Energy Efficient Ethernet Overview

Mike Bennett

Lawrence Berkeley National Laboratory

ITU-T SG15

Geneva, Switzerland

February 15, 2008

Disclaimer

- The information you're about to hear is my own point of view and does not represent an official position of the IEEE

The bigger picture

- LBNL's work on Energy Efficient Ethernet is a part of the Energy Efficient Digital Networks project
- Goal:
  - This project aims to reduce electricity use of electronics through a variety of methods, all with the common theme of digital networks.
- Sponsors:
  - California Energy Commission (CEC)
    - Public Interest Energy Research (PIER) Program
  - U.S. Environmental Protection Agency (EPA)
    - ENERGY STAR Program
- Website: http://efficientnetworks.lbl.gov/

Discussion

- What is Energy Efficient Ethernet?
- Background
- IEEE 802.3az Status Report

What is Energy Efficient Ethernet?

- A method to reduce energy use by an Ethernet interface.
  - This will be accomplished by facilitating transitions to and from lower power consumption in response to changes in network demand.
- Based on works of Dr. Ken Christensen from University of South Florida and Bruce Nordman from LBNL
  - Known as Adaptive Link Rate (ALR)
    - *Ethernet Adaptive Link Rate: System Design and Performance Evaluation*, Gunaratne, C.; Christensen, K.; Proceedings 2006 31st IEEE Conference on Local Computer Networks, Nov. 2006 Page(s):28 - 35

Background

The problem

- **Office equipment, network equipment, servers**
  - 97 TWh/year
    - **3% of national electricity**
    - **9% of commercial building electricity**
    - **Almost $8 billion/year**

Numbers represent U.S. only

	TWh/year
– **Network Equipment**	**13**
– **Servers**	**12**
– **PCs / Workstations**	**20**
– **Imaging (Printers, etc.)**	**15**
– Monitors / Displays	22
– UPS / Other	16

60% Networked Equipment

- **… However**
  - **Old data (energy use has risen)**
  - **Doesn't include residential IT or networked CE products**

Note: Year 2000 data taken from Energy Consumption by Office and Telecommunications Equipment in Commercial Buildings--Volume I: Energy Consumption Baseline Roth et al., 2002 Available at: http://www.eren.doe.gov/buildings/documents

