Energy Efficient Ethernet and 802.1

Mike Bennett
(unconfirmed) Chair 802.3az

Wael William Diab
Broadcom

IEEE 802 Plenary
Atlanta, GA
November 16, 2007

Discussion

- Brief review
    - What is EEE?
- Transition time and examples of latency-sensitive applications
- EEE Techniques Detail
    - Fast Start
    - Subset PHY
- Example of a speed shift
- Goals & Request for help
    - Straw Polls
- Additional Information

Contributors

- David Law, 3Com
- Bill Woodruff, Aquantia
- Howard Frazier, Broadcom
- Scott Powell, Broadcom
- Shalini Rajan, Solarflare
- George Zimmerman, Solarflare
- Mandeep Chadha, Vitesse

What is Energy Efficient Ethernet?

- A method to reduce energy use by an Ethernet interface by rapidly changing to a lower link speed during periods of low link utilization
- 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
- ALR = Rapid PHY Selection (RPS) + Control Policy
  - Generic RPS covers all of the techniques described later
  - Note: Control policy is the outside scope of 802.3

PHY Transition 101

- Transition time is time to switch between link rates
- A PHY has to keep "state information" about the link current for the link to be operational
- Initially acquired when PHY *trains* as part of link up
  - Information is dependent on the environment
  - When the link is up, the information is constantly updated
- When a different signaling scheme is used (e.g. a different standard PHY), the state will drift with time
  - Last stored state will become *stale* with time
  - This makes the link inoperable and requires full restart
  - May limit time different signaling scheme can be used
- 802.3az is exploring two possible solutions
  - FastStart (msecs transition) / Subset PHY (usecs transition)

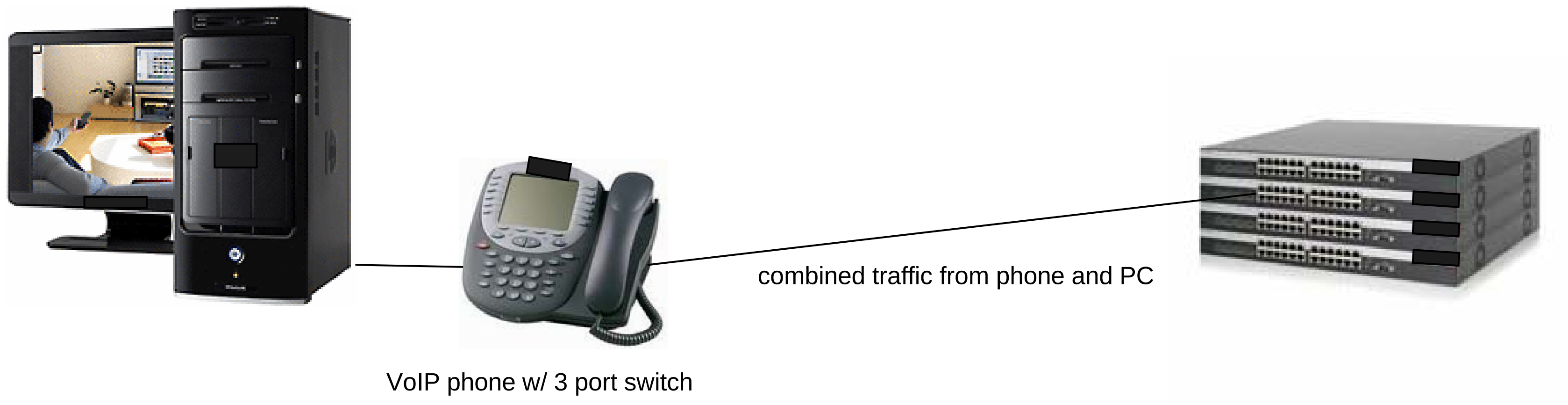
Transition Process and Upper Layers

- Control Policy
  - Decision when/how much (e.g. 10G to 1G vs. 10G to 100M) outside scope of 802.3az.
  - Once initiating side decides to switch over, 802.3az communication and transition mechanisms kick in
  - Would like to get feedback if .1 wants to take on policy
- Capability
  - A PHY with EEE capability will presumably advertise this to both the management layers and far end (perhaps at linkup)
- Actual transitions transparent to upper layers
  - Beyond the control entity, upper layers will be unaware of when the actual transition occurs
  - Upper layers continue to see higher capacity link
  - Link will not be dropped
  - Packets will not be dropped or corrupted

How Important is Transition Time?

