IEEE 802 Tutorial

Energy Efficient Ethernet

Hugh Barrass, Cisco
Mike Bennett, Lawrence Berkeley Lab
Wael William Diab, Broadcom
David Law, 3Com
Bruce Nordman, Lawrence Berkeley Lab
George Zimmerman, Solarflare

IEEE 802 Plenary
San Francisco, CA
July 16, 2007

Contributors

- David Law, 3Com
- Bill Woodruff, Aquantia
- Wael Diab, Broadcom
- Howard Frazier, Broadcom
- Scott Powell, Broadcom
- Pat Thaler, Broadcom
- Hugh Barrass, Cisco
- Rudy Klecka, Cisco
- Mike Bennett, LBNL
- Bruce Nordman, LBNL
- Shalini Rajan, Solarflare
- George Zimmerman, Solarflare
- Ken Christensen, USF
- Mandeep Chadha, Vitesse

Agenda

- Introduction
  - Mike Bennett, EEESG Chair
- Energy Use and Savings
  - Network Energy Use
    - Bruce Nordman, LBNL
  - Energy Efficient Ethernet: Beyond the PHY
    - Hugh Barrass, Cisco
- Feasibility
  - Transition Time Considerations and Calculations
    - David Law, 3Com
  - Technical Feasibility
    - Wael William Diab, Broadcom
- Objectives
  - Wael William Diab, Broadcom
- PAR & 5 Criteria
  - Mike Bennett, EEESG Chair
- Wrap-up

Introduction to Energy Efficient Ethernet

Mike Bennett

Lawrence Berkeley National Laboratory

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

Why Energy Efficiency Now?

- **Network industry has an opportunity to catch up with the server industry in this area**
  - **April 19, 2006 "Green Grid" formed**
    - **"**A group of technology industry leaders form The Green Grid to help reduce growing power and cooling demands in enterprise datacenters."
  - **Energy industry incentives for efficient products**
- **Energy Star**
  - **Requirements for PCs coming in 2009**
    - **"All computers shall reduce their network link speeds during times of low data traffic levels in accordance with any industry standards that provide for quick transitions among link rates"**
- **Customers like saving energy because it reduces operating costs**

Reference: ENERGY STAR® Program Requirements for Computers (V4.0, tier 2 requirements) available at http://www.energystar.gov/ia/partners/prod_development/revisions/downloads/computer/ComputerSpec_Final_Draft.pdf

Why Energy Efficiency Now?

- **Energy Efficiency gets U.S. congressional recognition**
  - **December 20, 2006 House Resolution 5646 signed into law**
    - **"To study and promote the use of energy efficient computer servers in the United States"**
    - **EPA hired LBNL to write the report, which was submitted in June**

- **The market for energy efficient Ethernet**
  - **Driven by customer's desire to save energy costs**
  - **Ethernet is used in markets where saving energy is crucial**
  - **Accelerate deployment for new applications**
  - **Enables use of incentives by energy industry**
  - **Ultimately these translate to increased demand**

Link utilization

- **Desktop-to-switch links**
  - **Are mostly idle**
    - **Lots of very low bandwidth "chatter"**
  - **High bandwidth needed for bursts**
    - **Bursts are often seconds to hours apart**
- **Server links are also often not fully utilized**
  - **Higher speed links offer more opportunity to save energy**
  - **This is an area where more data is needed**
- **Evidence of low utilization (desktop users)**
  - **LAN link utilization is generally in range 1 to 5% [1, 2]**
  - **Utilization for "busiest" user in USF was 4% of 100 Mb/s**

[1] A. Odlyzko, "Data Networks are Lightly Utilized, and Will Stay That Way", *Review of Network Economics*, Vol. 2, No. 3, pp. 210-237, September 2003.

[2] R. Pang, M. Allman, M. Bennett, J. Lee, V. Paxson, and B. Tierney, "A First Look at Modern Enterprise Traffic," *Proceedings of IMC 2005*, October 2005

Desktop links have low utilization

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

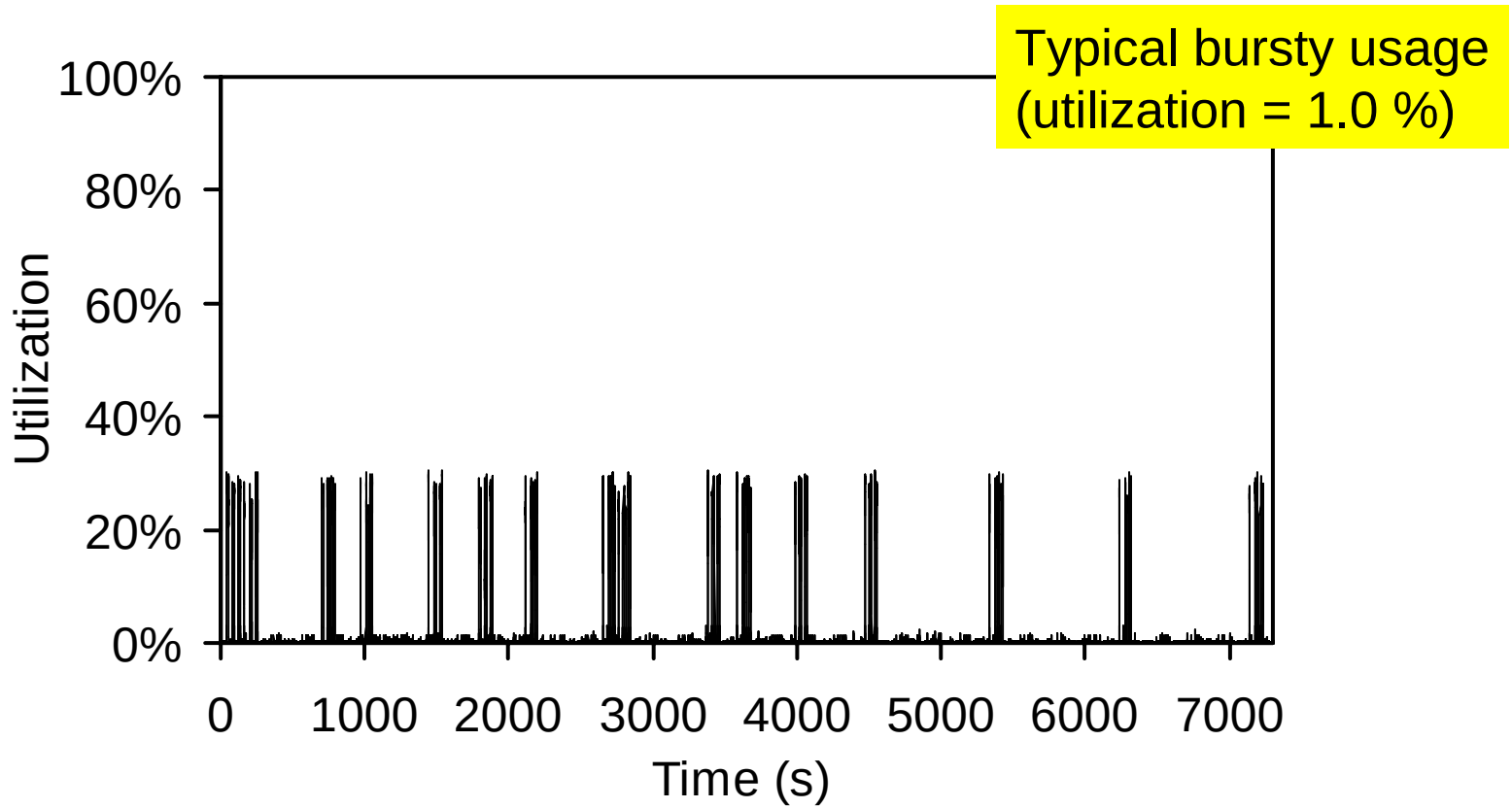
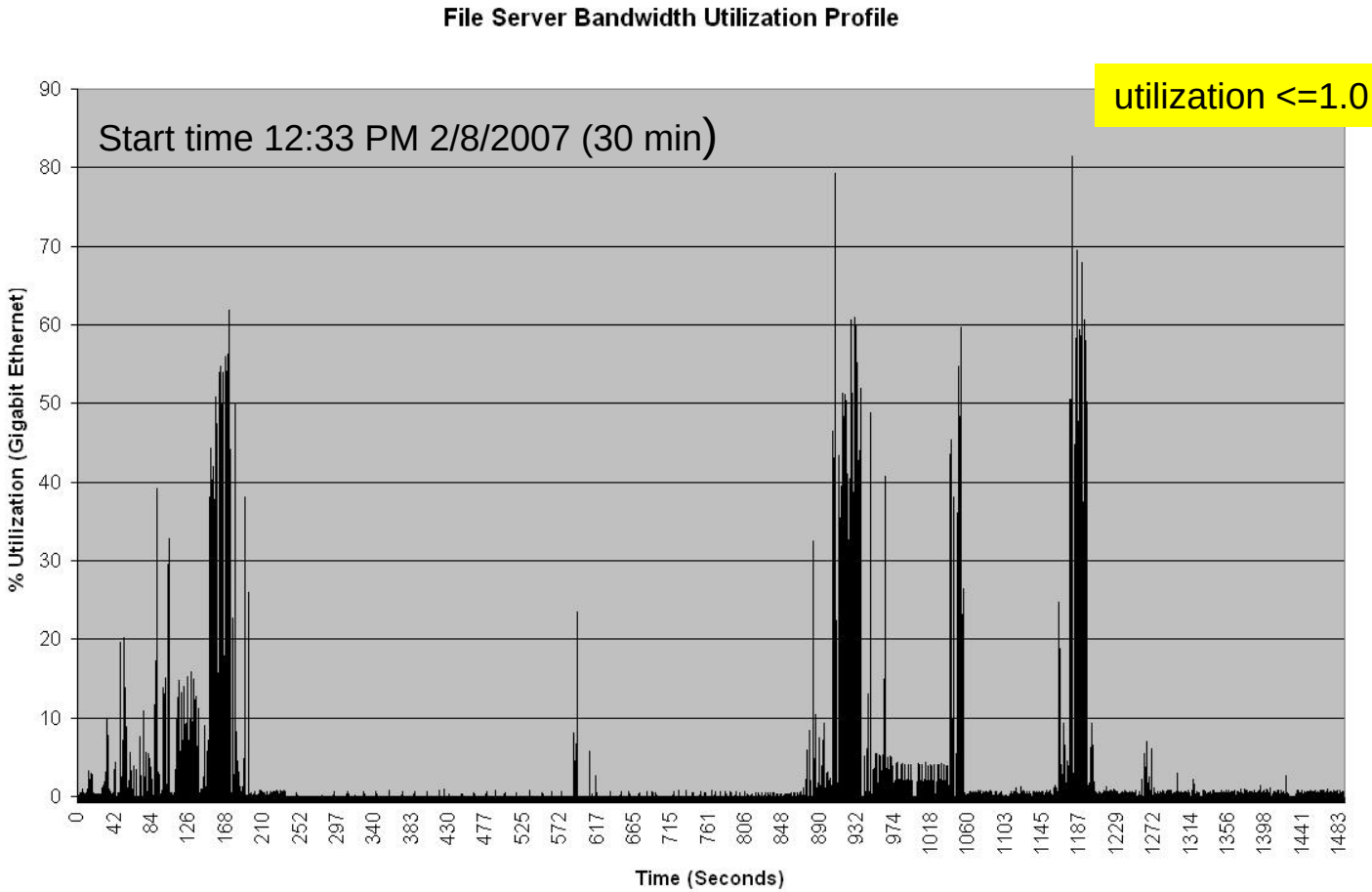