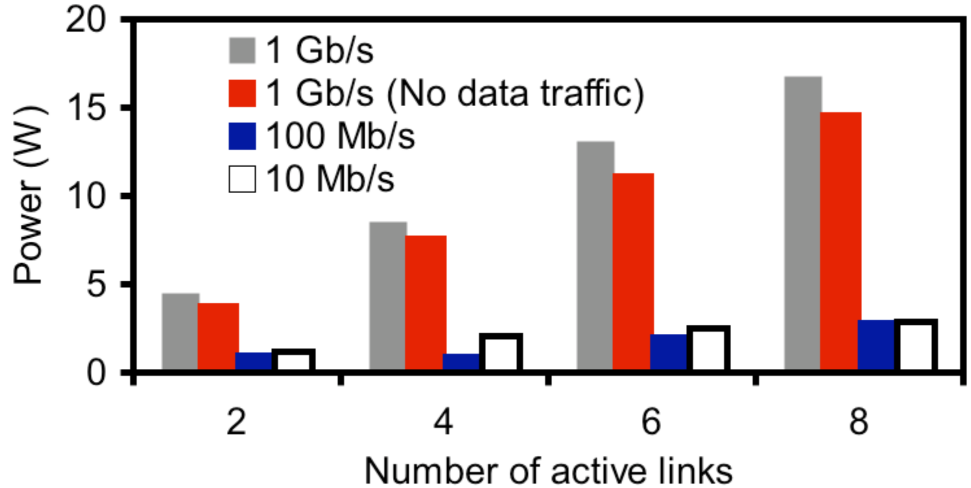
Link power

Results from (rough) measurements

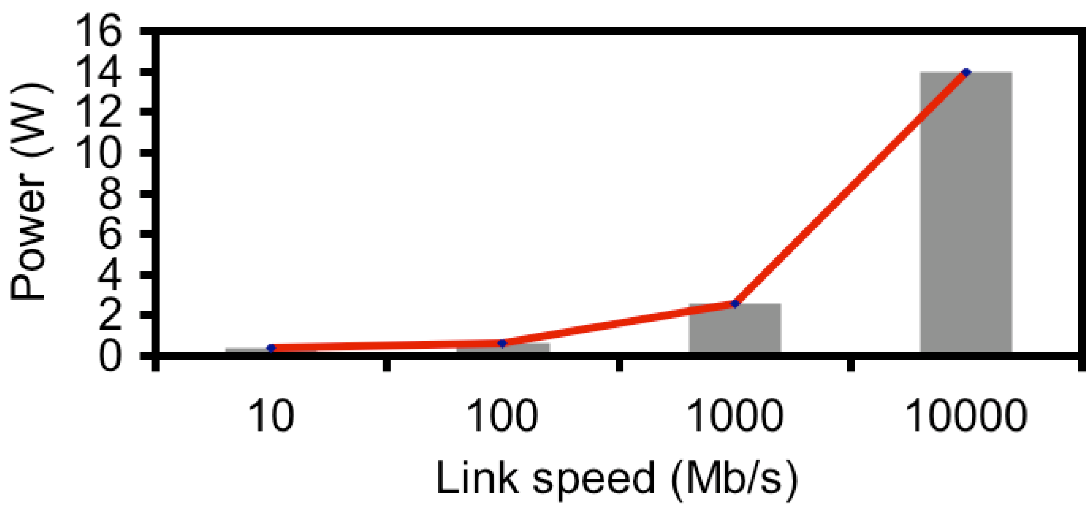
— all incremental AC power

— measuring 1st order

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



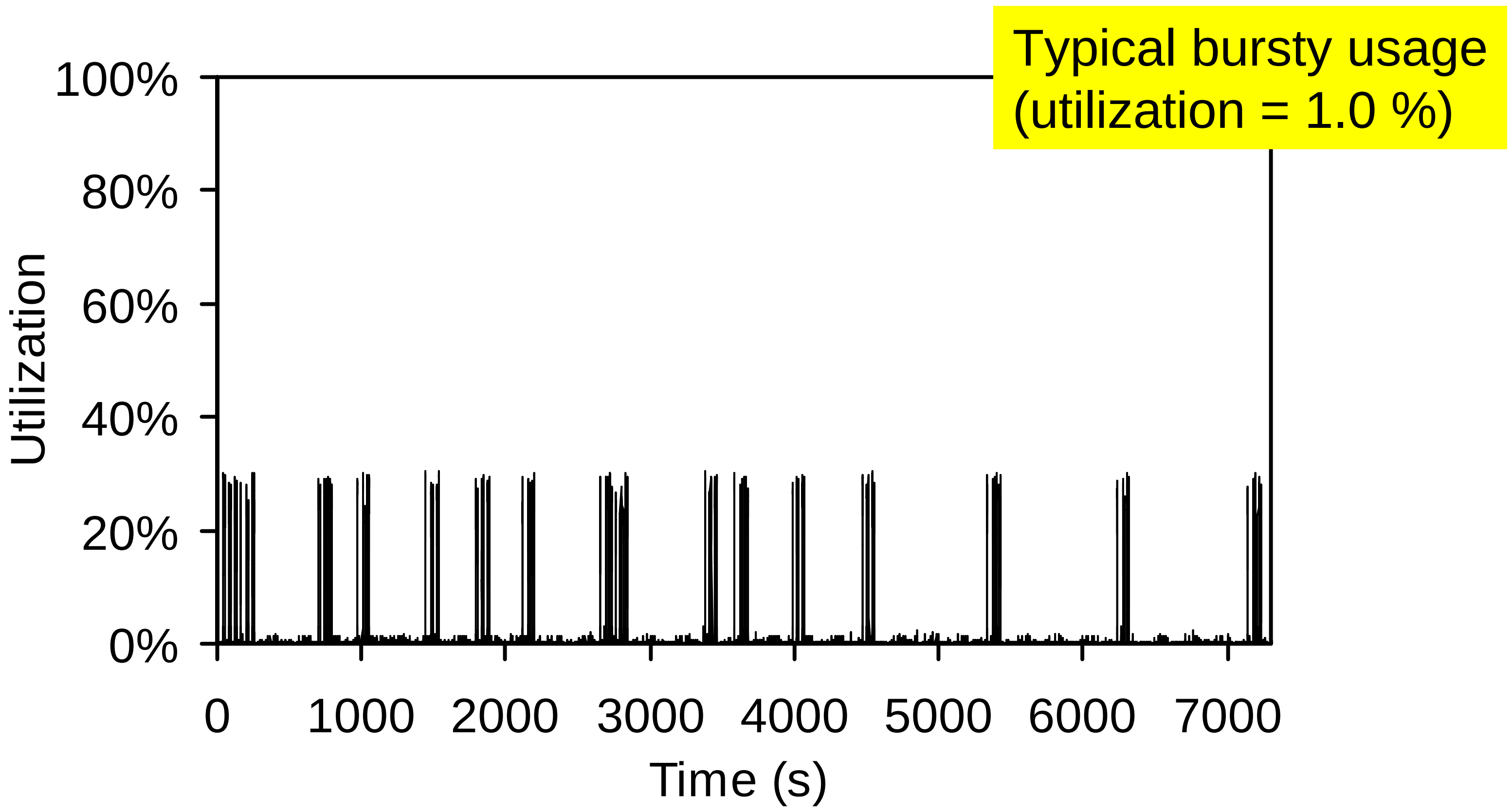
