Ethernet: An Energy-efficient Technology

International Workshop on Energy Efficiency and Networking

IMDEA

31 May 2010

Michael J. Bennett
Lawrence Berkeley National Laboratory
Chair IEEE P802.3az Task Force

These are my personal views

- Per IEEE-SA Standards Board Operations Manual, January 2005:
- “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.”
- The views expressed in this presentation are mine and not that of the IEEE or Ethernet Alliance

Topics

- Rationale for Energy-efficient Ethernet
- IEEE P802.3az development and status
  - Objectives
  - Timeline
  - Enhanced Energy Savings via Layer 2
  - Overview of Solutions
- Example applications
- Non-IEEE developments related to EEE
- Possible future development of Ethernet and energy efficiency

Rationale – Macro View

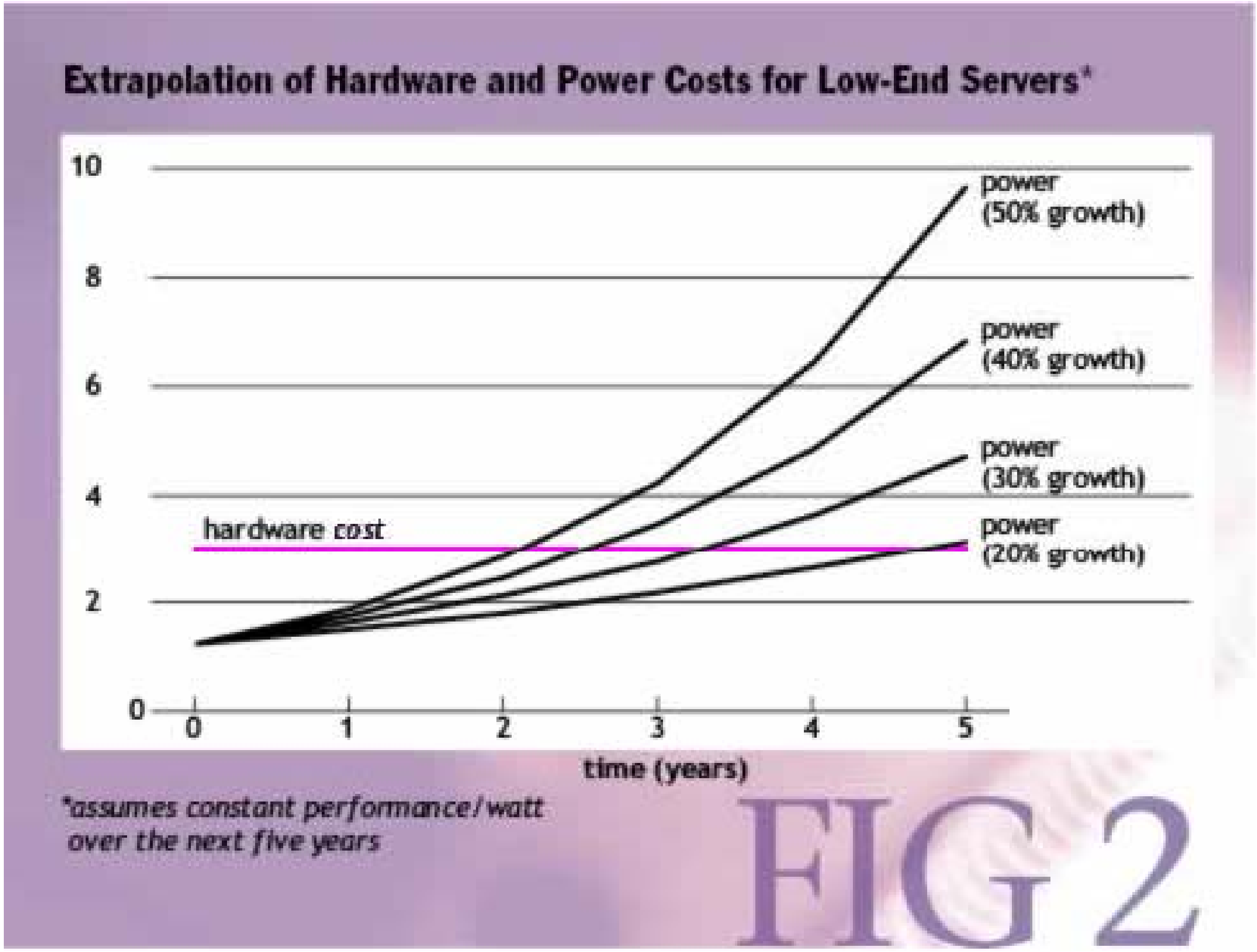
- "Big IT" – all electronics
  - PCs/etc., consumer electronics, telephony
    - Residential, commercial, industrial
  - 200 TWh/year
  - $16 billion/year
  - Nearly 150 million tons of CO_2 per year
    - Roughly equivalent to 30 million cars!

PCs etc. are digitally networked now — *Consumer Electronics* (CE) will be soon

Numbers represent U.S. only

One central baseload power plant
(about 7 TWh/yr)

Rationale – A Micro View



Unrestrained IT power consumption could eclipse hardware costs and put great pressure on affordability, data center infrastructure, and the environment.

Source: Luiz André Barroso, (Google) "The Price of Performance," *ACM Queue*, Vol. 2, No. 7, pp. 48-53, September 2005.

(Modified with permission.)

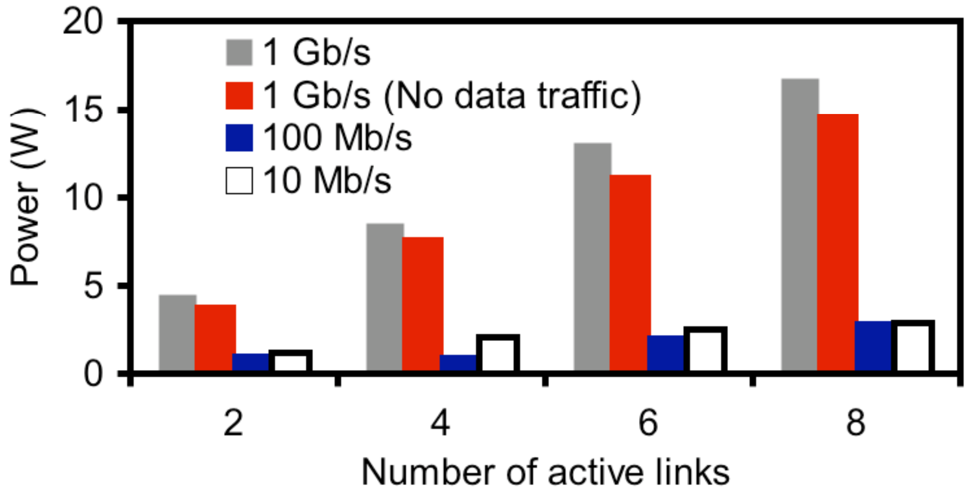
Industry and Regulatory Trends

- Government and Industry Recognition
  - April, 2006 “Green Grid” formed
  - December, 2006 U.S. House Resolution 5646 signed into law
  - European code of conduct
  - Japanese government initiative “Top Runner”
    - July, 2009 – routers and switches added as “target products” with target fiscal year 2011
- IEEE P802.3az – Energy Efficient Ethernet
  - Work began “officially” November, 2006
  - Cu based Ethernet interfaces will go green