Fig1.xls

Some Server links have low utilization

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



Reducing the link rate

- Can (and does) save energy

- Some NICs drop link rate when a laptop is battery-powered
  - Or, when a PC goes into sleep state
  - Turns-off PHY if no signal on link

- Match the link rate to utilization
  - High utilization = high link rate
  - Low utilization = low link rate
  - Do this within the capabilities established by Auto-negotiation

- We need fast transitions
  - Auto-negotiation won't accomplish this as it is not transparent
    - Can't change speeds without dropping the link and it takes too long
    - Can't advertise the desire to change to a higher speed

Study Group Progress

- **Study Group formed in November, 2006**
  - "Move that the IEEE 802.3 working group request formation of an *Energy Efficient Ethernet* IEEE 802.3 study group to evaluate methods to reduce energy use by reduction of link speed during periods of low link utilization"
- **4 meetings to date**
  - **28 presentations supporting Project Authorization Request (PAR), 5 criteria, and objectives**
  - **Study Group voted to submit PAR for consideration at July meeting**
- **The group has been focused mostly on**
  - **RPS (Copper PHYs)**
  - **Transition time**
- **Before getting into the technical details, let's have a look at energy use**

Network Energy Use

Bruce Nordman
Lawrence Berkeley National Laboratory

The problem

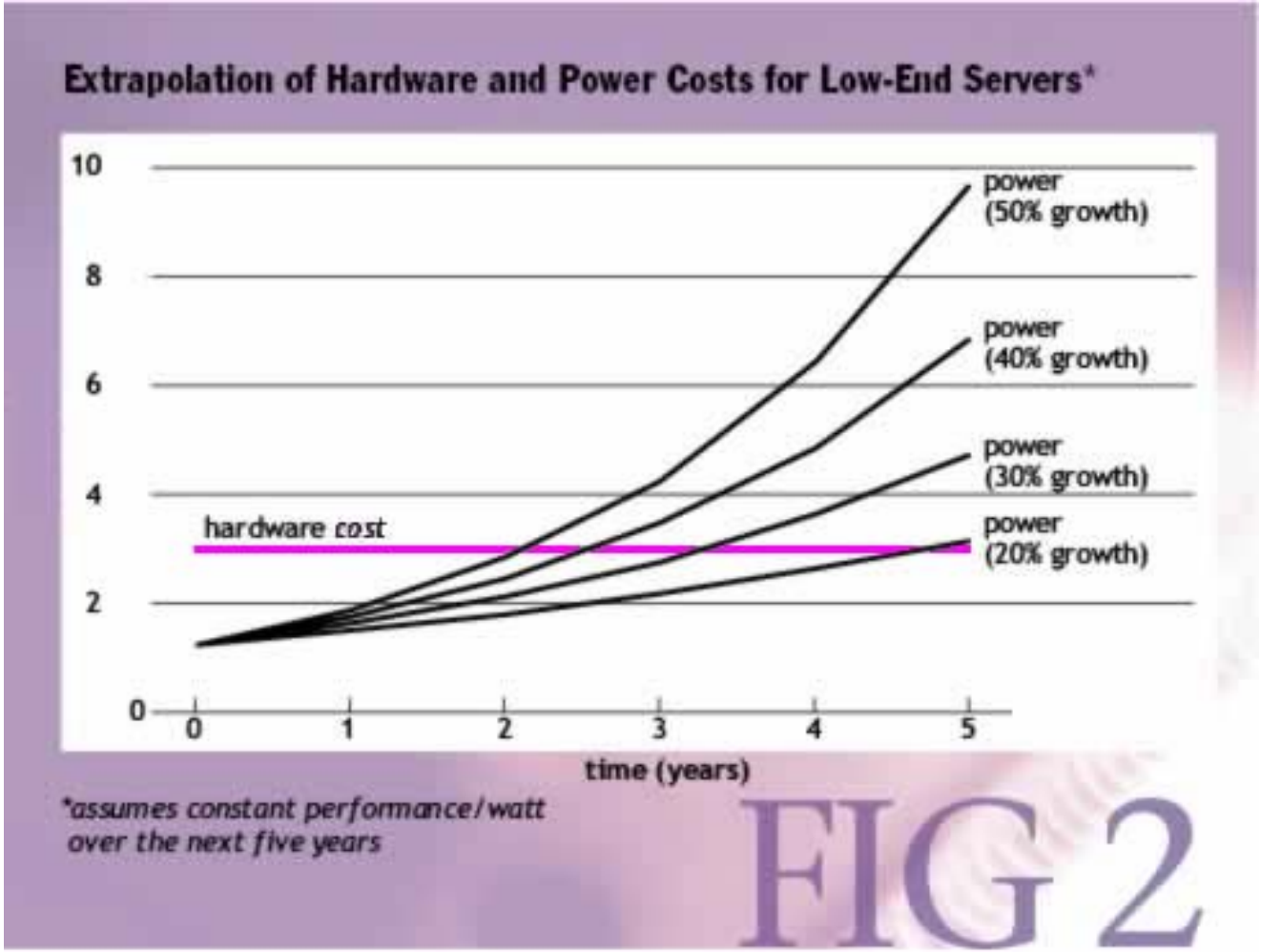
Numbers represent U.S. only

- **All electronics**
  - **IT equipment, consumer electronics, telephony**
    - **Residential, commercial, industrial**
  - **At least 250 TWh/year**
  - **$20 billion/year**
    - **Based on .08$/kWh; rates are rising**
  - **Over 180 million tons of CO_2 per year**
    - **Roughly equivalent to 35 million cars!**
- **IT equipment about half of this**
  - **PCs, displays, printers, servers, network equipment**