- **Various computer NICs averaged**



Based on initial numbers 10GBASE-T expected to be in the order of 25W AC

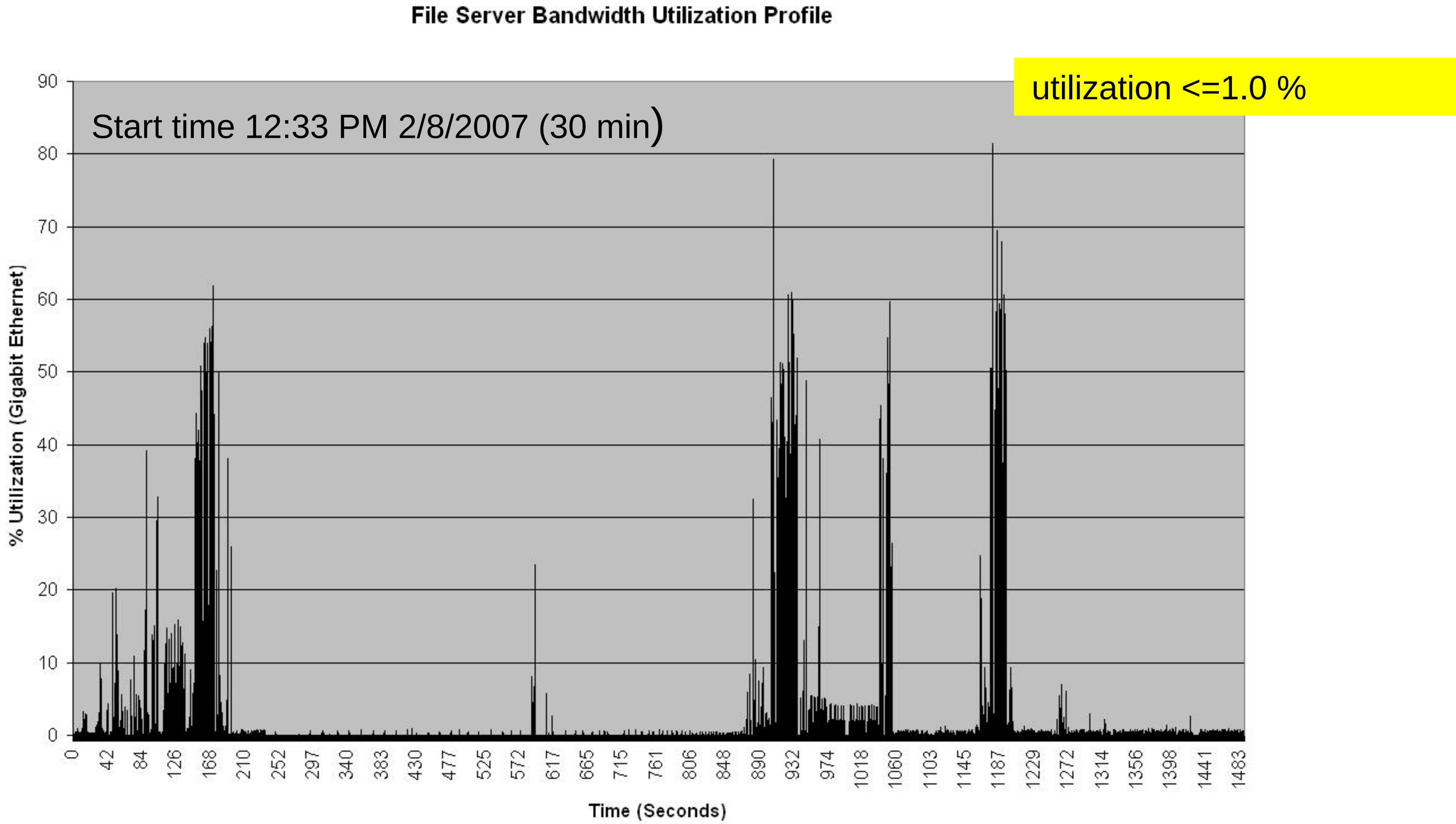
Desktop links have low utilization

- **Snapshot of a typical 100 Mb Ethernet link**
  - **Shows time versus utilization (trace from Portland State Univ.)**