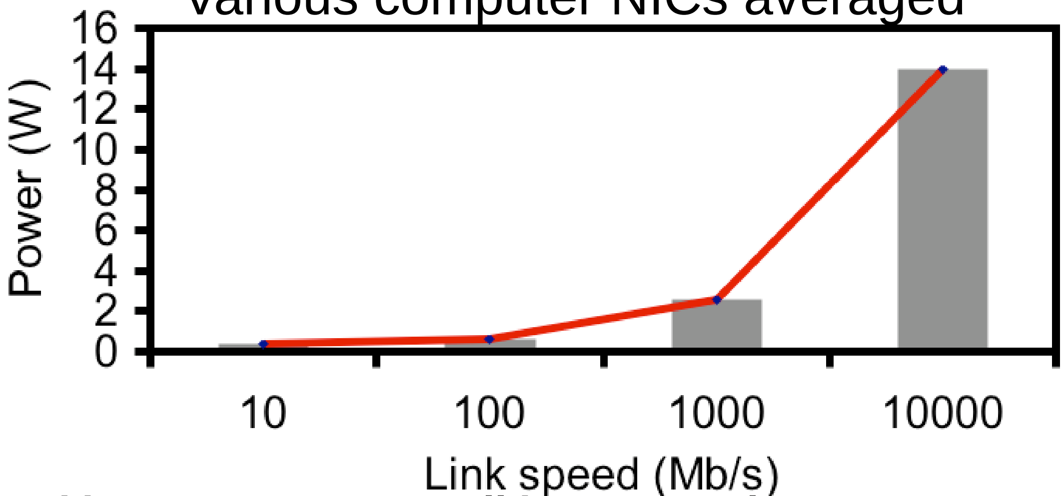
Rationale – A Link Perspective

- High port count triple speed switches
  - Linear relationship of power consumption to number of active links
  - Aggregate savings attractive in putting inactive links in LPI
- Low port count 10G systems
  - Idle power savings on a single link attractive

Typical switch with 24 ports 10/100/1000 Mb/s



Various computer NICs averaged



Results from 1st order (rough) measurements – all incremental *AC* power

Rationale – A Link Perspective

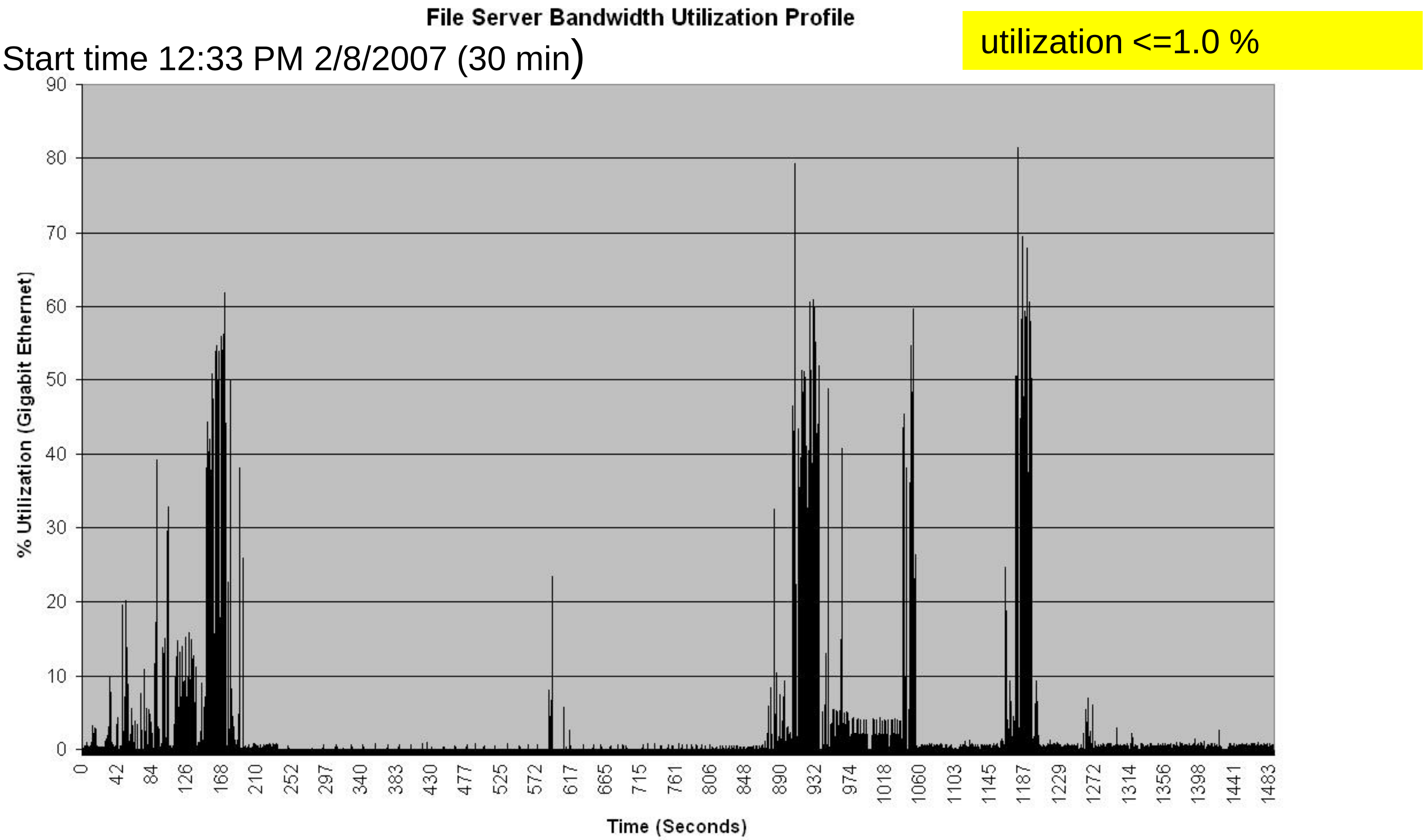
A data point on 10GBASE-T power (estimated Q1, 2008)

- Single Port 10GBASE-T Adapter
  - 3W – Single Port XAUI ASIC Controller
  - 5W – Teranetics PHY
  - 2W – Power management & miscellaneous
  - 10W – Total Single Port 10GBASE-T Adapter
- Dual Port 10GBASE-T Adapter
  - 3W – Dual Port XAUI ASIC Controller
  - 10W (2x5W) – Teranetics PHY
  - 2W – Power management & miscellaneous
  - 15W – Total Dual Port 10GBASE-T Adapter

Source: http://www.ieee802.org/3/eee_study/public/mar07/kohl_01_0307.pdf

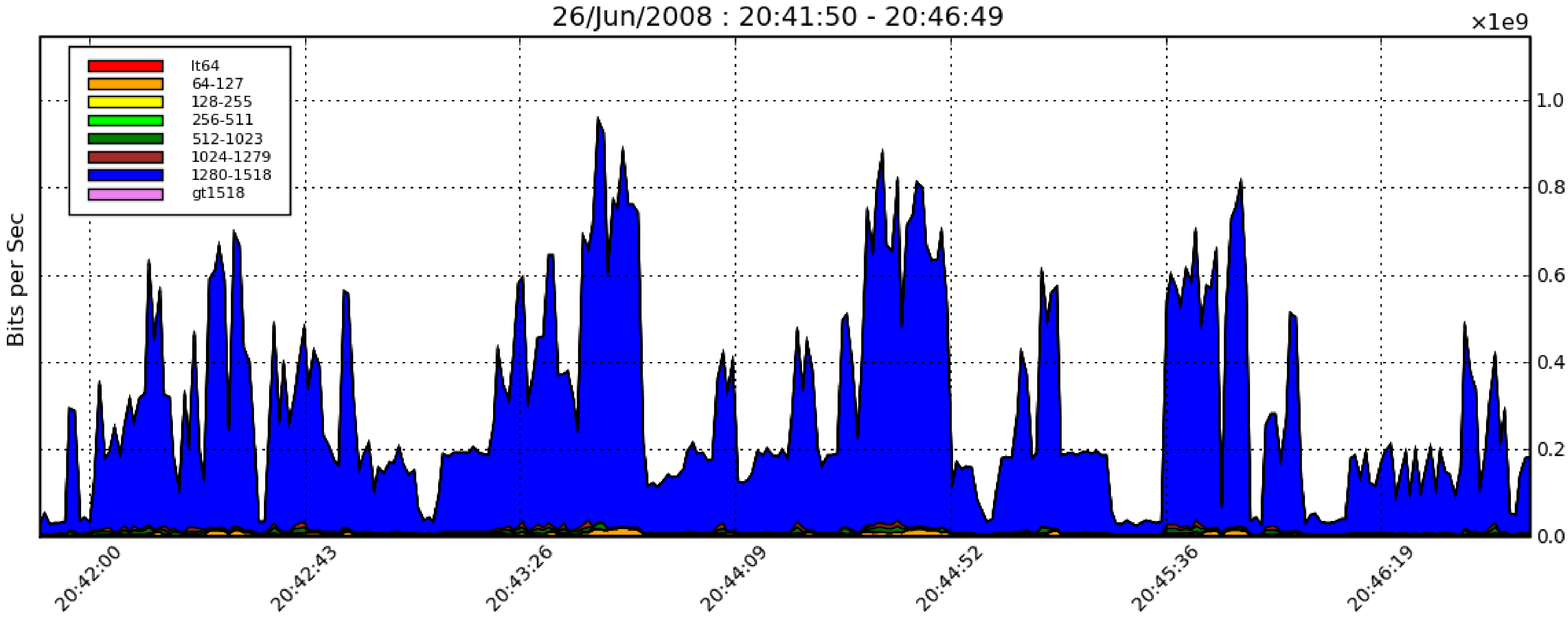
Ethernet Traffic Profiles

- Snapshot of a File Server with <u>1 Gb</u> Ethernet link
  - Shows time versus utilization (trace from LBNL)