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

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

The problem

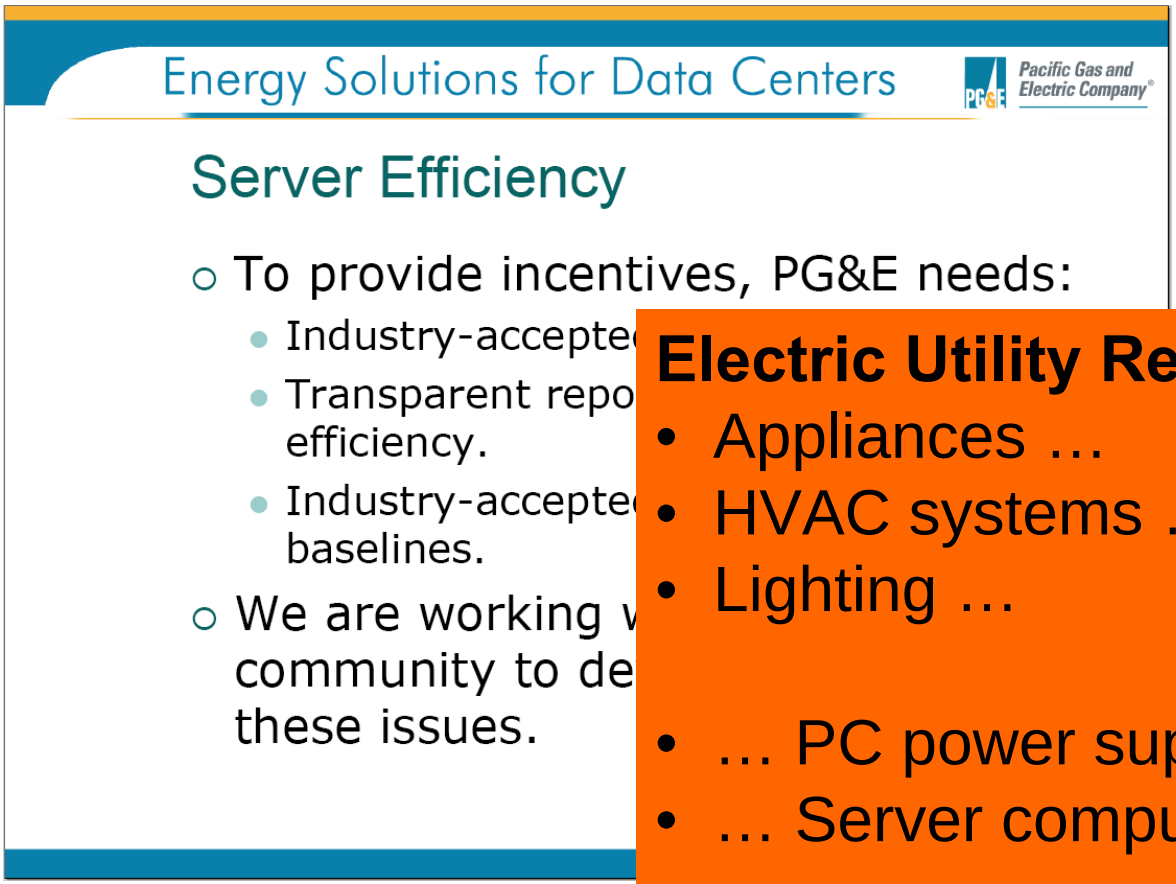


Unrestrained IT power consumption could eclipse hardware costs and put great pressure on affordability, data center infrastructure, and the environment.

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

(Modified with permission.)

Energy industry responds



Electric Utility Rebates

- Appliances …
- HVAC systems …
- Lighting …

- … PC power supplies (2005)
- … Server computers (2006)
- …

Reference: http://www.pge.com/docs/pdfs/biz/rebates/hightech/DataCenters_slides.pdf

PG&E (California) provides rebates for more energy-efficient servers

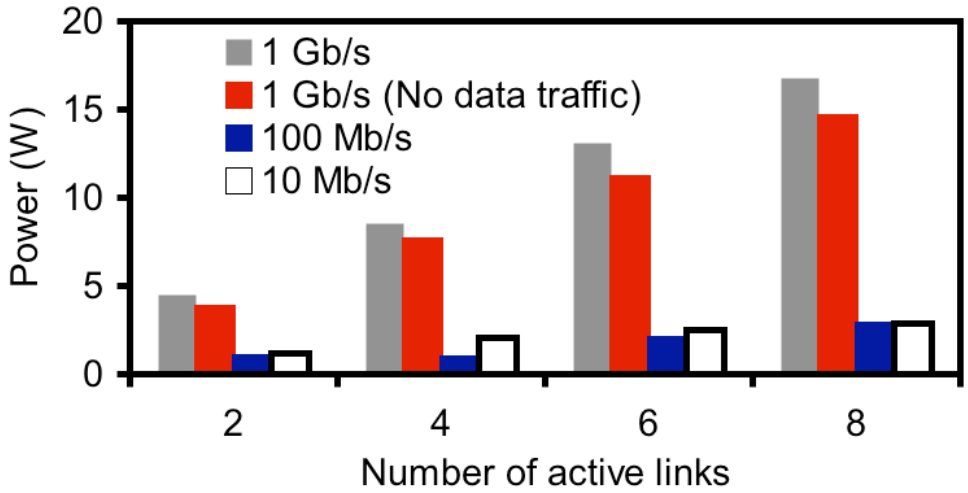
Link power

Results from (rough) measurements

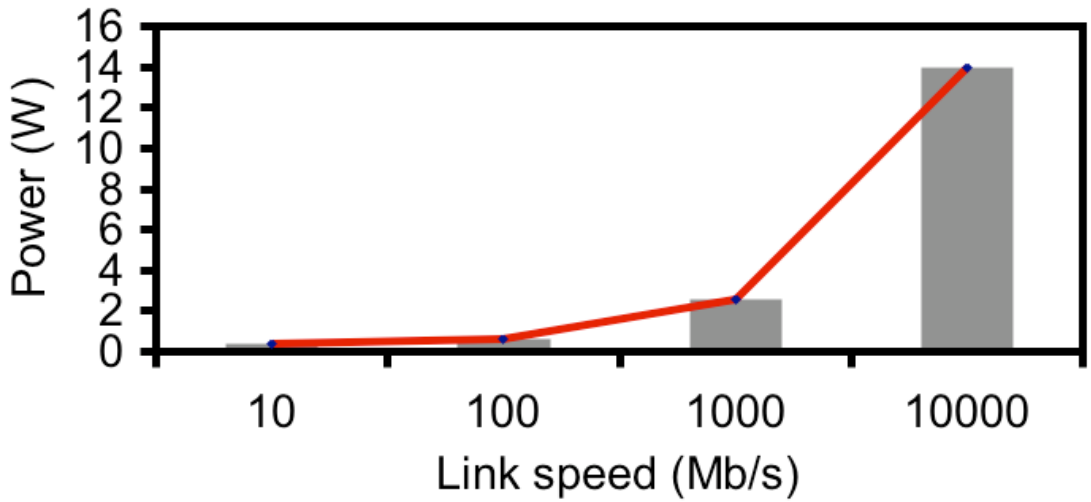
— all incremental AC power

— measuring 1st order

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



- **Various computer NICs averaged**



Potential Savings

Assumes ALL links become EEE capable and operable — actual savings will be some portion of this

- **Based on $0.08/kWh**
- **U.S. only**
- **Cooling savings excluded**

- **Estimate for 1 G**
  - **$250 million/year**
  - **Most NICs and most energy to be saved**
  - **Substantial benefits for homes and offices**
  - **Battery life benefit for notebooks**
- **Estimates for 10 G**
  - **$40 - $200 million/year**
  - **Depends on # of servers and # of 10G BASE-T ports/server**
  - **Reduces power burden in data centers**
  - **Reduces cooling burden in data centers**
  - **May increase switch/router port capacity**
- **Bottom line**
  - **Provides real economic benefit through substantial energy savings**
  - **Refinement of savings estimates will not affect standard rationale or content**

Energy Efficient Ethernet Beyond the PHY

Hugh Barrass, Cisco
Rudy Klecka, Cisco

A wider view

EEE study group has discussed saving power in the PHY

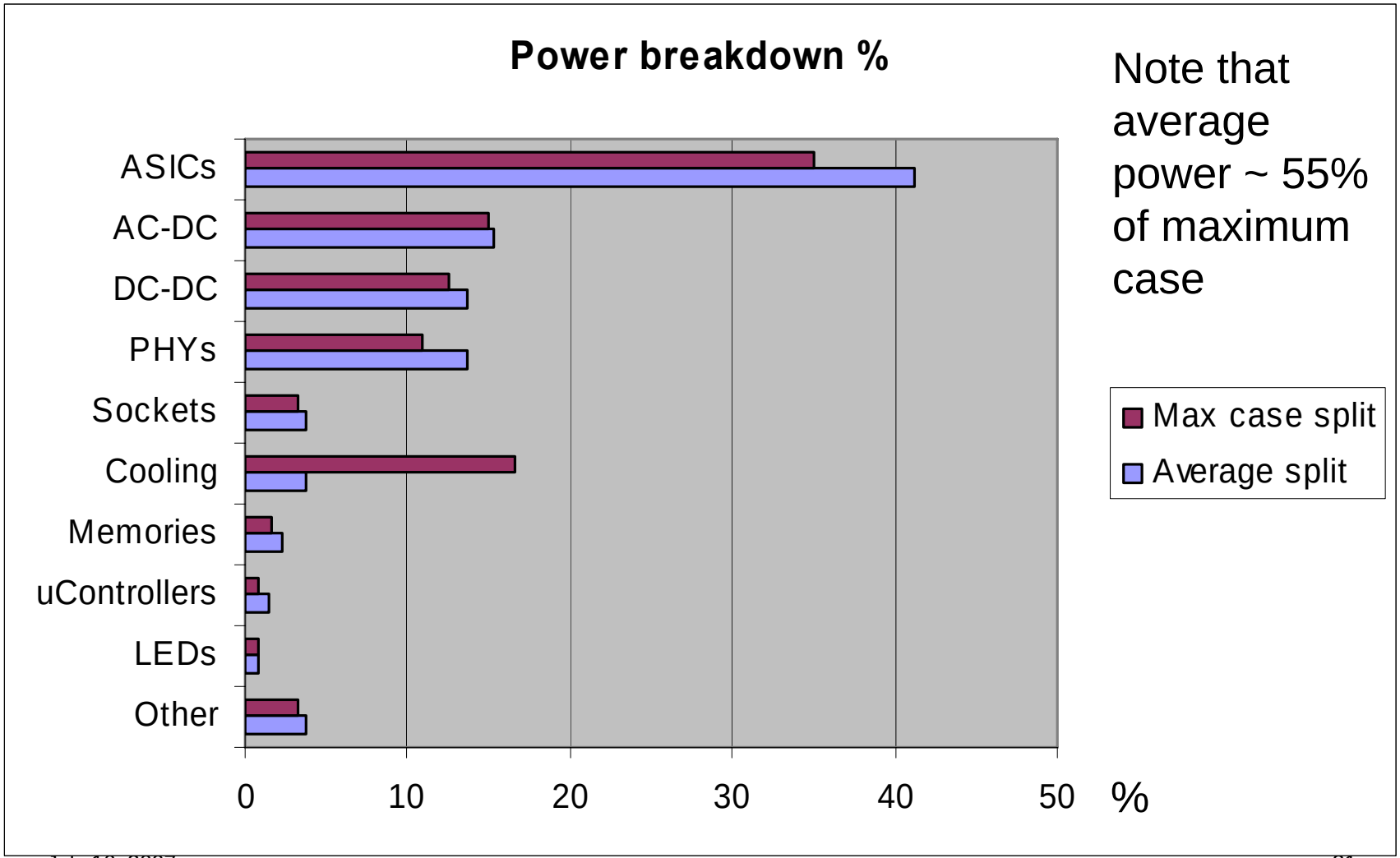
But whole system power measurements shown in CFI