- EEE looked at a number of issues. Issues fall into two categories: App jitter and buffering
- Latency sensitive applications
  - Rate change may introduce unwanted application jitter. For example in VoIP call or AVB network
- Switch buffering
  - All frames have to be preserved during transition
  - During transition node has to hold line-rate *@higher rate*
  - Transition time may have an impact on buffer size
- Would like to get feedback from 802.1 on transition time

Application – IP phone

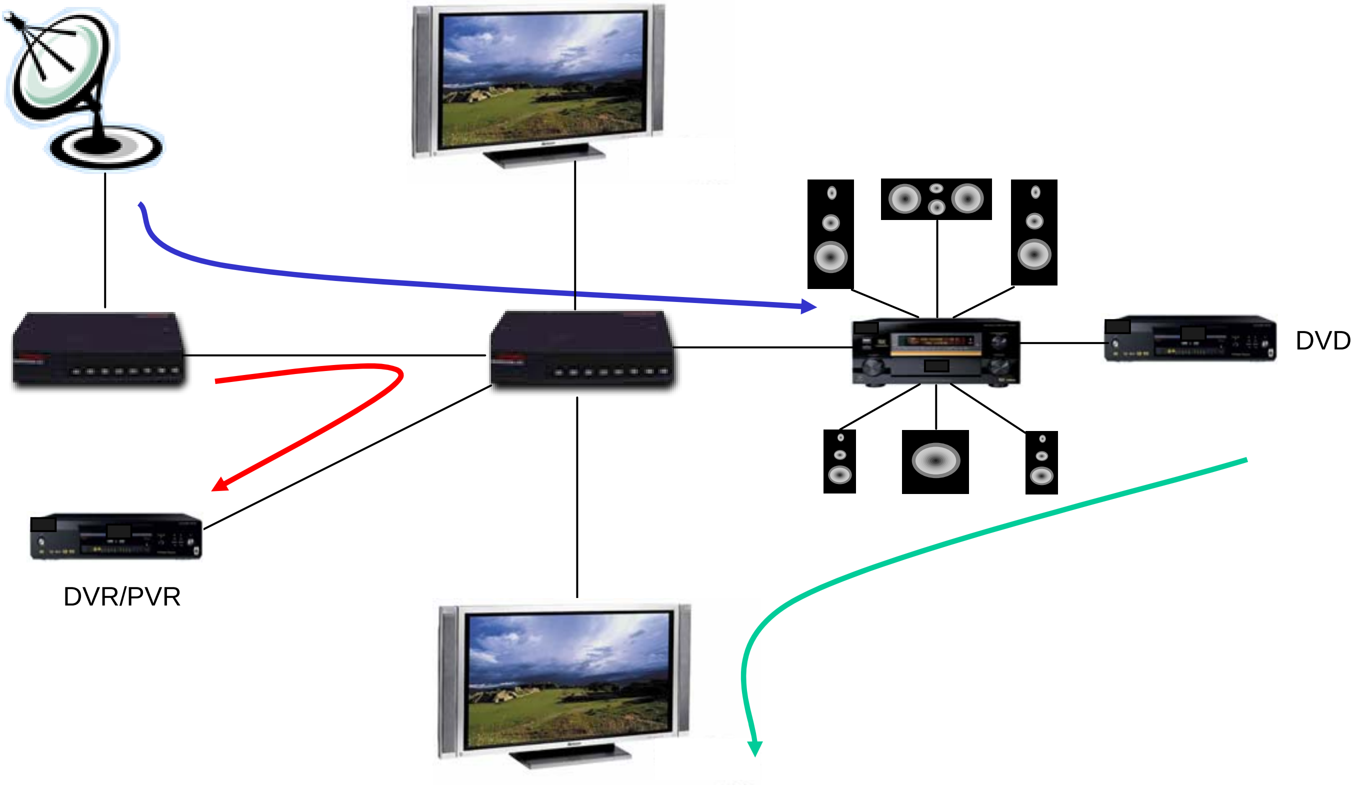


combined traffic from phone and PC

VoIP phone w/ 3 port switch

1. VoIP traffic low, link between phone and switch operates at 100 Mbps
2. Application on PC initiates data transfer
3. Link between phone and wiring closet switch transitions from 100 Mbps to 1000 Mbps
4. Transition time must be less than 10 ms to avoid audible disruption of phone call
5. Application on PC finishes data transfer
6. Link between phone and wiring closet switch transitions back to 100 Mbps
7. Transition time must be less than 10 ms to avoid audible disruption of phone call