Ethernet Traffic Profiles

- Snapshot of a 10 Gb Ethernet link



10

Ethernet Traffic Profiles

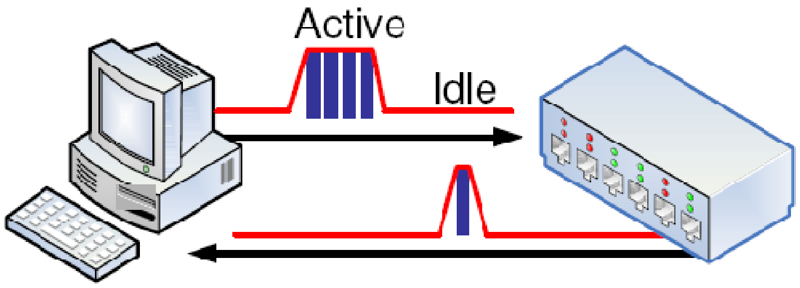
- Traffic profiles used for the Energy Efficient Ethernet Study Group were primarily enterprise-centric
  - Lawrence Berkeley National Laboratory Campus Area Network
  - Abilene backbone
  - Intel enterprise network
- We found it difficult to get data center traces
  - Some work on EEE after we selected the technology used data center traffic and indeed found opportunities to achieve even more efficiencies – more about that later

What is Energy-efficient Ethernet?

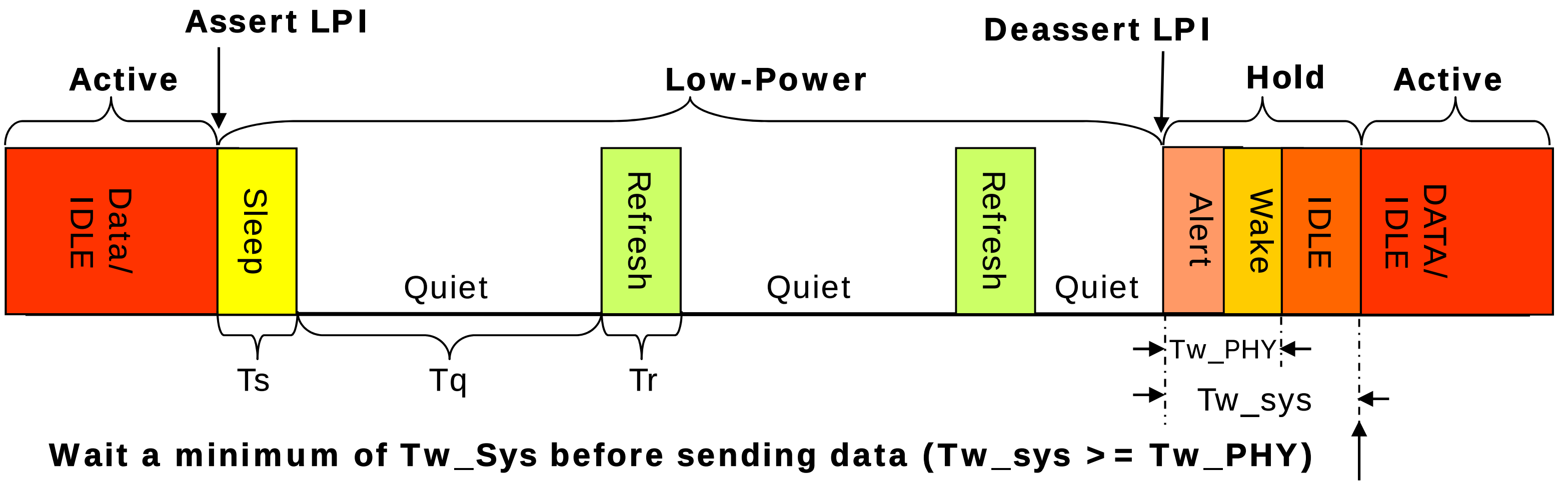
- Energy Efficient Ethernet (EEE) is a method to reduce energy use by an Ethernet device during periods of low link utilization
- The premise for EEE is that Ethernet links are under utilized
- Specified for copper interfaces
  - “BASE-T’s’
  - Backplane
- The method we’re using is called Low Power Idle

What is Low Power Idle?

- 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



Low Power Idle Overview



- Low Power Idle (LPI) – PHY powers down during idle periods
- During power-down, maintain coefficients and synchronization to allow rapid return to Active state
- Wake times for the respective twisted-pair PHYs:
  - 100BASE-TX: Tw_PHY <= 20.5 usec
  - 1000BASE-T: Tw_PHY <= 16.5 usec
  - 10GBASE-T: Tw_PHY < ~8 usec
- PHY power in LPI mode ~20-40% of normal (depends on type and implementation)

What is 10BASE-Te?

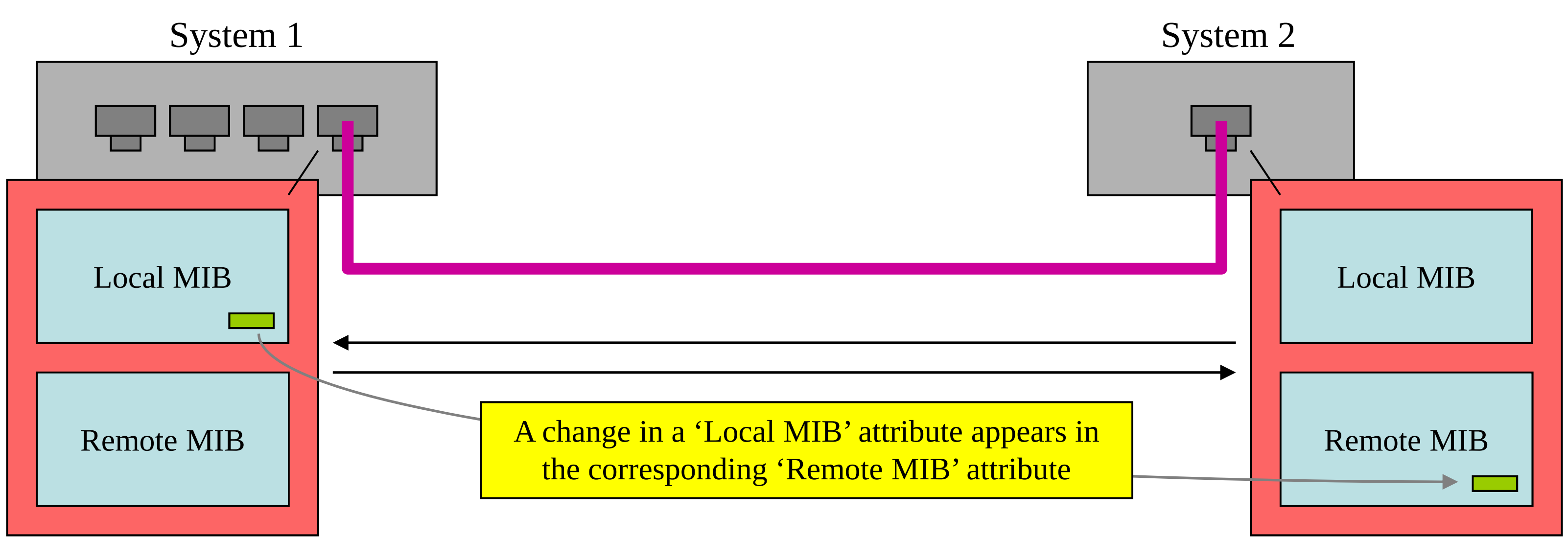
- The original definition for 10BASE-T specified transmission of large voltage swings to overcome the losses of Class C cable
- 802.3az specifies 10BASE-Te with a smaller transmission voltage that is compatible with legacy 10BASE-T on any cable that has Class D (or better) characteristics.
  - Provides for energy savings
  - Allow manufacturers to use the latest high density processes that will save power on multi-speed devices
  - Enables a reduction in voltage supplies / convertors