- Power savings vs PHY speed > expected PHY power
- Even existing systems are saving more than PHY power

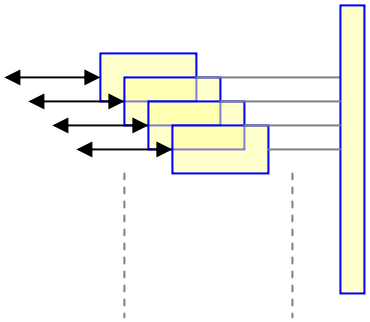
Examine current and potential system power savings

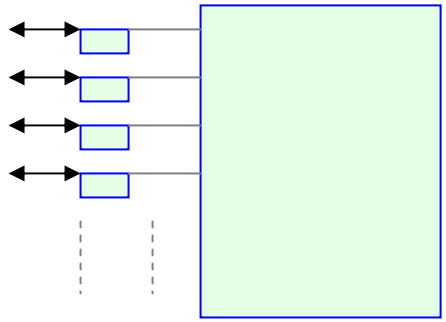
- “Reduction of power during low link-utilization”
- Where will this benefit from standards-based control?

Where does the system power go?



ASIC architectures





Port-based

- **Distributed memory structures & data paths**
- **Easy to reduce single port, link speed or RPS**
- **Power savings smaller – law of small numbers**
- **If memory structures used to absorb "return to activity" burst, cannot be powered down**

Centralized

- **Large central memories & data paths**
- **Power saving modes depend on thresholds**
- **Large power savings % for v. low activity**
- **May require port memories to absorb bursts, hinders efficiency**
- **Traffic characteristics = aggregation**

NB – power savings in ASIC memory structures discussed later

ASIC power savings

Existing architectures save power with link speed

- Static functions, based around auto-neg speed
- Saves 0 – 25% of ASIC power (up to 10% of system)
- Could benefit from RPS without architecture change

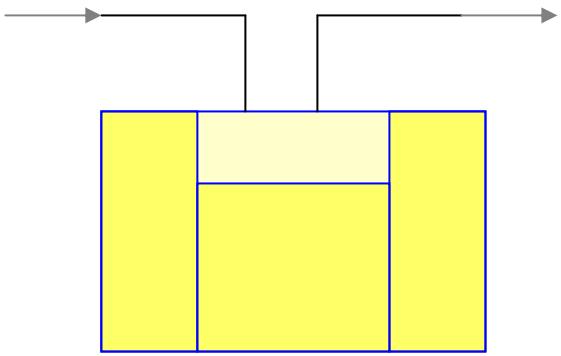
Newer architectures could increase savings

- If known changeover time before high speed burst…
- … allow more widespread shutdown (without buffer wastage)

High speed, centralized designs could save >75% of ASIC

- Up to 40% of system power
- Controlled speed change allows larger savings…
- … uncontrolled requires more conservative policy

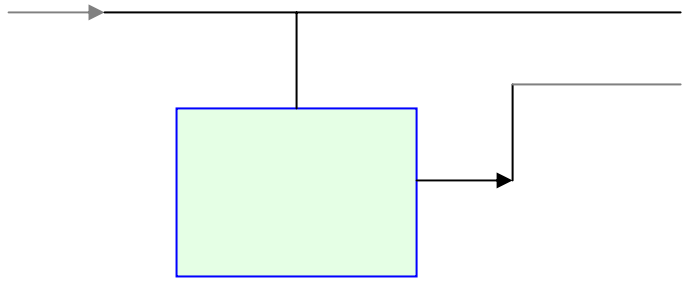
Memory types



Store-and-forward memory

Some savings come from reducing width

Bandwidth.delay sizing allows depth reduction, saving in exercising columns and refreshing



Lookup memory

Reduction in width for lower bandwidth

Some architectures may eliminate parallel copies for high bandwidth support

Dynamic copies may need reload for "return to activity"

Static memory power scales with activity

SSRAM benefits from reduction in clock speed or width

In conclusion

Significant power savings can be achieved with EEE!

EEE facilitates power savings throughout the system

- Beyond PHY power, savings >50% of typical

Next generation architecture should offer more savings

Secondary benefits of standard

- Common behavioral expectations, clearer benchmarking

Transition Time Considerations and Calculations

David Law, 3Com

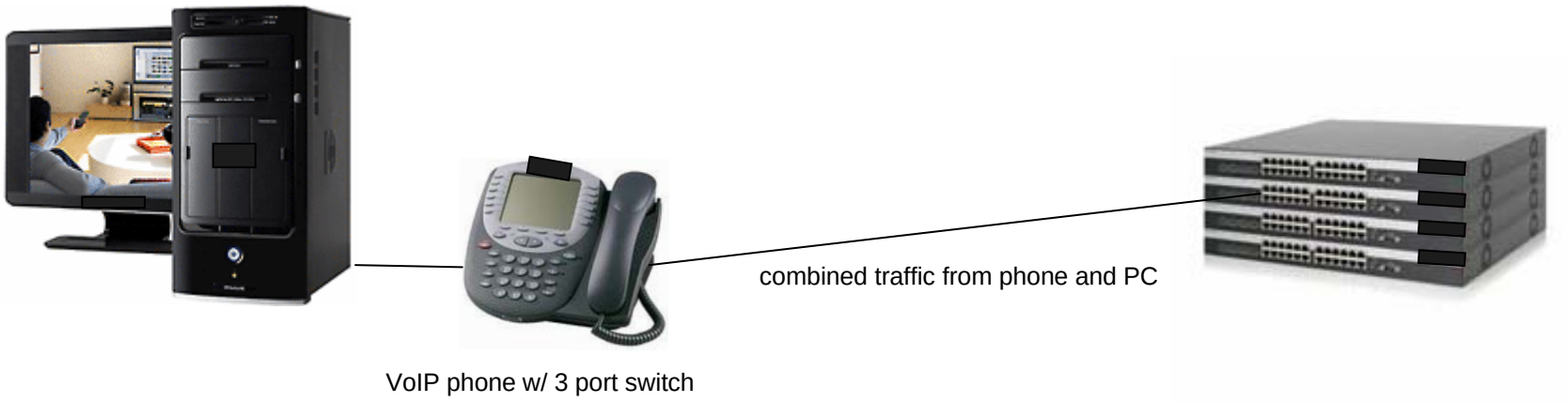
Wael Diab, Broadcom

Transition Time Considerations

- Ethernet links are widely deployed. They carry very diverse applications and higher layer protocols

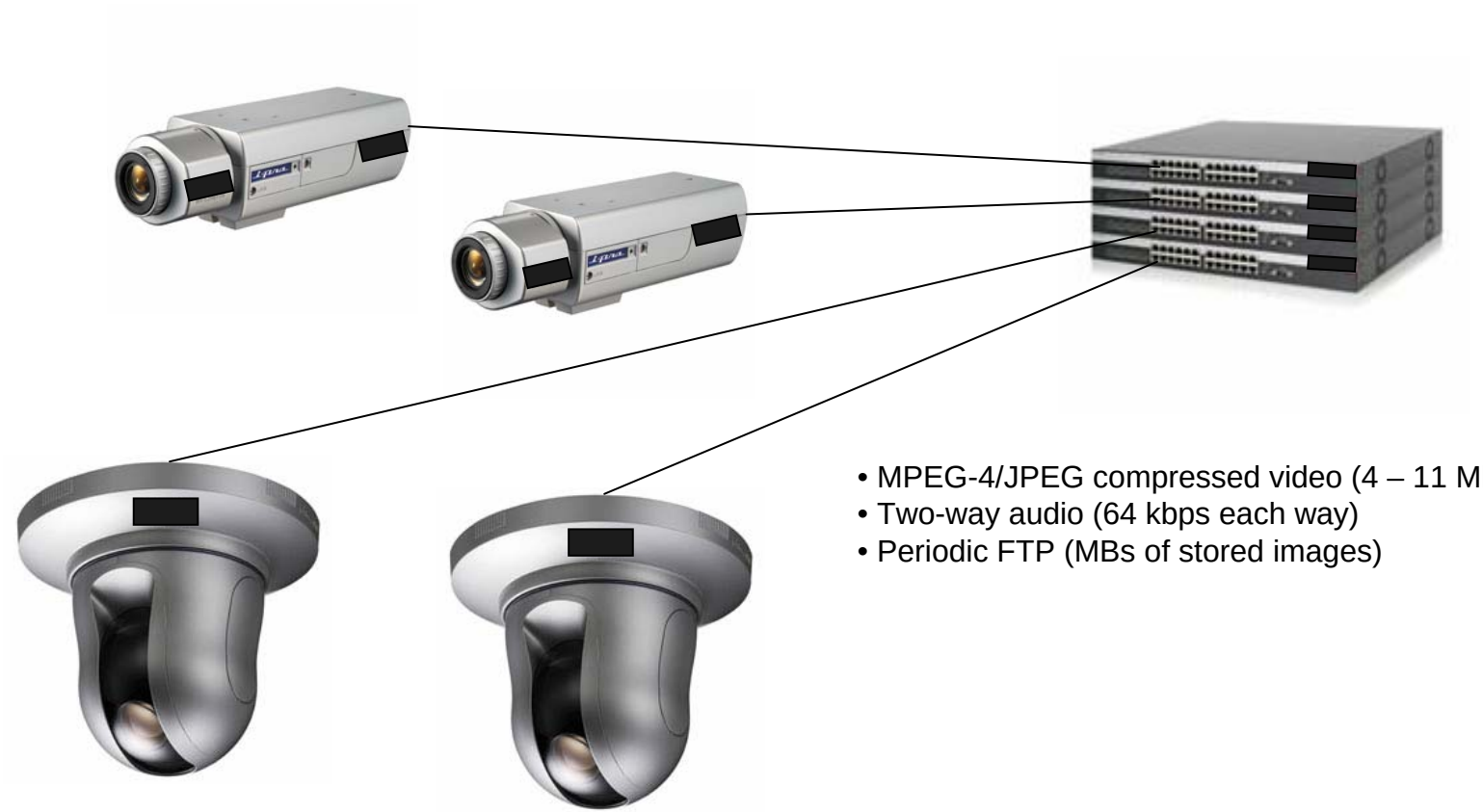
- A tradeoff exists for transition time
  - At one extreme, some applications may not be sensitive to transition time
    - E.g. Wake on LAN in consumer space
  - At another extreme, complete transparency favors shorter times
    - Latency sensitive applications E.g. AV Bridging, VoIP Phones etc.
    - Never need to “turn the feature off”
  - There are a number of ways to achieve the correct balance