Application - AVB home network



1. Listening to satellite radio on AVB receiver, link between receiver and switch operating at 10 Mbps
2. Start playing DVD on a screen in another room
3. Link between receiver and switch must transition from 10 Mbps to 100 or 1000 Mbps
4. Transition time must be less than 10 ms to avoid audible disruption
5. DVR/PVR set to record “Survivor” from satellite receiver at 8:00 pm on Thursday
6. Link between satellite receiver and AVB switch must transition from 10 Mbps to 100 or 1000 Mbps
7. Transition time must be less than 10 ms to avoid audible disruption

Categories of Solutions

1. Standard Autoneg + startup ("Std Autoneg")
   - aka, reset and re-establish at the new speed

2. Skip unnecessary autoneg steps ("Fast AN")
   - Speed, duplex, M/S resolution, etc are all established on first link up
   - No need to re-negotiate after an EEE speed change

3. Skip unnecessary start-up steps ("Fast Start")
   - Power backoff, precoder coefficient exchange, etc (10G)
   - Initialize filters, cancellers, control loops from last known state

4. Switch between 802.3 PHY and subset PHY ("Subset PHY")
   - Define lower power PHY as a subset of the higher speed standard PHY

Fast Start: Overview

- Technique switches between different standard PHY signaling
  - E.g. 10GBASE-T / 1000BASE-T. 10GBASE-T / 100BASE-TX
  - E.g. 1000BASE-T / 100BASE-TX. 1000BASE-T / 10BASE-T
- Amount of time PHY can spend in the different signaling states is bound by the *aging* of the state information.
- Transition can take advantage of the fact the PHY had been operating on the link already allowing *Fast Start / Training*
  - This allows optimization of training / start sequences

