on creative *system* energy management
  - When the link transitions from Active to “fast wake”, turn off non-essential *system* circuits and transceiver except the laser
  - Recall the system diagram
    - It’s the only way there will be network energy savings for optical Ethernet

- EEE will be applicable for any optical link as long as there are no extreme low latency requirements

EEE for Optical Networks

- When will EEE for Optical networks be available?
  - Current timeline has the standard completed in 2015[1]
- The tools are being created
  - The next thing that has to happen is for manufacturers to make them useful
  - Users need to ask for the features
  - Plenty of room for innovation

http://www.ieee802.org/3/bm/timeline_0912.pdf

Opportunities for Development

- EEE is being developed in some form for many different IEEE projects
  - IEEE P802.3bp, Single Twisted Pair Gigabit Ethernet
    - Applications include automotive and industrial
  - IEEE P802.3bm, Next Generation Optical 100 Gigabit Ethernet
  - IEEE P802.3bj, 40 and 100 Gigabit Backplane Ethernet
  - IEEE P802.3bq, 40 Gigabit Ethernet on Twisted Pair (40GBASE-T)
  - IEEE P802.3bs (not a typo), 400 Gigabit Ethernet

Opportunities for Development

- Fast-wake is either being specified or considered in each of those projects
  - In order for EEE to achieve maximum impact, network equipment manufacturers must design their systems to take advantage of LPI
  - We need to make systems like the one in Dan Dove's example a reality
- It's also good to go back and reexamine EEE
  - Can it be improved?
  - This is being done in 400GbE

Opportunities for Development

- The 400 Gigabit Ethernet project is going to specify a new PHY for a new speed
  - Considering to shut down lanes to save energy
  - For example, if you only need 200G on a link, shut down 2 lanes (in a 4x100G lane design)
- The problem of ever-increasing power consumption must be solved
  - In Joel Georgen's presentation, *400GE Electrical Interface Thoughts*, he mentioned if we keep building systems the way we have been, the next generation systems will approach 100 KW per rack!
  - He suggests using a different approach to reduce SERializer-DESerializer (SERDES) power
  - This may lead to a new form of EEE for chip-to-chip applications

Wrap-up

- EEE will improve the energy efficiency of ICT equipment
  - The amount of energy that can be saved depends on the amount of time the network is not fully utilized
    - Make the network energy-use proportional to useful work
  - The trade-off for improved energy efficiency is an increase in latency
    - In the order of single digit microseconds or less
- EEE for optical networks coming in 2015

Wrap-up

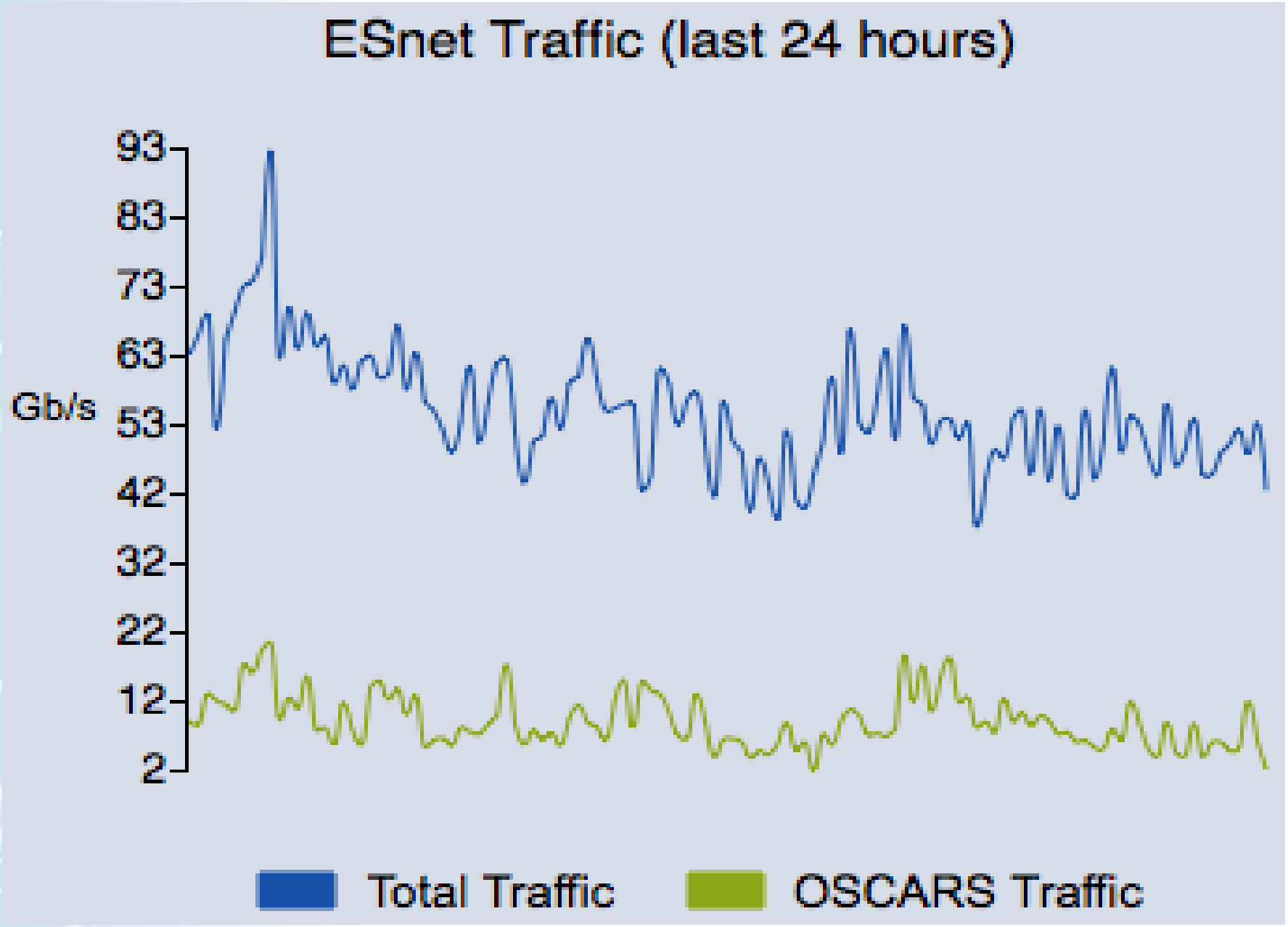
- Possible improvements to EEE coming for 400GbE
- Overall, good things happening for EEE
- End users need to demand more efficient network equipment to motivate NEMs to make it happen

3MG Consulting

Questions?

Thanks!

Michael Bennett – mjbennett@ieee.org


ESnet Traffic (last 24 hours)
Gb/s
93
83
73
63
53
42
32
22
12
2
Total Traffic
OSCARS Traffic

ESnet Top Traffic

Networking for the Future of Science.