Application – IP phone



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

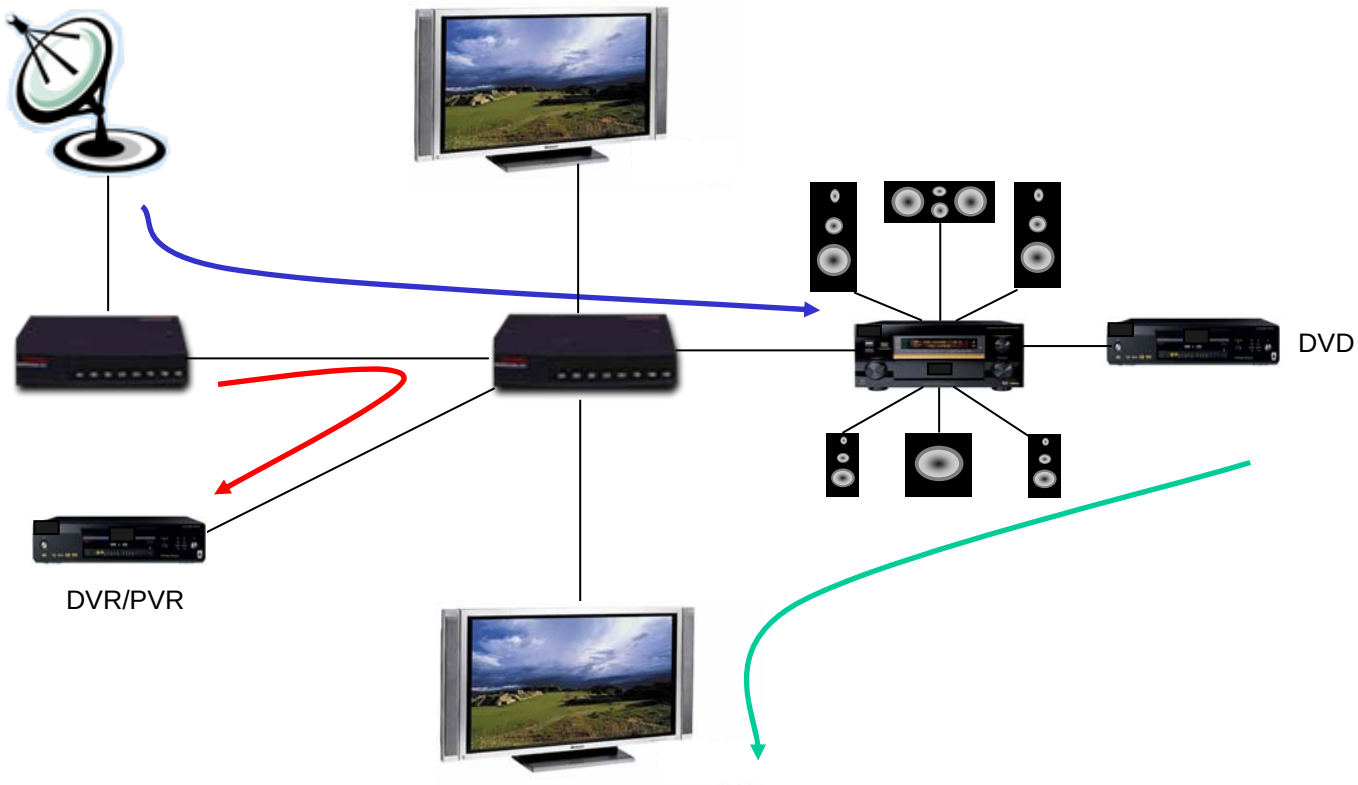
Application – surveillance camera



- MPEG-4/JPEG compressed video (4 – 11 Mbps)
- Two-way audio (64 kbps each way)
- Periodic FTP (MBs of stored images)

1. Cameras send MPEG-4 video to server and display consoles at 30 fps, ~4 Mbps
2. Cameras periodically send JPEG still images to server using FTP
3. When FTP session initiates, link will transition from 10 Mbps to 100 Mbps (possibly 1000 Mbps in the future)
4. Transition time must be less than ~15 ms to prevent frame loss

Application - EAV home network



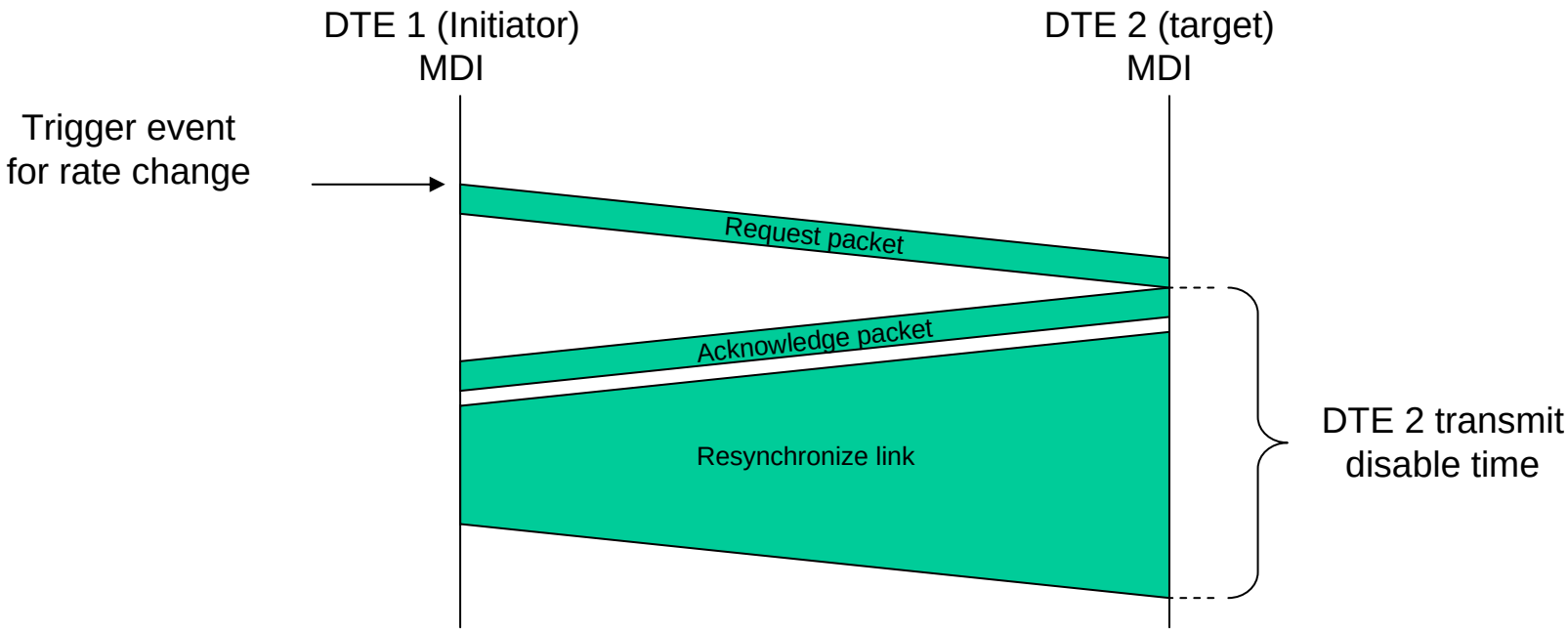
1. Listening to satellite radio on EAV receiver, link between receiver and switch operating at 10 Mbps
2. Start playing DVD on a screen in another room
3. Link between receiver and switch must transition from 10 Mbps to 100 or 1000 Mbps
4. Transition time must be less than 10 ms to avoid audible disruption
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

Application – file transfer

- Assume that a file transfer will invoke a transition from low power to higher bandwidth operation
  - depends on the control policy
- Depending on the transition time and file size, the transition time may be a significant fraction of the transfer time
  - The transition to higher bandwidth might actually increase the file transfer time