Optimizing Energy Efficiency

- Energy Efficiency can be optimized by using link-partner communications after the link is established
  - Use Link Layer Discovery Protocol (LLDP) to change wake times.
  - The longer the wake time, the longer the delay till frames can pass, i.e. latency variation increases
  - Trade-off between energy savings and latency
- There are system power savings opportunities in addition to PHY power

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)



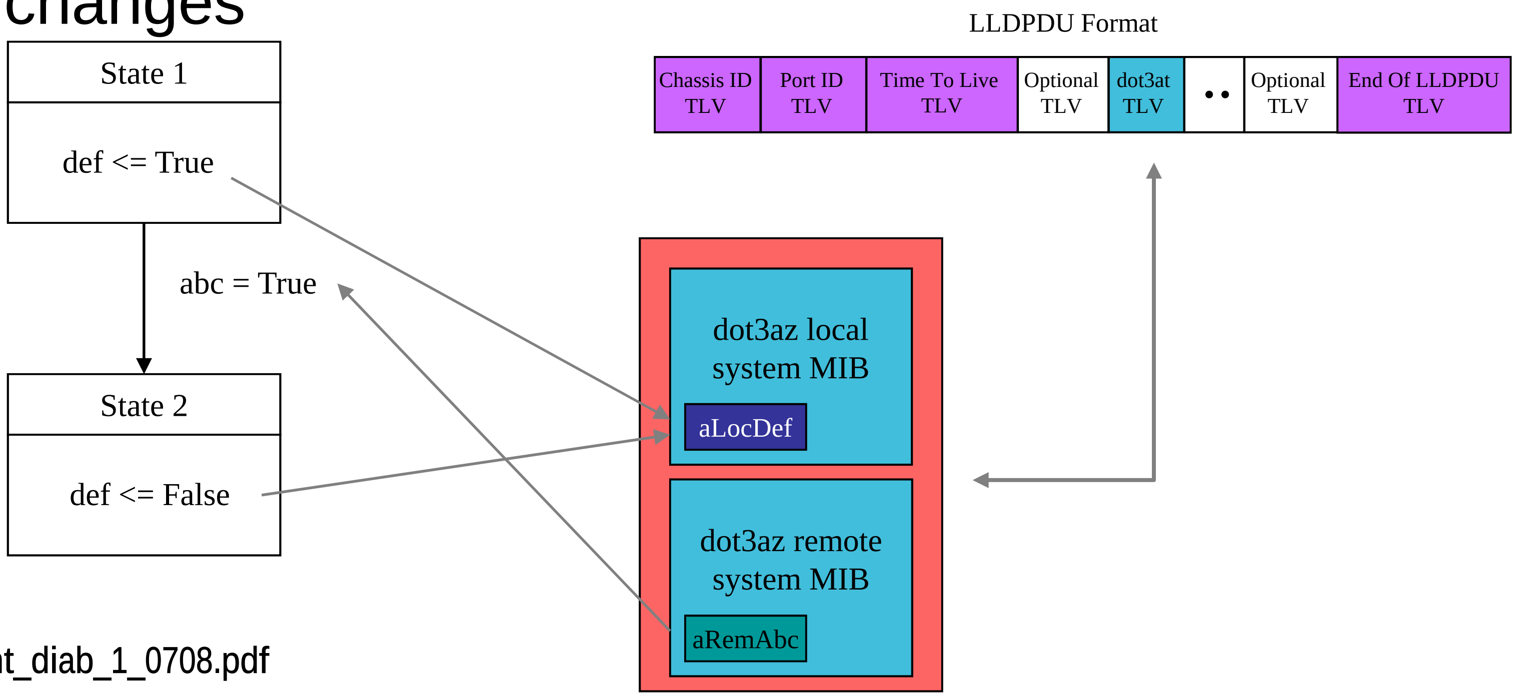
Source: law_01_0508.pdf

802.3az - Layer 2

- 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 – E.g. Fallback states
- Mandatory for 10G and above speeds. Optional for lower speeds
- 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 the RX to turn off more circuitry when it goes to sleep as it knows it has additional time from when the PHY is woken up

LLDP and State diagrams

- Can't map directly to type-length-value (TLV) contents
  - Map through objects in dot3az local and remote MIB
  - Define MIB attribute to variable mapping
  - Allows .3 layers to take action based on variable changes