Some Server links have low utilization

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



Potential Savings from EEE

Assume 100% adoption (U.S. Only), 90% operation at lower speed

- **Residential**
  - **PCs, network equipment, other**
  - **1.73 to 2.60 TWh/year**
  - **$139 to $208 million/year**
- **Commercial (Office)**
  - **PCs, switches, printers, etc.**
  - **1.47 to 2.21 TWh/year**
  - **$118 to $177 million/year**
- **Data Centers**
  - **Servers, storage, switches, routers, etc.**
  - **0.53 to 1.05 TWh/year**
  - **$42 to $84 million/year**

These figures do **not** include savings from cooling/power infrastructure

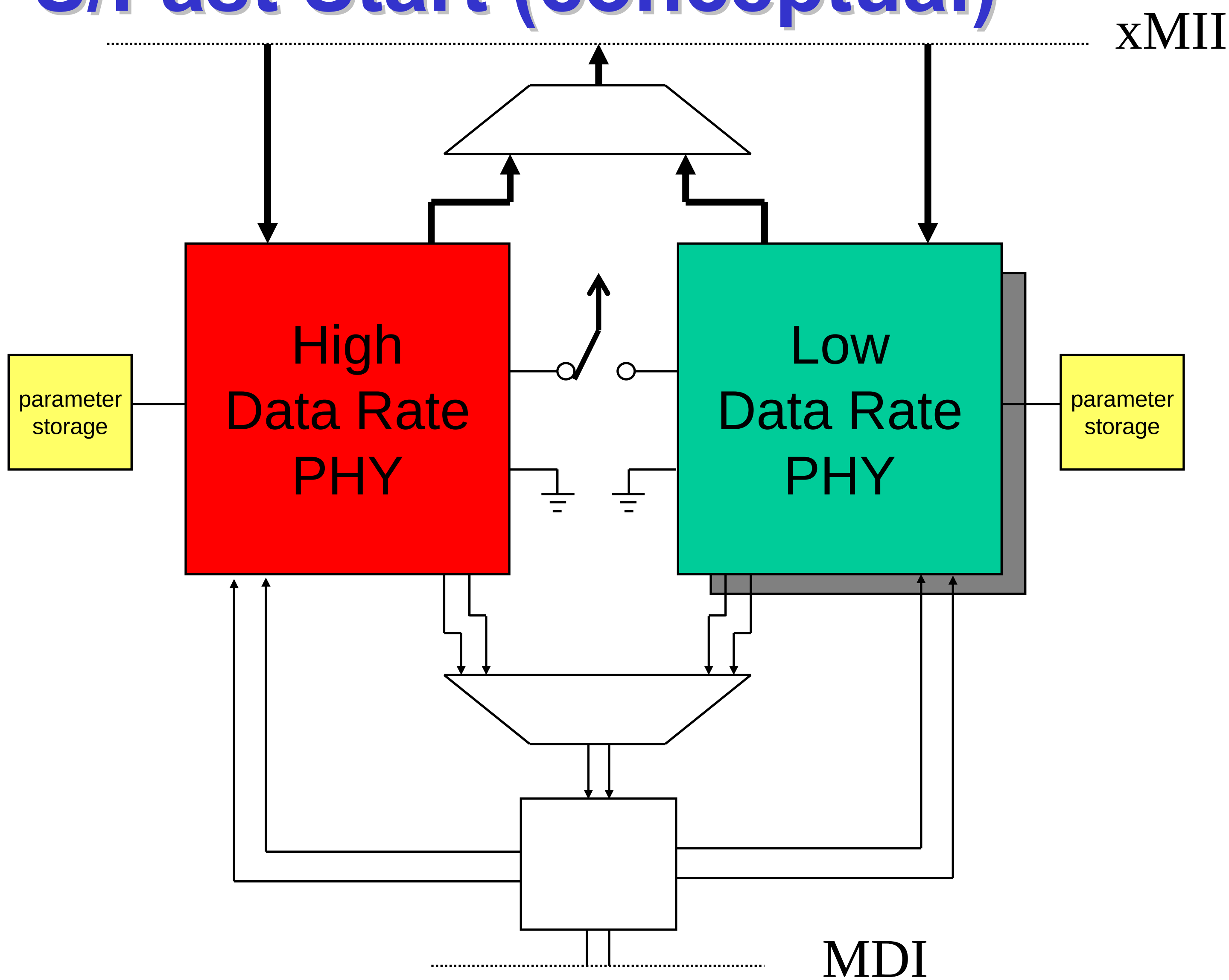
Total: $298 to $469 million/year