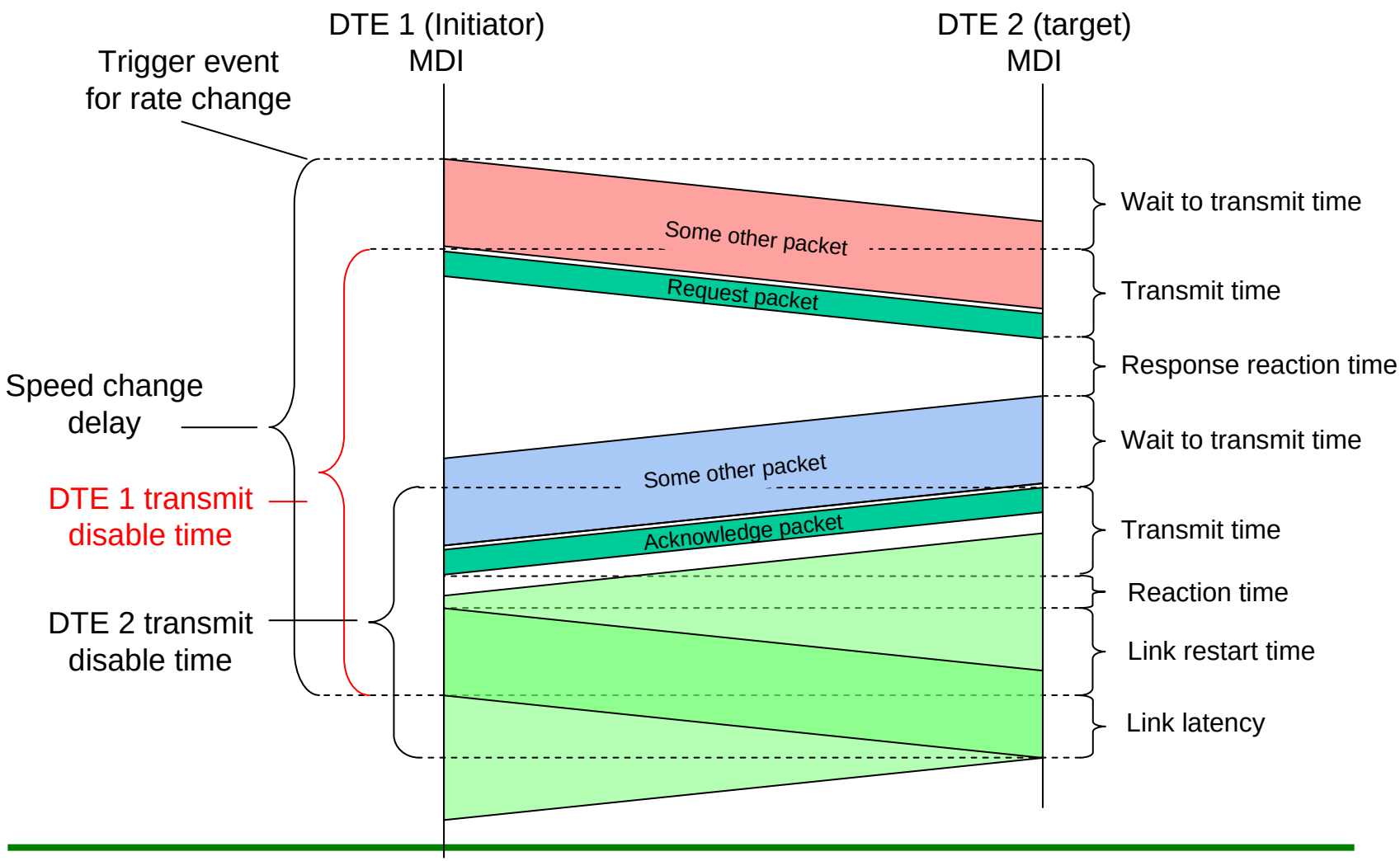
File size (MB)	Transfer Time @ 10 Mbps (ms)	Transfer Time @ 100 Mbps (ms)	Transfer Time @ 1000 Mbps (ms)	Transfer Time @ 10000 Mbps (ms)
0.01	8	0.8	0.08	0.008
0.1	80	8	0.8	0.08
1	800	80	8	0.8
10	8000	800	80	8

Link disable time – model so far



Notes:
1 - Ignores error recovery such as Acknowledge packet being lost

Transmit disable calculation



Example calculation – worst case

Parameter	Calculation	Bits	10Mb/s	100Mb/s	1Gb/s	10Gb/s
			us	us	us	us
Wait to transmit time Note 1	**(Max Packet + IPG) x 8 x BT**	**16160**	**1616**	**161.6**	**16.16**	**1.616**
Transmit time Note 2	**Min Packet x 8 x BT + link delay**	**576**	**58.17**	**6.33**	**1.146**	**0.627**
Response reaction time	**Same as required for pause**	**Note 3**	**57.6**	**5.76**	**1.024**	**3.072**
Wait to transmit time	**(Max Packet + IPG) x 8 x BT**	**16160**	**1616**	**161.6**	**16.16**	**1.616**
Transmit time	**Min Packet x 8 x BT + link delay**	**576**	**58.17**	**6.33**	**1.146**	**0.627**
Reaction time Note 4	**(Same as required for pause)/2**	**Note 3**	**28.8**	**2.88**	**0.512**	**1.536**
Link restart time	**Link speed increase**	**Note 5**	**2500**	**2500**	**1000**	**n/a**
	Link speed decrease	**Note 5**	**n/a**	**1**	**2500**	**2500**
Totals for link speed increase						
Speed change delay during link speed increase				**5934.7**	**2844.5**	**1036.0**
DTE 1 (initiator) transmit disable time during link speed increase				**4318.7**	**2682.9**	**1020.0**
DTE 2 (target) transmit disable time during link speed increase				**2587.6**	**2509.8**	**1002.2**

Notes:

1 **Maximum packet size = Envelope frame + SFD + Preamble = 2008 Bytes**

2 **Minimum packet size = Minimum frame + SFD + preamble = 72 Bytes; Link delay = 5.7ns/meter x 100 = 0.57 us**

3 **For operating speeds of 100 Mb/s or less response time = pause_quantum + 64 = 512 + 64 = 576 Bits**
For an operating speeds of 1000 Mb/s response time = two pause_quantum = 2 x 512 = 1024 Bits
For an operating speeds of 10 Gb/s response time = sixty pause_quantum = 60 x 512 = 30,720 Bits

4 **This delay could be included in link restart time but some delay has to be allocated for PHY latency and packet processing that is discrete from PHY restart**

5 **Based on data in chadha_1_0407.pdf, suggestion EEE 1000BASE-T mode in woodruff_01_0307.pdf and nominal value for 10BASE-T of 1us.**

Some observations

- Transmit disable time on single link can be a number of milliseconds
  - Is this acceptable for upper layer protocols
- Transmit disable time asymmetric
  - Longer for the requesting end
- Transmit disable time packet size dependant at slower speeds
  - Average time will be less than maximum
- Transmit disable time PHY restart dependant at higher speeds
  - Average time very similar to maximum
- Assuming EEE 10BASE-T is lowest power by significant margin
  - 10/100/1000BASE-T ports will have broad market potential
    - 1000BASE-T to provide best performance for cost
    - EEE 10BASE-T to provide lowest power operation

Some observations

- Hence 10BASE-T to 1000BASE-T speed change will be important
  - 10BASE-T to 1000BASE-T via 100BASE-T
    - 10BASE-T to 100BASE-T
      - 6 ms maximum, 3ms min
    - 100BASE-T to 1000BASE-T
      - 2.5 ms minimum (no packets other than request to increase to 1000BASE-T)
    - Total 8.5ms maximum, 5.5ms minimum
  - 10BASE-T to 1000BASE-T direct
    - 6 ms maximum, 3ms min

Transition Time Conclusions

- **Applications require sub 10 ms transition time**
- **Recommend that the EEE TF retain the goal of achieving a transition time of less than or equal to 1 ms**

Technical Feasibility of Energy Efficient Ethernet

Wael Diab, Broadcom

George Zimmerman, Solarflare

Technical Feasibility for Link Utilization Mechanisms

- A number of mechanisms have been presented to the SG that address the mechanism to achieve lower power consumption during period of low link utilization

- There is a tradeoff between the following factors
  - Simplicity of implementation and technology reuse
  - Simplicity of modifications to the standards specification
  - Extent of power conservation
  - Quickness of the transition time

- This list attempts to categorize the possible solution space
  - Standard auto-negotiation
  - Fast auto-negotiation
  - Fast start
  - Subset PHY

Possible 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 a lower power consumption PHY during periods of low link utilization as a subset of the higher speed standard PHY**

Fast Start: Overview

- Be safe: do no harm
  - Base results on WORKING systems
  - MAC interfaces are assumed constant
  - No change to operational mode of existing PHYs
- Be lazy: don't invent unnecessary things
  - Transition would minimally impact existing specifications
    - Reuse of existing 802.3an PHY control as much as possible
- Be quick: get PHY transition times down
  - Need for transitions of <10msec, pref ~1msec
  - Transition time and feasibility are controlled by re-entry to the high rate state
  - Need to minimize retraining time

Fast Start: Approach and Feasibility Questions

- **Freeze stored PHY state while lower speed was running**
  - **Feasibility Question: How long before the transceiver state typically gets stale?**
- **Restart 10GBASE-T transmission by entering final stages of PHY-control startup sequence**
  - **Feasibility Question: How short might a transition be made using the existing or minimally changed framework**