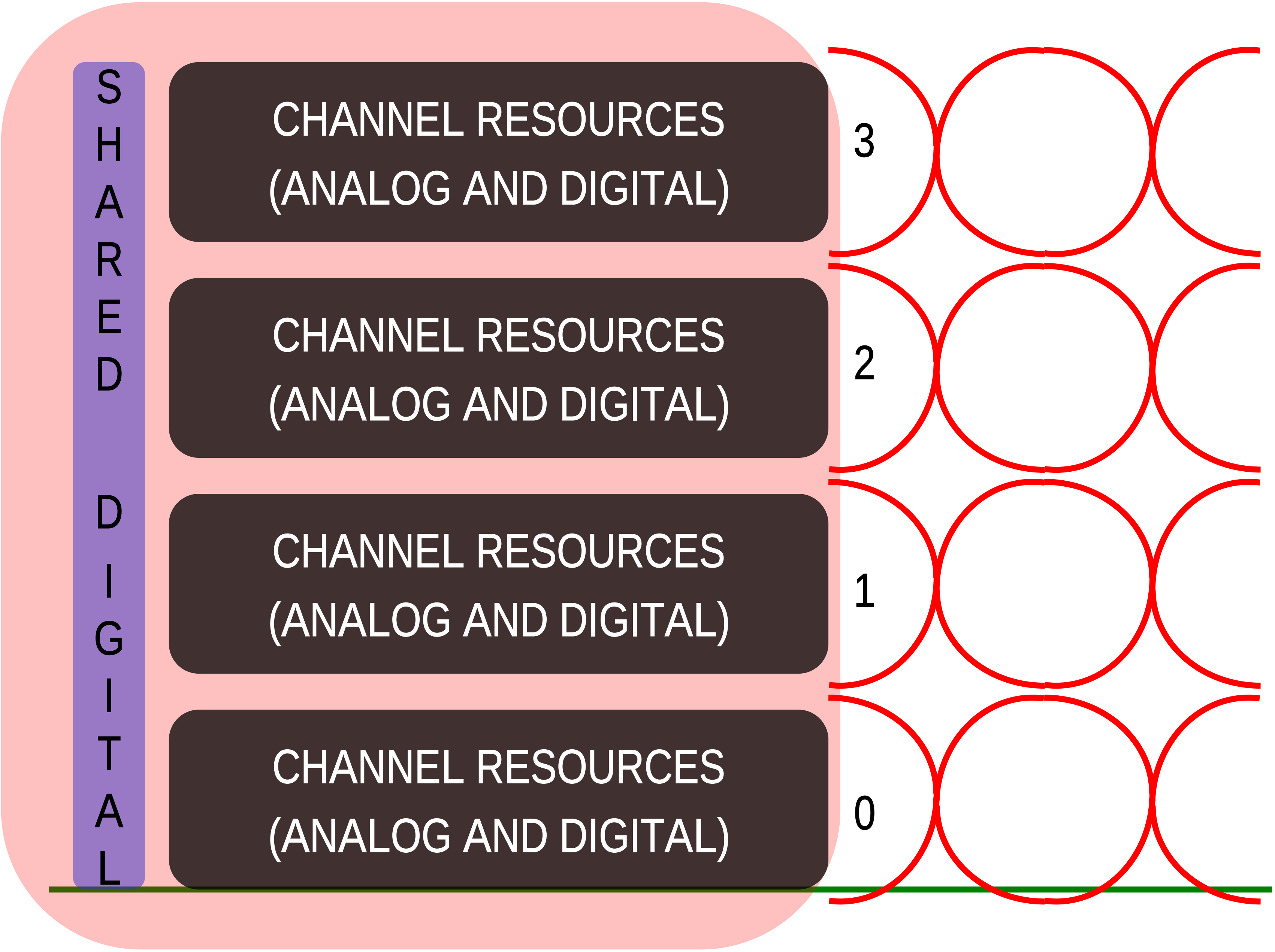
Fast Start: Conclusions

- **Fast restart of 10GBASE-T and 1000BASE-T from stored state appear feasible**
  - **Preliminary experiments suggest state should be current within 3 minutes for entry *without* retraining, >5 minutes requires a fast restart/retrain**
- **Fast restart can reuse existing PHY standards**
  - **Simple changes to the transition counter in the standard allow development of transitions within ~1-2msec**
- **Laboratory results demonstrate the feasibility of 3-4msec retrains today, even on 100m 10GBASE-T links**
  - **Similar results for 1000BASE-T**
- **These are *without* the PHY implementations being modified to improve fast acquisition & training**