IEEE 802.3az Update

Study Group Overview

- **Formed in November, 2006**
- **6 meetings**
  - **39 presentations supporting Project Authorization Request (PAR), 5 criteria, and objectives**
    - **11 presentations on Rapid PHY Selection (RPS)**
    - **4 presentations on Subset PHY**
    - **3 presentations on modification of 10BASE-T**
    - **Remaining presentations focused on link utilization, power consumption, impact of transition time on application performance**
  - **Study Group voted to submit PAR for consideration at July 2007 meeting**
    - **PAR was approved by 802.3 in July, NesCom/SASB in Sept. 07**
- **The group focused mostly focus on RPS**

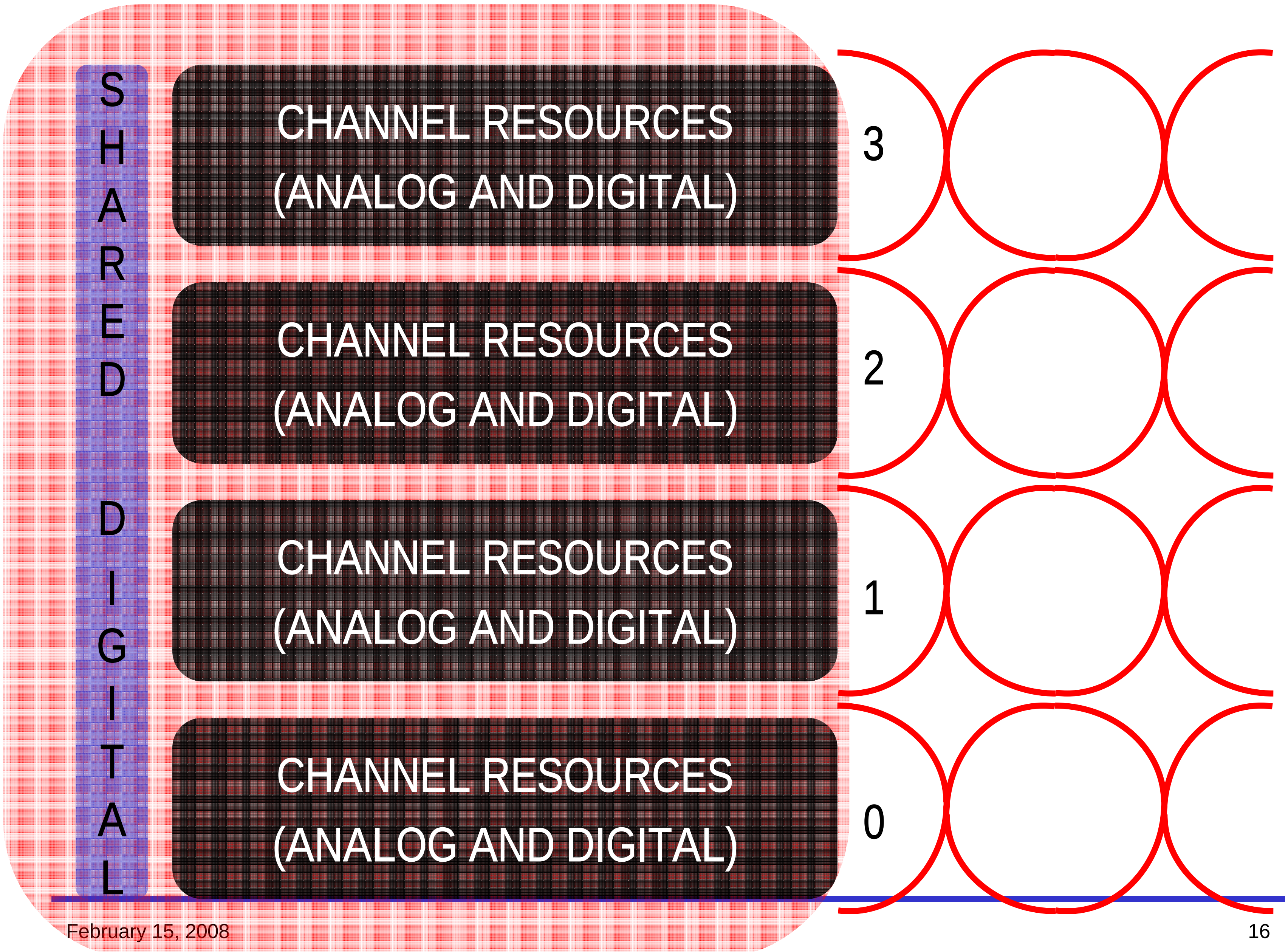
RPS/Fast Start (conceptual)