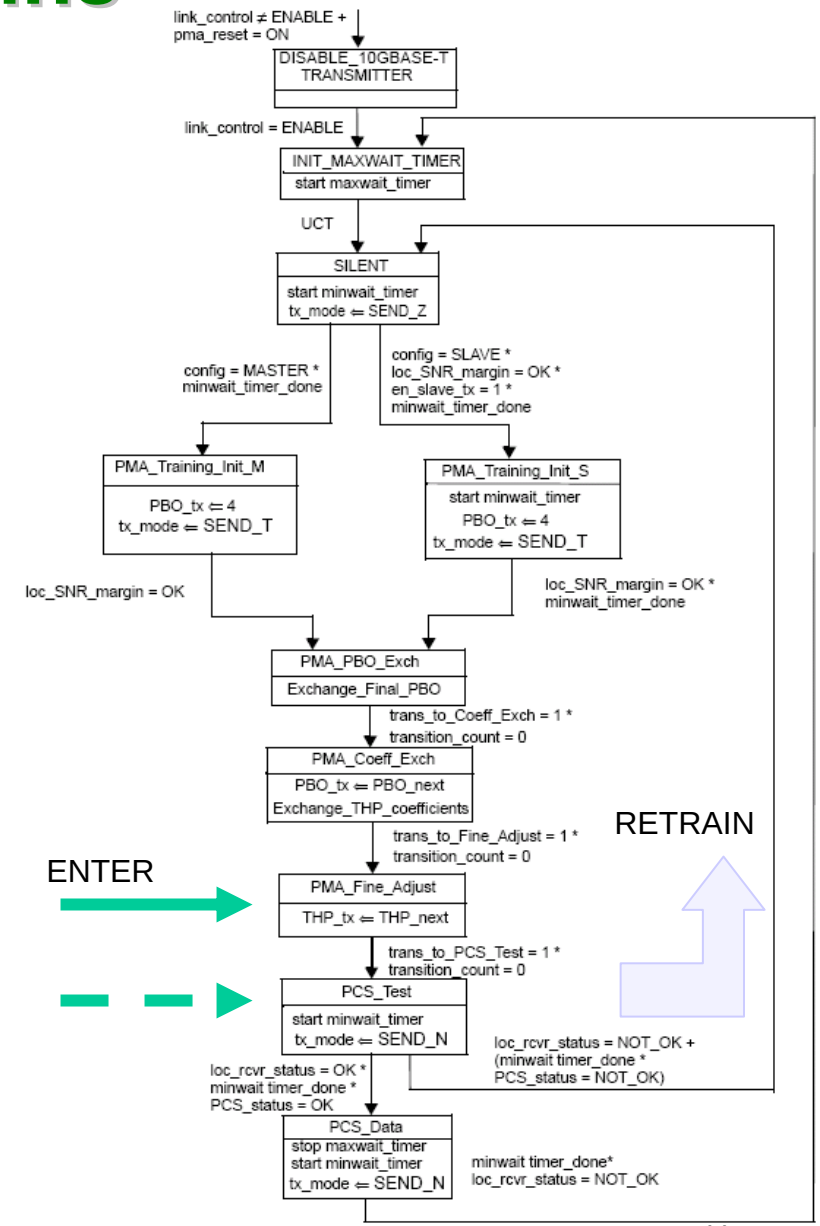
Fast Start: Stored-State Longevity*

- Experimental data gathered showed longevity of stored adaptive filter state:
  - 10GBASE-T PHYs operating on 100m Cat 6a (zimmerman_1_0307)
  - 1000BASE-T PHYs (chadha_1_0407)
- Error-free link established on worst-case link segments
  - Receiver adaptive filters were frozen and time to link errors was observed
  - 10GBASE-T Link degradation began ~ 3 minutes after freeze, recoverable up to ~ 5 minutes
    - Note: Timing recovery was also disabled
  - 1000BASE-T link solid overnight (but had timing recovery going)
- Conclusion:
  - Stored state can be used with minimal retune up to 3 minutes, staged, fast retrain may be required after 5 minutes.

***Note: this summarizes separate vendors' experiments**

Fast Start: Fast Retrains

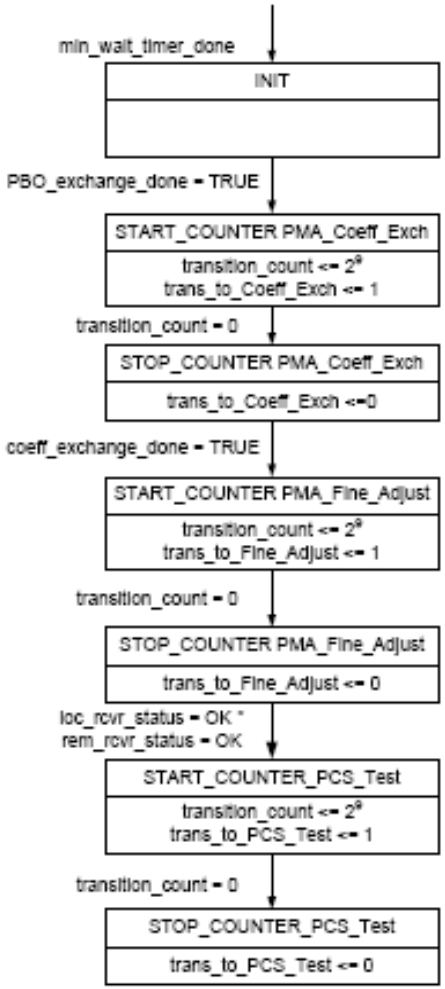
- 10GBASE-T PHY Control State Machine (Fig. 55.4.6.1)
- Entrance points for EEE state-restoral:
  - PMA_Fine_Adjust or
  - PCS_Test (1msec fixed)
    - Required to maintain quality
    - Test time limited by desire to see enough LDPC frames
  - PMA_Fine_Adjust entry needed after 5 minute point
- Full Retrain triggered if PCS_Test fails, dropping link
  - May be true for reentry from subset PHYs as well



Fast Start: Minimal Changes

- **Transition time is controlled by 512 infofield countdown (each count tick = 20.48usec)**
  - **Prior rationale allowed for controller sync to be sloppy – not consistent with EEE assumptions of ~1msec transitions**
- **Change transition_count value in Fig. 55-25 (MASTER transition count) from 2^9 to 2^3**
  - **Corresponding change of transition count for response in Fig. 55-26 (Slave) from 2^6 to 2^2**
- **Minimum PMA_Fine_Adjust time reduced from 10.5 msec to 164 usec**
  - **Limitation should now be training time, not protocol**
  - **Still allows plenty of time (>50usec) for Master-Slave state change synchronization**
- **1msec PCS_Test state time remains**

- Enables transitions down to ~1-2 msec by reducing unnecessary overhead with minimal standards changes



Fast Start: Fast Training Test Results

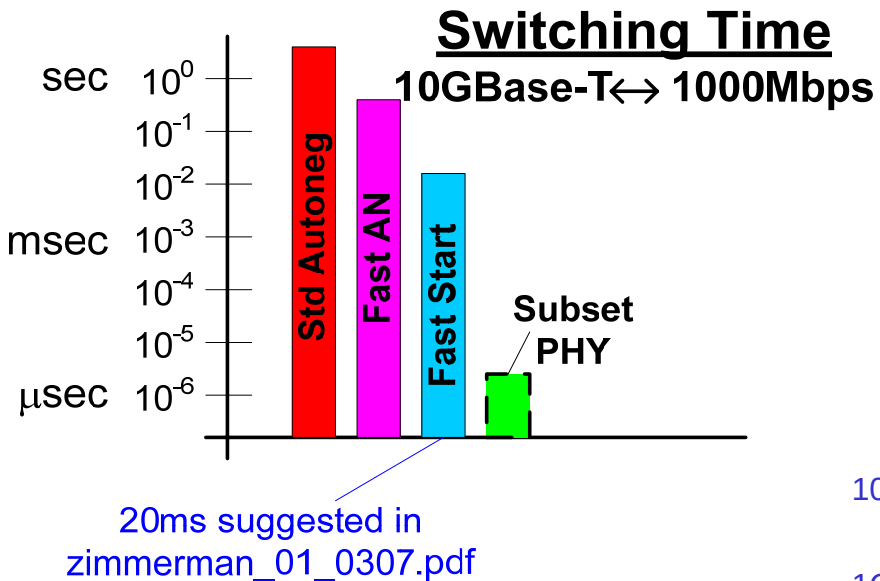
- **Question: Can a transceiver be fine-trained in 1-5 msec?**
- **Experiment set up to mimic worst-case retrain (all equalizers and cancellers need adjustment)**
  - **10GBASE-T link setup with 4 connector, 100m channel**
  - **Link trained with modified counters on transition to PCS_test**
  - **Timing and phase readjust at entry to PMA_Fine_Adjust state**
  - **Training time at PMA_Fine_Adjust varied to determine limitations**
- **Consistent demonstrations show SNR and Ethernet Frame Error Rate are uncompromised by 3-4 msec retrain time**
- **Similar results for 1000BASE-T showed timing reacquisition at 2-2.5msec for existing PHYs (chadha_1_0407.pdf)**
  - **Conservative estimate of <5msec rate transitions**

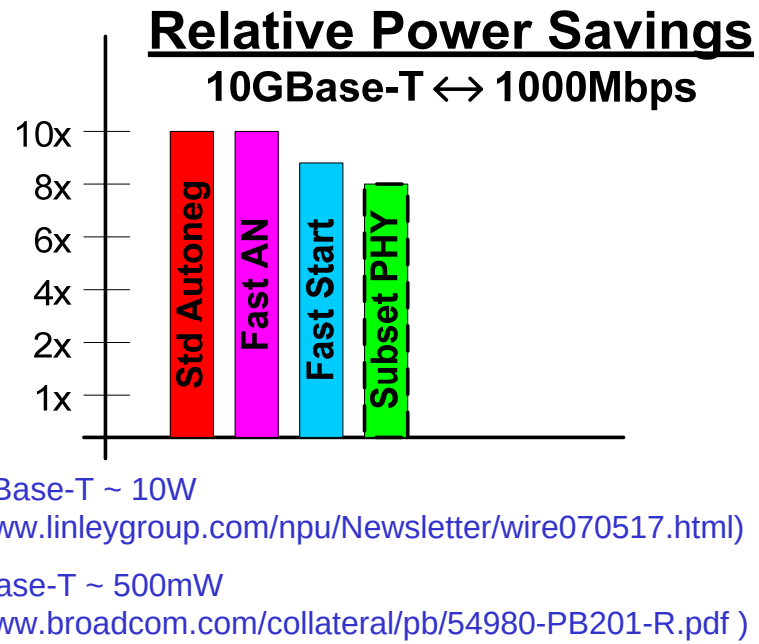
Fast Start: Conclusions

- **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: 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**