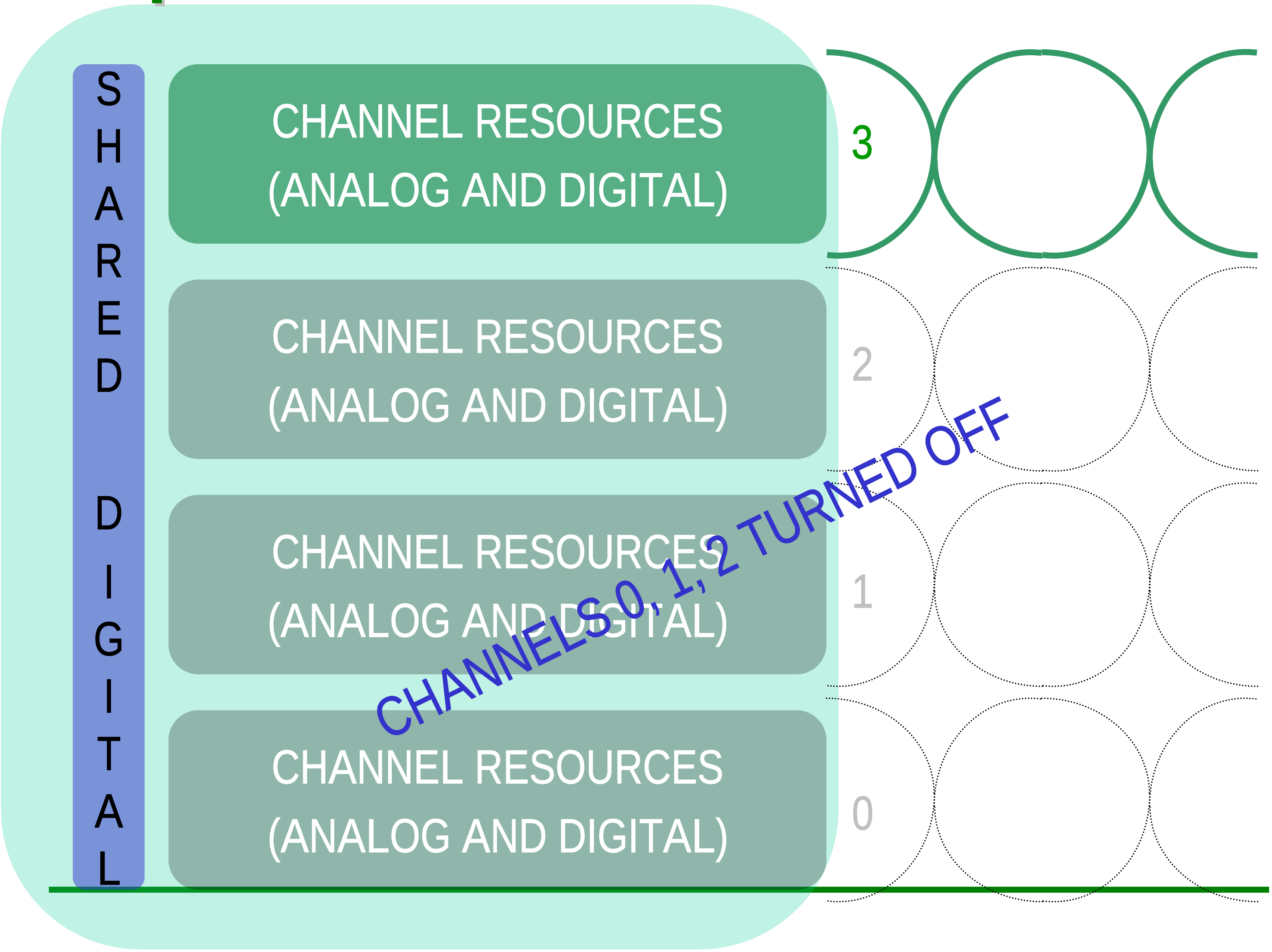
Subset PHY: Overview

- PHY resources are channel dependent and shared
  - Bulk of resources such as DSP and Analog are per channel
- Minimizes power while maintaining the current state
  - PHY stays within current signaling technique using a *subset*
- PHY turns off channels (and their associated dependent resources) to achieve lower rate
- Shared & ON resource power-optimized for lower rate
  - E.g. clocking of shared resource slowed, cancellers OFF
- 10GBASE-T to 1000BASE-T Example
  - Line code: PAM-16 -> PAM-4; Reduced number of channels (each direction): 4 -> 1. Produces simplex operation @800 MSymbols/S
  - Also 10GBASE-T->10/100BASE-TX. 1000BASE-T->10/100BASE-T
- PHY not bound in time in lower state due to aging
- Transition very quick as *no re-training* is necessary

10GBASE-T PHY


SHARED DIGITAL
CHANNEL RESOURCES (ANALOG AND DIGITAL)
3
CHANNEL RESOURCES (ANALOG AND DIGITAL)
2
CHANNEL RESOURCES (ANALOG AND DIGITAL)
1
CHANNEL RESOURCES (ANALOG AND DIGITAL)
0

Simple 10GBASE-T Subset PHY



SHARED DIGITAL

CHANNEL RESOURCES (ANALOG AND DIGITAL)

3

CHANNEL RESOURCES (ANALOG AND DIGITAL)

2

CHANNEL RESOURCES (ANALOG AND DIGITAL)