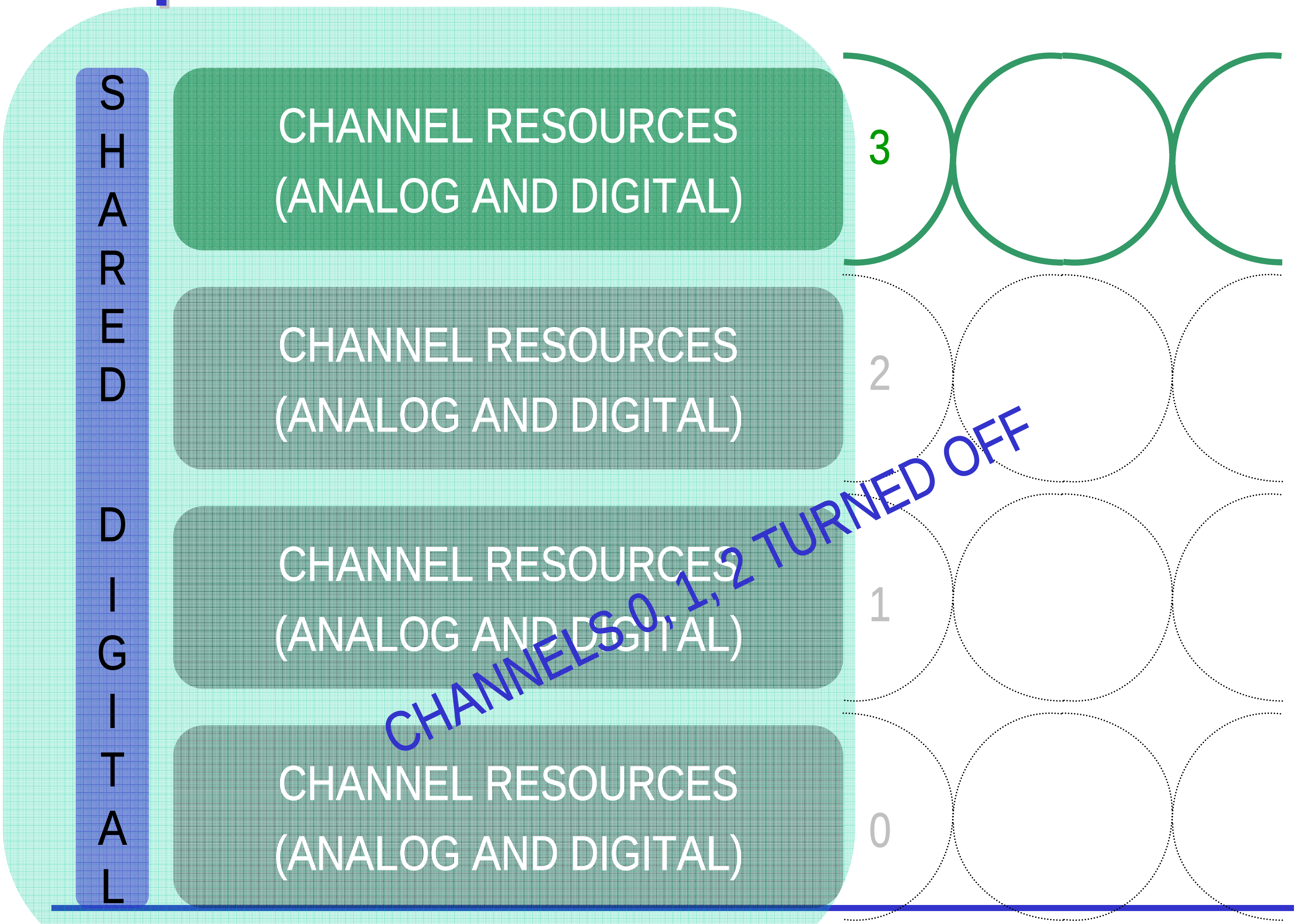
Transition time

- Several people concerned about the impact of transition time on applications
- An initial study on feasibility of 1 ms transition from lower speed to 10GBASE-T suggested 20 ms was feasible, 1 ms was not
- More concerns raised regarding impact on real-time applications such as Audio Video Bridging (AVB)
  - Transition time needs to be at most 1 ms
  - The problem is PHY change testing suggested 20 ms
  - What to do?