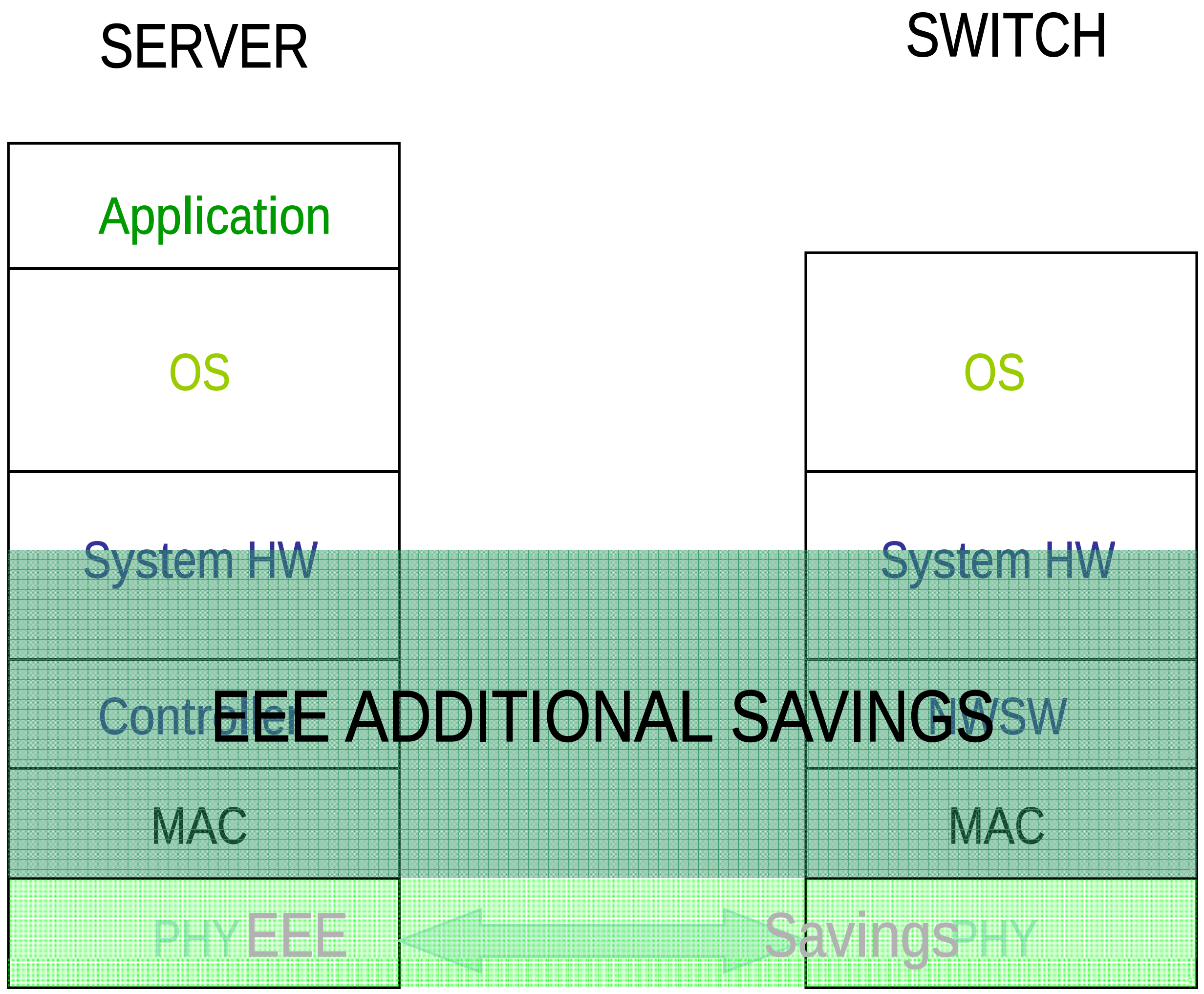
Source: joint_diab_1_0708.pdf

Energy Efficient Ethernet TLV

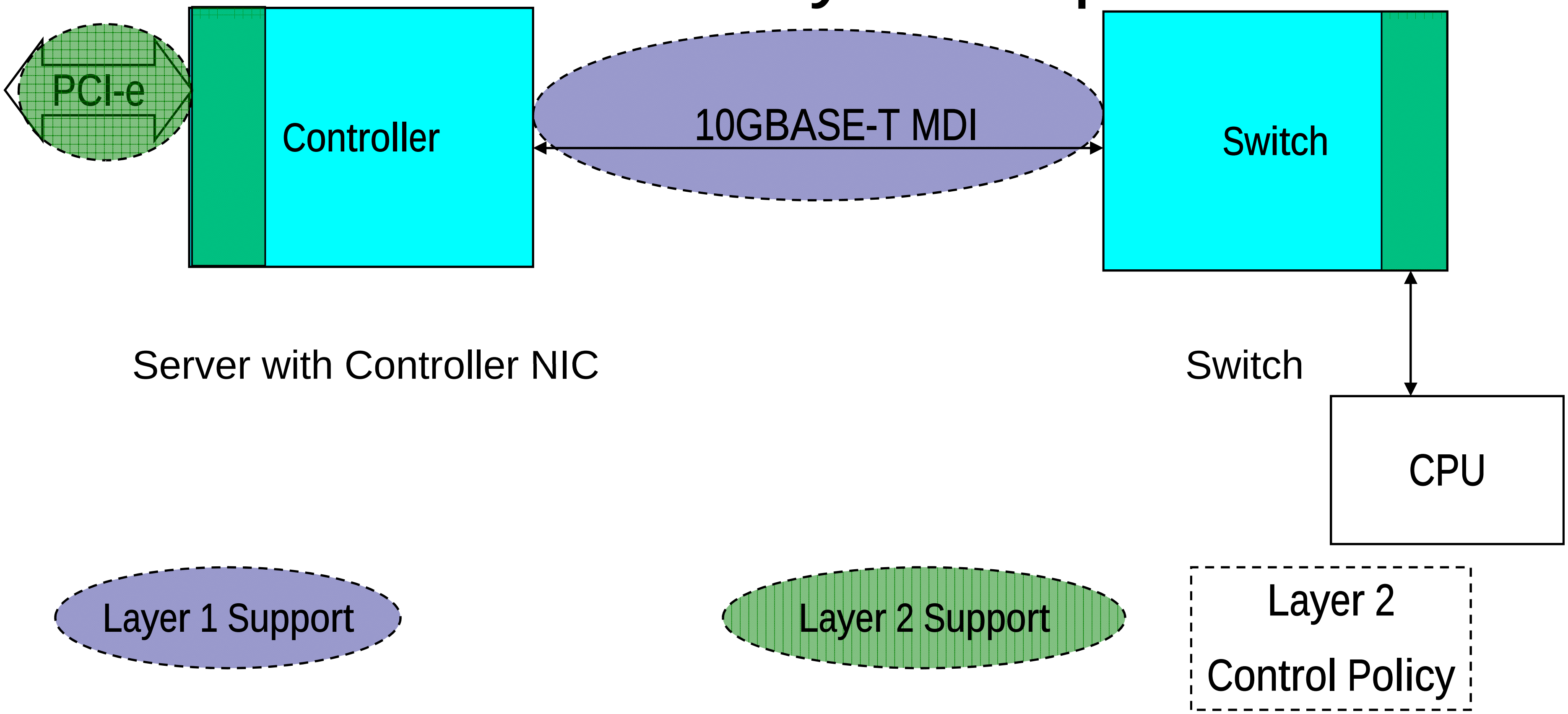
TLV Type = 127	TLV information String length = 12	802.3 OUI 00-12-0F	802.3 Subtype = TBA	…
7 bits	9 bits	3 octets	1 octet	
TLV header		TLV information string		

…	TX Tw_sys	RX Tw_sys	Echo TX Tw_sys	Echo RX Tw_sys
	2 octets	2 octets	2 octets	2 octets
	TLV information string (continued)			

Example EEE End-to-End Savings

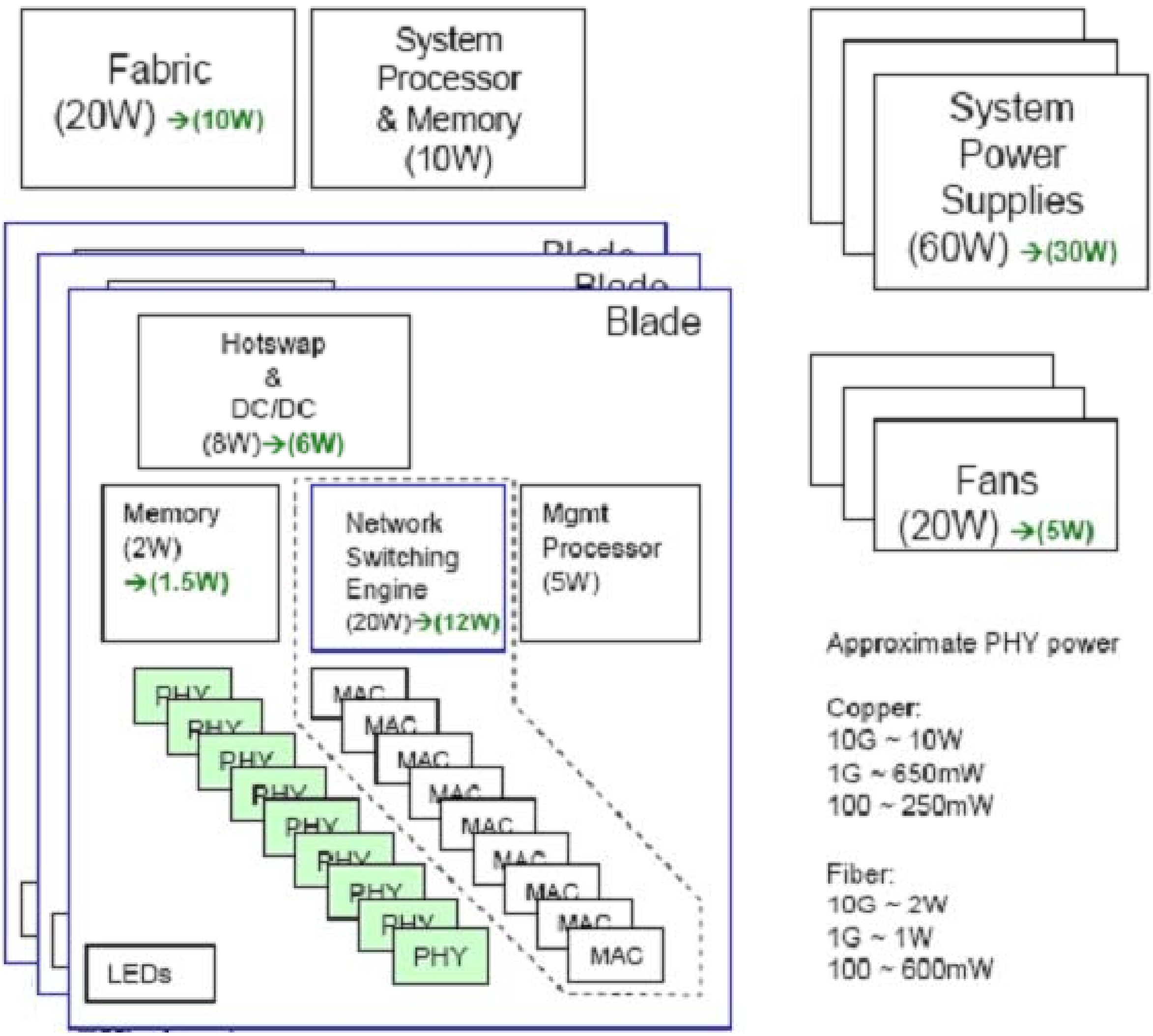


EEE Enhanced Layer 2 Operations



- Opportunity to save additional power within a box (link partner)
  - Additional circuits beyond the PHY can be turned off
- Additional RX wakeup time negotiated using 802.3az’s Layer 2 --- *standards based*