1

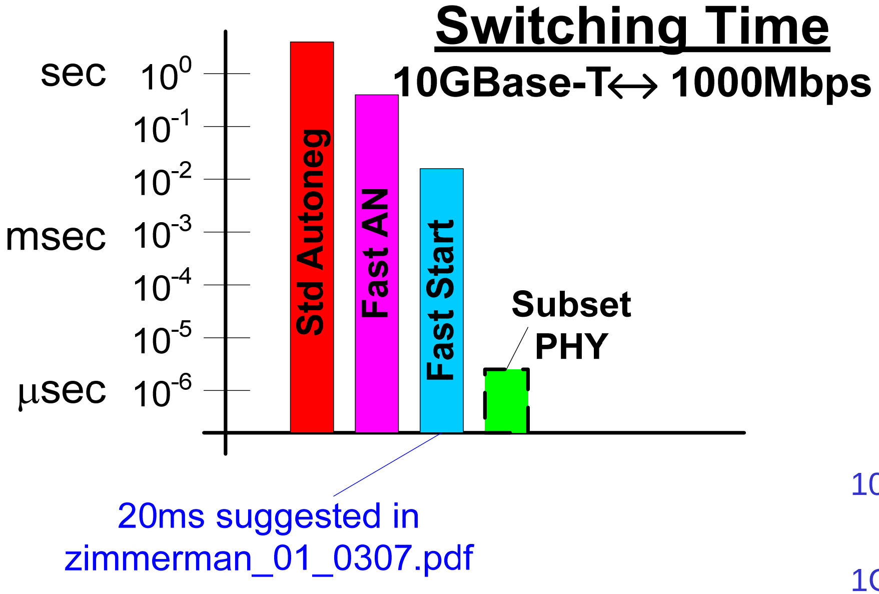
CHANNEL RESOURCES (ANALOG AND DIGITAL)

0

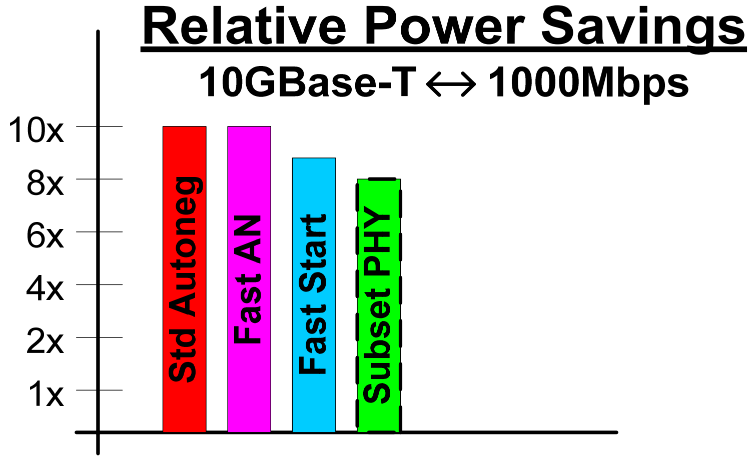
CHANNELS 0, 1, 2 TURNED OFF

Subset PHY: An Example of the Tradeoff

- **Assume 10GBASE-T is the highest negotiated speed**
- **Speed and power of subset PHY are an early estimate of what's possible**



20ms suggested in zimmerman_01_0307.pdf



10GBase-T ~ 10W
(www.linleygroup.com/npu/Newsletter/wire070517.html)