10GBASE-T PHY

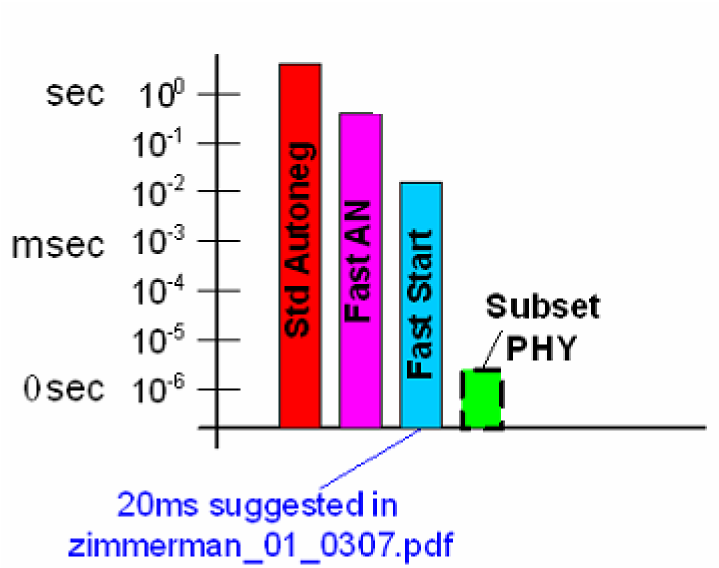


Simple 10GBASE-T Subset PHY



Transition time comparison

- **Assumptions**
  - **10GBASE-T is the highest negotiated speed**
  - **Power savings for various options is comparable**



- **Subset PHY offers potential to improve transition time by over 3 orders of magnitude**
  - **µS instead of mS**

Study group summary

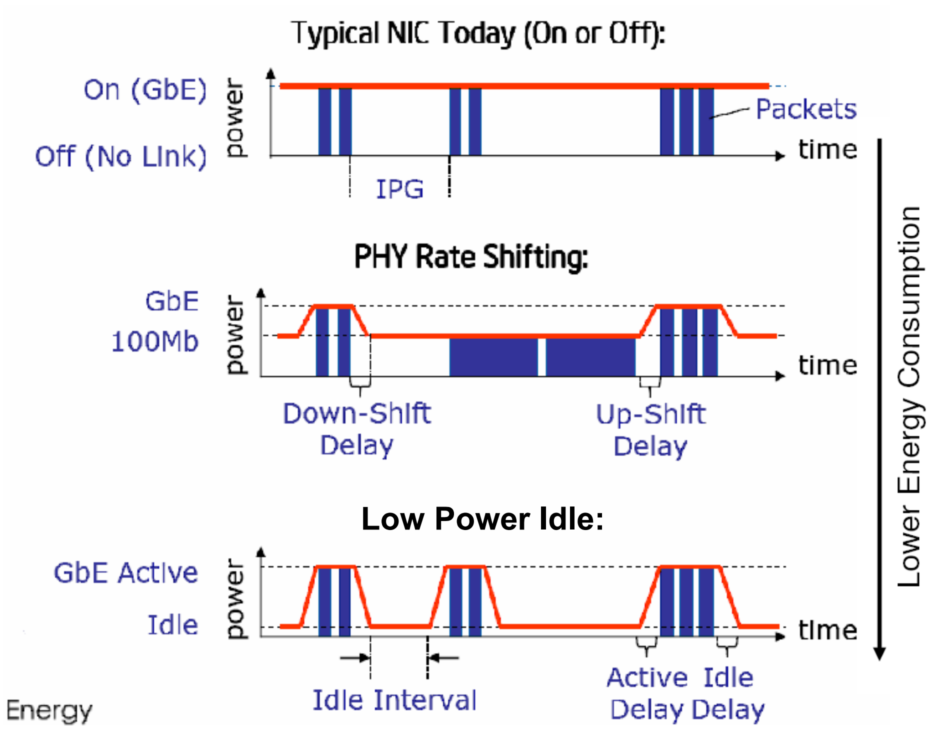
- During the Study group phase of project, we investigated:
  - Protocol to negotiate the speed change, stop transmission, change speeds and resume transmission at new speed
    - Impact of Frame-based protocol exchange on transition time
      - » At slower speed, waiting on "normal" frames to finish before speed change dominates transition time
      - » At higher speed, time to resume transmission dominates
  - Rapid PHY Selection / Fast Start (modified RPS)
    - Main difference between RPS and "Fast Start" is with the latter, state of channel characteristics is saved, entry points in state machines are optimized to minimize start-up time, thus minimizing total transition time
      - » Transition time in the order of 10's of milliseconds feasible
  - Subset PHY
    - Operate at lower speed by using a "subset" of a standard PHY
      - E.g. operate 1000BASE-T as a subset of a 10GBASE-T
      - Transition time in the order of 10's of microseconds feasible
  - Also working on Backplane Ethernet and 10BASE-T