A system view (switch centric)



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

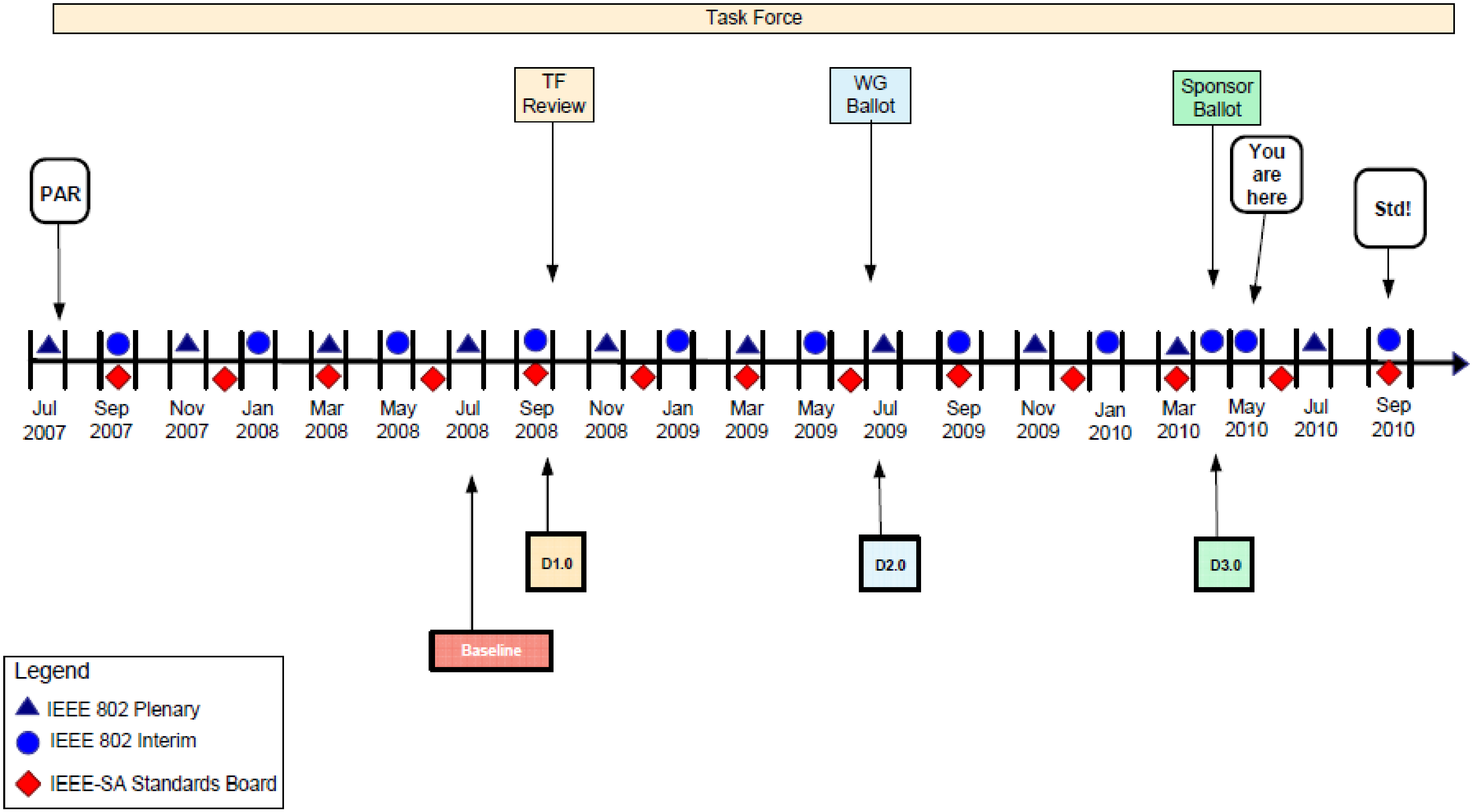
EEE Objectives

- Define a mechanism to reduce power consumption during periods of low link utilization for the following PHYs
  - 100BASE-TX (Full Duplex)
  - 1000BASE-T (Full Duplex)
  - 10GBASE-T
  - 10GBASE-KR
  - 10GBASE-KX4
  - 1000BASE-KX
- Define a protocol to coordinate transitions to or from a lower level of power consumption
- The link status should not change as a result of the transition
- No frames in transit shall be dropped or corrupted during the transition to and from the lower level of power consumption
- The transition time to and from the lower level of power consumption should be transparent to upper layer protocols and applications