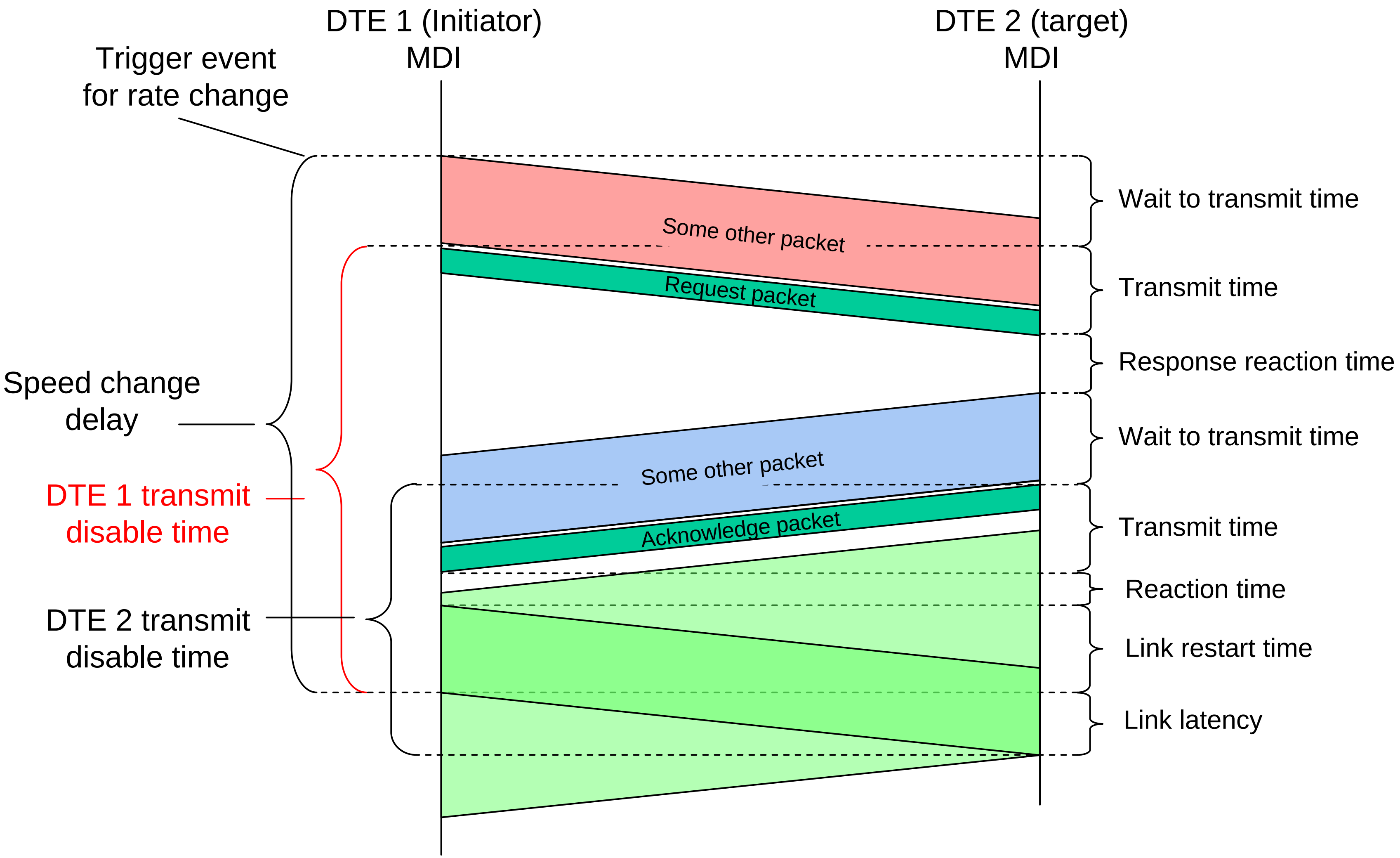
1GBase-T ~ 500mW
(www.broadcom.com/collateral/pb/54980-PB201-R.pdf)

- **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**

Subset PHY: Summary & Conclusions

- It is possible to achieve very fast transitions with a Subset PHY
  - usec range feasible
- Similar techniques can be applied to 10GBASE-T and to 1000BASE-T
- Significant PHY power savings can be achieved
- Implementation complexity potentially less than shifting between standard PHYs

Transmit disable process*



*Example assumes packet based control protocol. TF also considering physical layer control

Goals

- We want to make sure that the decisions we make in 802.3 have minimal impact on work in 802.1

- We would like to be aware of decisions in 802.1 that may impact us

- Open communication with 802.1
  - E.g. We have not yet decided whether or not to use a framed-based protocol to communicate speed changes
    - If we choose a frame-based protocol, we'll inform you

- EEE and 802.1 will have to work together to make sure we minimize impact on each other's work

Things we could use help with

This week: We would like to get your feedback

- Control policy
  - Should 802.1 work on any or all of a policy?
- Transition time
  - What transition time is deemed transparent to 802.1?
  - uSecs, mSecs or Secs

Long term: Ongoing dialogue

- Coordination with time-sensitive protocols, e.g. for AVB
- How do we communicate with the upper layers?
  - Default: MIBs accessible via Layer Management Entity.
  - MAC Service interface?
  - SNMP?

Straw Poll: Sense of 802.1 on Transition Time

- Preferred transition time of

	One Vote
• usecs	__39__
• msecs	__3__
• secs	__0__
• Don’t care	__1__

Straw Poll:
Sense of 802.1 on Control Policy

- Would 802.1 like to take on the EEE Control Policy?

	One Vote
Y	__21__
N	__6__
Don’t care	__15__

Additional Information

- For a more detailed presentation on EEE please refer to the 802 tutorial http://www.ieee802.org/802_tutorials/july07/IEEE-tutorial-energy-efficient-ethernet.pdf

- The EEE website can be found at
  - Study Group: http://www.ieee802.org/3/eee_study/index.html
  - Transitioning to Task Force

    http://www.ieee802.org/3/az/index.html

Thank you!