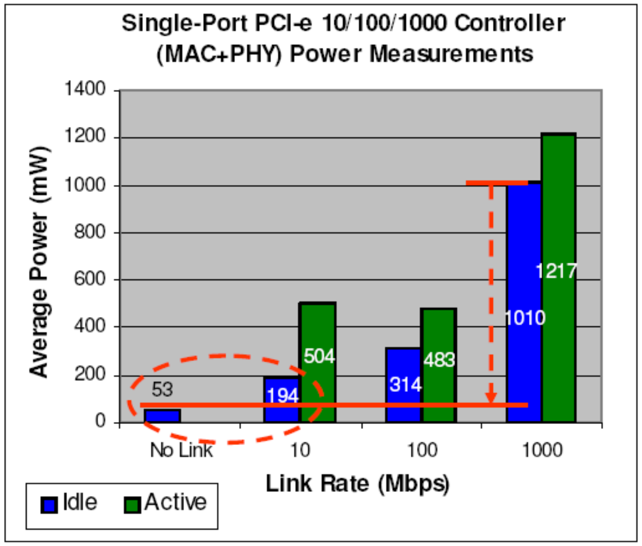
Task Force Overview

- **Formed in November, 2007**
- **2 meetings**
  - **24 presentations**
    - **Digging deeper into the technical details**
      - **More work done on Subset PHY approach**
    - **Working towards developing a baseline set of proposals**
  - **Introduction of a new concept**
    - **Low Power Idle (began as “Active Idle toggling”)**
      - **Simple concept: transmit when there is data to send, reduced power when there is not**
        - » Add a counting state machine for idle modes to wake up periodically
        - » Turn off receivers, transmitters for N frames
        - » Turn on receiver (or transmitter) on schedule for 1 (or M) frames
        - » Check for “wake-up” codeword
        - » Continue activity transitioning back to active mode or go back to “counting sleep” depending on codeword received

Low Power Idle

- Energy use is lower than typical NIC and RPS (rate shifting)
  - Transition time in the order if microseconds feasible





Source: Intel labs. Intel® 82573L Gigabit Ethernet Controller, 0.13µm, "Idle" = no traffic, "Active" = line-rate, bi-directional

Task Force Summary

- We're making good progress
  - Lots of good ideas
- We have a number of open questions to answer and issues to deal with
  - Low Power Idle will be efficient in bursty traffic
    - What happens when the traffic is real time and / or streaming?
    - Might require switch vendors to add buffers
  - Subset PHY approach will need a means to keep channel characteristics relatively stable
    - Send "refresh" signals over unused pairs periodically
  - There needs to be a means for applications to communicate with the network interface

Estimated Timeline

- **PAR approved by 802.3/EC July 2007**
- **Project 802.3az approved**
- **1st Task Force Meeting: November 2007**
- **Last new proposal: March 2008**
- **1st Draft done: May 2008**
- **2nd Draft done/Task Force Review: November 2008**
- **3rd Draft done/Working Group Ballot: March 2009**
- **4th Draft done/LMSC Ballot: July 2009**
- **5th Draft done: November 2009**
- **Standard done: March 2010**
- ***Note: timeline not adopted by the task force***

Acknowledgements

- ***Bruce Nordman and Ken Christensen***
- ***Howard Frazier, Wael William Diab, David Law, Bill Woodruff, George Zimmerman, Rob Hays, Mandeep Chadha***
- ***Energy Efficient Ethernet Study Group and 802.3az Task Force members, for their hard work and dedication to this project***

Thank You!

Extras

Objectives – what we've agreed to do

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

• 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

Objectives – what we've agreed to do

- **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**