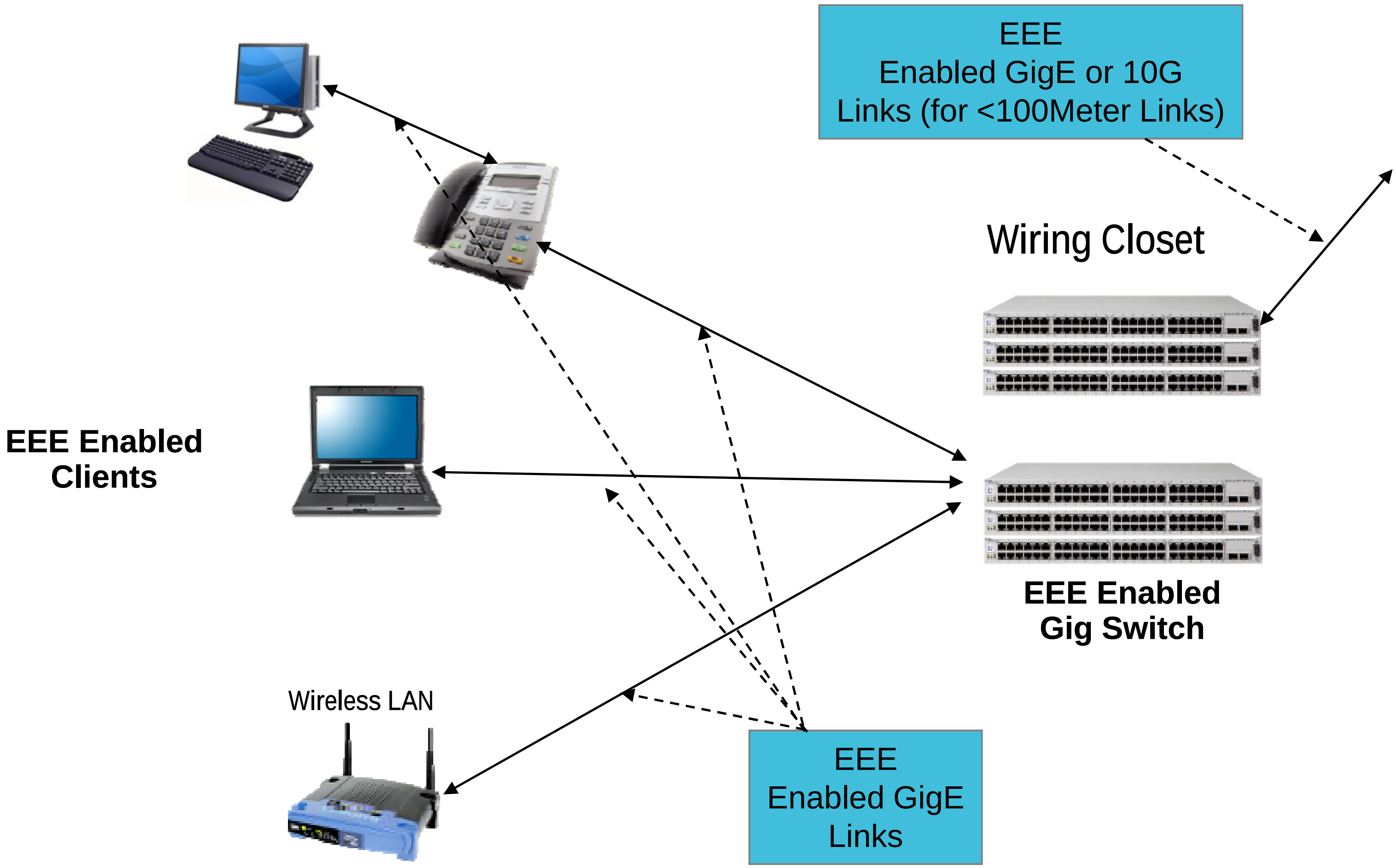
EEE Objectives

- Define a 10 megabit PHY with a reduced transmit amplitude requirement such that it shall be fully interoperable with legacy 10BASE-T PHYs over 100 m of Class D (Category 5) or better cabling to enable reduced power implementations
- Any new twisted-pair and/or backplane PHY for EEE shall include legacy compatible auto negotiation

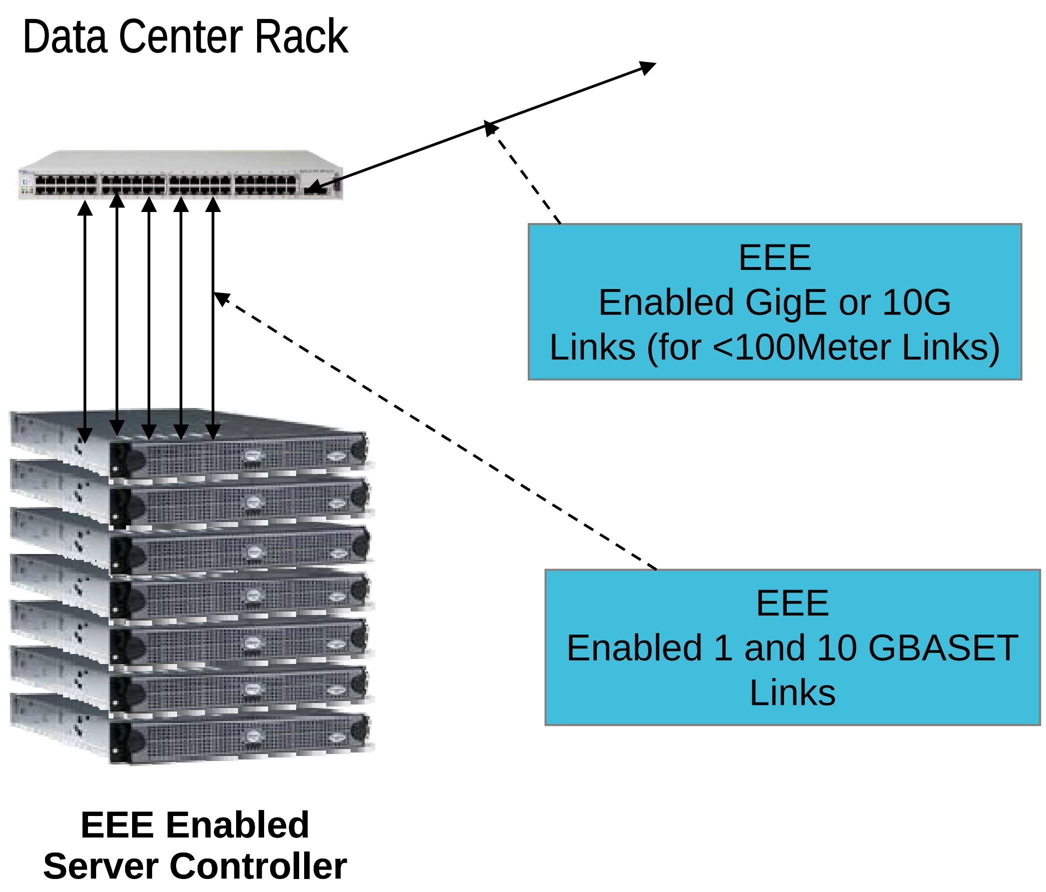
IEEE P802.3az Task Force Timeline



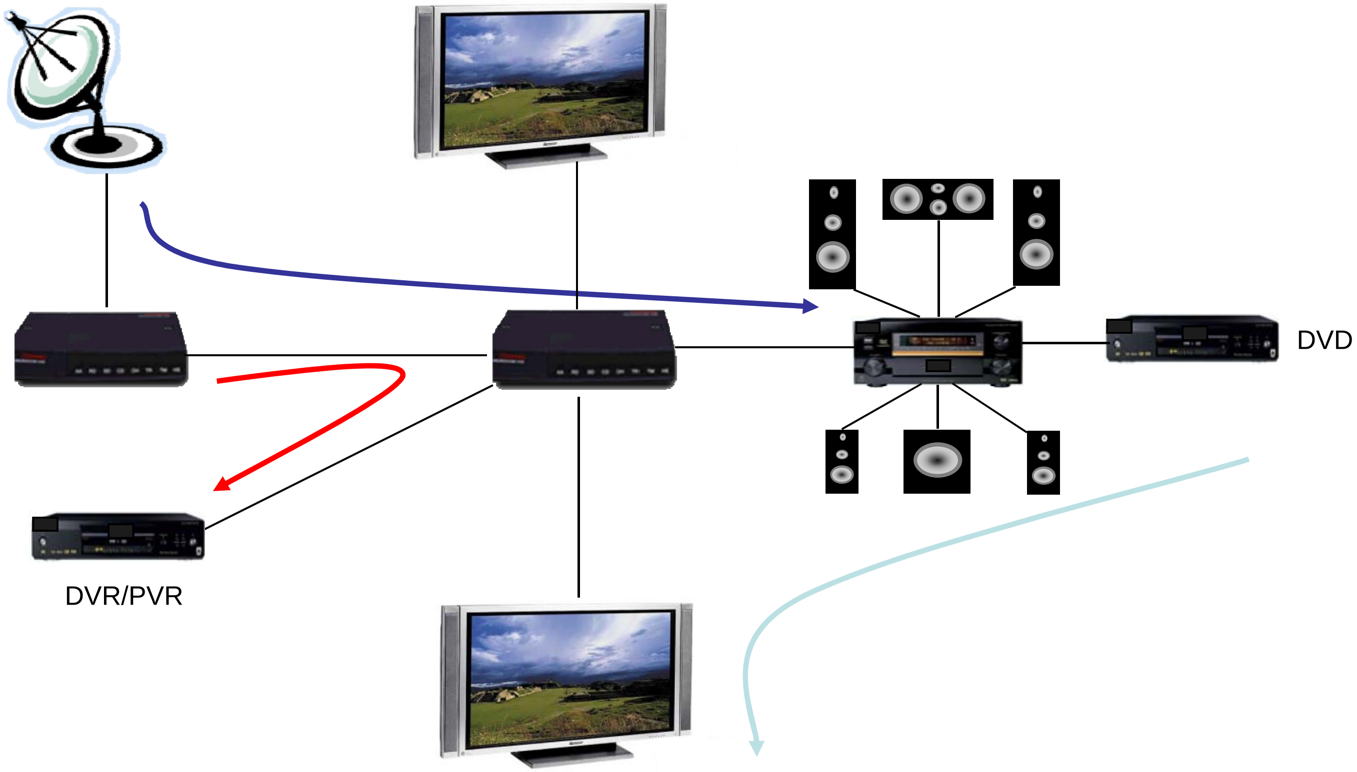
Application – Wiring Closet



Application – Data Center and TOR



Application - EAV home network



1. Listening to satellite radio on Ethernet AV receiver, link between receiver and switch
2. Start playing DVD on a screen in another room
3. DVR/PVR set to record favorite show from satellite receiver at 8:00 pm on Thursday