- **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: A 10G Example

- Adaptations of 10GBASE-T to operate at GE rates
- Line code: PAM-16 -> PAM-4
- Reduced number of channels (each direction): 4 -> 1
- Simplex operation at 800MS/S
  - Produces 1.6Gb/s raw bit rate
- Zero Stuffing, or equivalent, to match rates

Note signaling rate remains at 800MS/S

- Intent is to minimize changes in behavior as an aggressor

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

Control Policy and Coordination Protocol for Transition

Control Policy

- PHY transition is triggered by the system's control policy
- Definition of control policy is outside the scope of IEEE 802.3

Coordination Protocol

- When an EEE capable device decides to make a transition based on its control policy, it needs to coordinate with its link partner
- EEE will define a protocol to do this

Energy Efficient Ethernet Objectives

Wael Diab

Broadcom

Link Utilization Objectives 1 of 4

- 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

Link Utilization Objectives 2 of 4

- 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

- Goal is to allow consideration of a mechanism that can address high power consumption on widely deployed links during low utilization
- Listed PHY types are types that would be considered for low link utilization. This allows falling back to
  - one of the listed types
  - non-listed standard type (e.g. 10BASE-T)
  - a subset type (e.g. subset PHY)
  - a new type (e.g. “0BASE-T”)
- Objective allows various mechanisms to be considered

Link Utilization Objectives 3 of 4

- 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

- Goal of the 1st listed objective here is to allow both ends of the link to communicate to one another about a desire to move from one level of power consumption to the next. A non-exhaustive list of example protocols could be
  - LLDP (802.1AB / 802.1AB-REV)
  - OAM (802.3ah)
  - other slow protocol (802.3ad)
  - MAC Control frame
  - Physical layer signaling
- Goal of the last 2 listed objectives is to shield upper layers from transition. Specifically, a change in link state or dropped frame may be detected, reported and/or trigger an upper layer change

Link Utilization Objectives 4 of 4

- The transition time to and from the lower level of power consumption should be transparent to upper layer protocols and applications

- Goal of this objective is at best to minimize the impact to upper layer protocols and applications from the transition time by bounding it
- There are a number of scenarios that have been identified and under study. E.g.
  - VoIP Phones
  - Surveillance cameras
  - AV Bridging
  - File transfers
  - TCP/IP
- The choice of a transition time will affect the choice of the solution. The group is working on understanding these scenarios and establishing a framework to select a time or set of time(s)

10BASE-T Voltage Objective

- **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**

- **Goal of this objective is reduce power consumption of 10BASE-T by reducing the envelope voltage**
- **Additionally, this allows for potentially easier implementations in silicon as it removes the requirement for legacy voltage which in some cases make legacy compatibility economically prohibitive**
- **Note that this objective will not support 100m over Category 3**

Auto Negotiation Objective

- **Any new twisted-pair and/or backplane PHY for EEE shall include legacy compatible auto negotiation**

- **Goal of this objective is to do no harm to the auto negotiation in place today for PHY types over twisted-pair and/or backplane by requiring compatibility with legacy auto negotiation**
- **This objective does not imply that EEE uses auto negotiation**

PAR and 5 Criteria

Mike Bennett
Lawrence Berkeley National Laboratory

PAR Scope

The proposed standard will include a symmetric protocol to facilitate transition to and from lower power consumption in response to changes in network demand. The transition will not cause loss of link as observed by higher layer protocols. The project will also specify PHY enhancements as required for a selected subset of PHY types to improve energy efficiency.

PAR Purpose

Most Ethernet links have significant periods of low utilization or no utilization for application data traffic. This project will take advantage of this to provide energy savings in the PHY and enable energy savings in the system which will deliver reduction in total cost of operation.

PAR Need

Market pressure and legislative action worldwide is demanding improvements in energy efficiency of networked systems. Energy costs are a major component of operating cost. Energy Efficient Ethernet (EEE) features will be explicitly or implicitly required by a significant fraction of Ethernet edge connections in the future. Energy consumption and efficiency will become a major factor in the choice of network solutions, especially in data centers. EEE capabilities will be important as Ethernet becomes an enabler for low duty cycle, consumer class applications. EEE capabilities will enable new system level energy management techniques that will save energy beyond the network interface. EEE will address interface changes required to improve energy efficiency.

PAR Stakeholders

Ethernet is pervasive, with a consequent pervasive set of stakeholders. This includes and is not limited to: component providers (e.g., cabling and integrated circuit), system product providers (e.g., switch and NIC), network providers (e.g. installers, network support, enterprise network implementers), bandwidth providers (e.g., carriers), software providers (e.g., network management), and the users of any of these products or services.

Broad Market Potential

- Broad set(s) of applications
- Multiple vendors, multiple users
- Balanced cost (LAN vs. attached stations)

Market pressure and legislative action worldwide is demanding improvements in energy efficiency of networked systems. Energy costs are a major component of operating cost. EEE features will be explicitly or implicitly required by a significant fraction of Ethernet edge connections in the future.

Energy consumption and efficiency will become a major factor in the choice of network solutions, especially in data centers. EEE capabilities will be important as Ethernet becomes an enabler for low duty cycle, consumer class applications.

EEE capabilities will enable new system level energy management techniques that will save energy beyond the network interface. EEE will address interface changes required to improve energy efficiency.

Ethernet equipment vendors and customers are able to achieve an optimal cost balance between the network infrastructure components and the attached stations.
(Adopted 4/18/07)

Compatibility

- IEEE 802 defines a family of standards. All standards shall be in conformance with the IEEE 802.1 Architecture, Management, and Inter-working documents as follows: 802. Overview and Architecture, 802.1D, 802.1Q, and parts of 802.1f. If any variances in conformance emerge, they shall be thoroughly disclosed and reviewed with 802.

  •Each standard in the IEEE 802 family of standards shall include a definition of managed objects that are compatible with systems management standards.

It is expected that Energy Efficient Ethernet will conform with the 802 Overview and Architecture and remain compatible with 802.1D, 802.1Q and 802.1f. The project will work with 802.1 to address any extensions to these standards if required and to encourage their work to take advantage of the features that this project will provide.

As an amendment to IEEE Std 802.3, the proposed project will follow the existing format and structure of 802.3 MIB definitions.

Incompatibility with legacy PHYs (e.g. operational conditions and media types) will be addressed in terms of market relevance. The proposed standard will include a 10 Mb/s PHY that may not support full 100m of category 3 cable.

(Adopted 5/30/07)

Distinct Identity

a) Substantially different from other IEEE 802 standards
b) One unique solution per problem (not two solutions to a problem)
c) Easy for the document reader to select the relevant specification

---

This project will provide capabilities that are specifically for IEEE 802.3 links and IEEE Std 802.3 does not address energy efficiency. For example, there is no mechanism to allow a change of PHY speed without dropping link and renegotiation.

We may introduce specifications to optimize existing PHYs. Where appropriate, these optimized PHYs will only be accessed through EEE.

The proposed project will be formatted as a amendment to IEEE Std 802.3, making it easy for the document reader to select the EEE specification.

(Adopted 5/30/07)

Technical Feasibility

a) Demonstrated system feasibility
b) Proven technology, reasonable testing
c) Confidence in reliability

Energy efficiency techniques based on reducing capabilities to lower power consumption have been broadly deployed and used. The technology to be utilized in the realization of the EEE PHY will rely heavily on previous 802.3 standards.

The study group expects the proposed standard to use existing PHYs where possible. When necessary to meet the objectives, the proposed standard may include modified PHYs.

The control mechanism will build upon well known simple protocols.

The latency variation introduced by EEE is expected to be transparent to most upper layer protocols. EEE will define control, status, and management so that other protocols can be informed of the state of EEE.

Confidence in the energy saving effectiveness and system feasibility of selected proposals will be demonstrated through simulation of typical applications and usage; in conjunction with input from higher layer networking experts.
(Adopted 5/30/07)

Economic Feasibility

a) Known cost factors, reliable data
b) Reasonable cost for performance
c) Consideration of installation costs

EEE will not materially impact component or installation costs, and may provide cost savings opportunities.

While EEE is within IEEE 802.3, the creation of EEE provides opportunities for energy savings beyond the PHY, potentially of much greater magnitude than the PHY itself.

The control mechanism will use similar functions to those already included in most Ethernet equipment and therefore will not add any significant cost.

The energy savings achieved will result in lower operating costs.

(Adopted 5/30/07)

Wrap-up

Mike Bennett
Lawrence Berkeley National Laboratory

Summary

- EEE can and will save energy
- Our goal is to achieve these savings with minimal impact on the standard and on the industry
- EEE is feasible
- EEE enables:
  - Energy savings beyond the PHY
  - Energy Industry incentives for savings
- Estimated completion date: March 2010
- Website: http://grouper.ieee.org/groups/802/3/eee_study/index.html

  Up to date PAR, 5 Criteria, objectives are available there

Questions?

Thank you!