Non-IEEE developments related to EEE

- In October, 2009 Riviergo, et. al. published a paper examining the energy efficiency of P802.3az (Draft 1.2.1)
- They observed that the wake and sleep times are high compared to the time it takes to transmit a frame. The performance was analyzed using a variety of traffic profiles *including traces from data centers*
- The analysis suggested there could improvements in efficiency by buffering and bursting frames
- This kind of work is essential to the development of control policies to maximize energy savings

Non-IEEE developments related to EEE

- Further work by Ken Christensen, et. al. examines the trade-offs in performance of energy-efficient Ethernet
- Builds from the work described in the previous paper
- Suggests a packet coalescence mechanism

Possible future developments

- How do we continue the energy-efficiency effort begun in P802.3az?
  - *In my opinion*, one way to do this would be add an energy-efficiency component to the economic feasibility criterion
    - That could take some time, if it happens at all, so what to do in the mean time?
  - The EPA (Energy Star) will look to IEEE 802.3 for guidance for their requirements on new versions of Ethernet
    - Incentives are good for the market

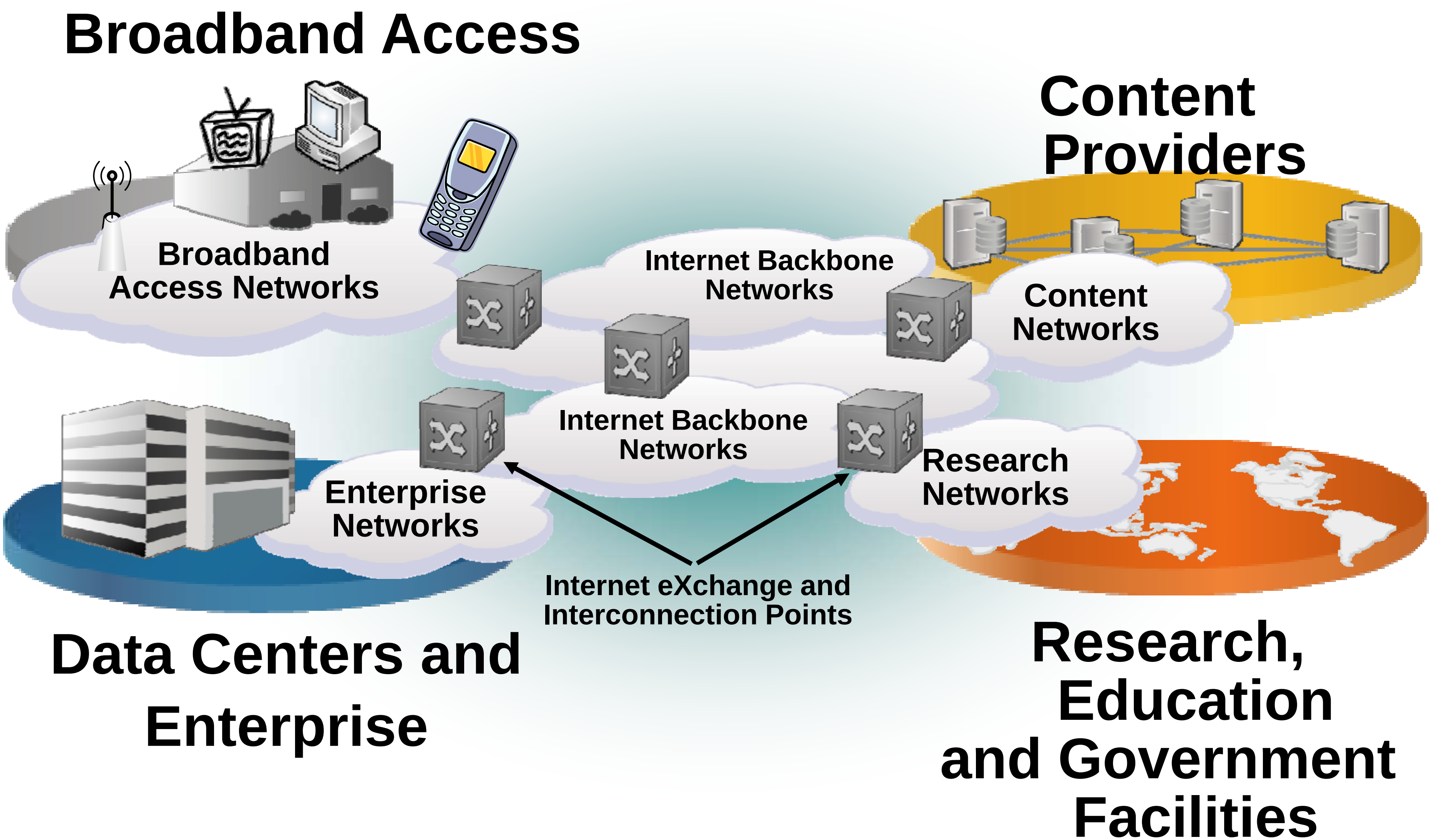
Possible future developments

- Optical Ethernet
  - Optical PHYs were not studied during the study group phase of the EEE. The following need to be studied:
    - Potential for energy savings
    - Whether or not lasers can be turned off (completely) and on
      - Any adverse affects?
    - Time to transition between states

Possible future developments

- Considerations for new Ethernet projects
  - Characterize the traffic
    - Defines opportunity to save energy
  - Minimize need to increase buffers/latency
  - Maximize energy savings
  - Remain transparent to upper layers
  - Ability to communicate changes after the link is up
    - LLDP

Possible future developments



Possible future developments

- Questions:
  - Is LPI the best method to achieve energy efficiency for future Ethernet technologies?
    - P802.3az serves most of the ecosystem, however
      - What are the utilization rates for ISP links and data centers?
        - » Translates to “how much energy can be saved?”

Possible future developments

- Questions:
  - One possible alternative could be Rapid PHY Selection (RPS)
    - Changes speed with demand
      - e.g. Drops to a lower speed (uses less energy) when the demand decreases
    - Can system energy savings be achieved with RPS?
  - Is there a better approach we haven't thought of?

Possible future developments

- There may be opportunities to discover a better approach to achieving energy efficiency in new projects

- Impact of other technologies
  - How will virtualization/consolidation affect traffic?
  - How will latency-sensitive applications work with EEE?

Summary

- Energy-efficient Ethernet will save energy
  - At the physical layer
  - In the system
- There are trade-offs for saving energy
  - Latency variation vs. energy use
- There are opportunities to develop the work done in P802.3az
  - Improvements in efficiency
  - Control policy and network management
  - Optical and higher speed Ethernet

Thank You!

References

- B. Nordman, Digital Networks: http://efficientnetworks.lbl.gov/enet.html
- K. Christensen, et.al., "IEEE 802.3az: The Road to Energy Efficient Ethernet," submitted to *IEEE Communications*, March 2010.
- W. Diab Use of LLDP, http://www.ieee802.org/3/az/public/jan09/diab_02_0109.pdf
- H. Barrass, et.al., AVB and EEE p.30 http://www.ieee802.org/802_tutorials/07-July/IEEE-tutorial-energy-efficient-ethernet.pdf
- Dove, Energy Efficient Ethernet: A switching Perspective http://www.ieee802.org/3/az/public/may08/dove_02_05_08.pdf
- P. Reviriego, J.A. Hernandez, D. Larrabeiti, J. A. Maestro, IEEE COMMUNICATIONS LETTERS, VOL. 13, NO. 9, SEPTEMBER 2009
- P. Reviriego, et.al., Reduce latency in energy efficient Ethernet switches with early destination lookup, http://www.embeddedinternetdesign.com/design/224400698
- P802.3az public page, http://ieee802.org/3/az/index.html