Reducing the Energy Consumption of Networked Devices

Bruce Nordman
Energy Analysis
Lawrence Berkeley National Laboratory
Berkeley, CA 94720
bnordman@lbl.gov

Ken Christensen
Computer Science and Engineering
University of South Florida
Tampa, FL 33620
christen@cse.usf.edu

 IEEE 802.3 tutorial – July 19, 2005 (San Francisco)

BERKELEY LAB

USF UNIVERSITY OF SOUTH FLORIDA

Acknowledgement

- We would like to thank Bob Grow for inviting us
- We hope that you will get useful information from this tutorial

Topics

- Energy use by IT equipment Part 1
- Overview of power management Part 2
- Reducing network induced energy use Part 3
- Reducing network direct energy use Part 4
- Potential energy savings Part 5
- Summary and next steps Part 6

Berkeley Lab
USF UNIVERSITY OF SOUTH FLORIDA

Background - Key Terms

Networked Device

- An electronic product with digital network connection, either a piece of network equipment or end use device.

Network Equipment

- Products whose only function is to enable network communications (Switches, routers, firewalls, modems, etc.)

Energy

- Direct electricity consumed by electronic devices. Does not include extra space conditioning energy, UPS, etc.

- All $ figures based on $0.08/kWh
  - 1 TWh = $80 million
  - $1 billion = 12.5 TWh
  - 1 W/year = 70 cents

BERKELEY LAB

USF UNIVERSITY OF SOUTH FLORIDA

Energy use by IT equipment

- Welcome to Part #1

In this part…the energy consumption of IT generally and PCs specifically.

Current IT energy use: All IT equipment

- **"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

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

PCs and etc. already digitally networked — *Consumer Electronics* (CE) will be soon

Current IT energy use: All IT equipment <u>continued</u>

- **" Little IT" — office equipment, telecom, data centers**
  - 97 TWh/year (2000) [Roth] — 3% of national electricity; 9% of commercial building electricity

Commercial buildings only

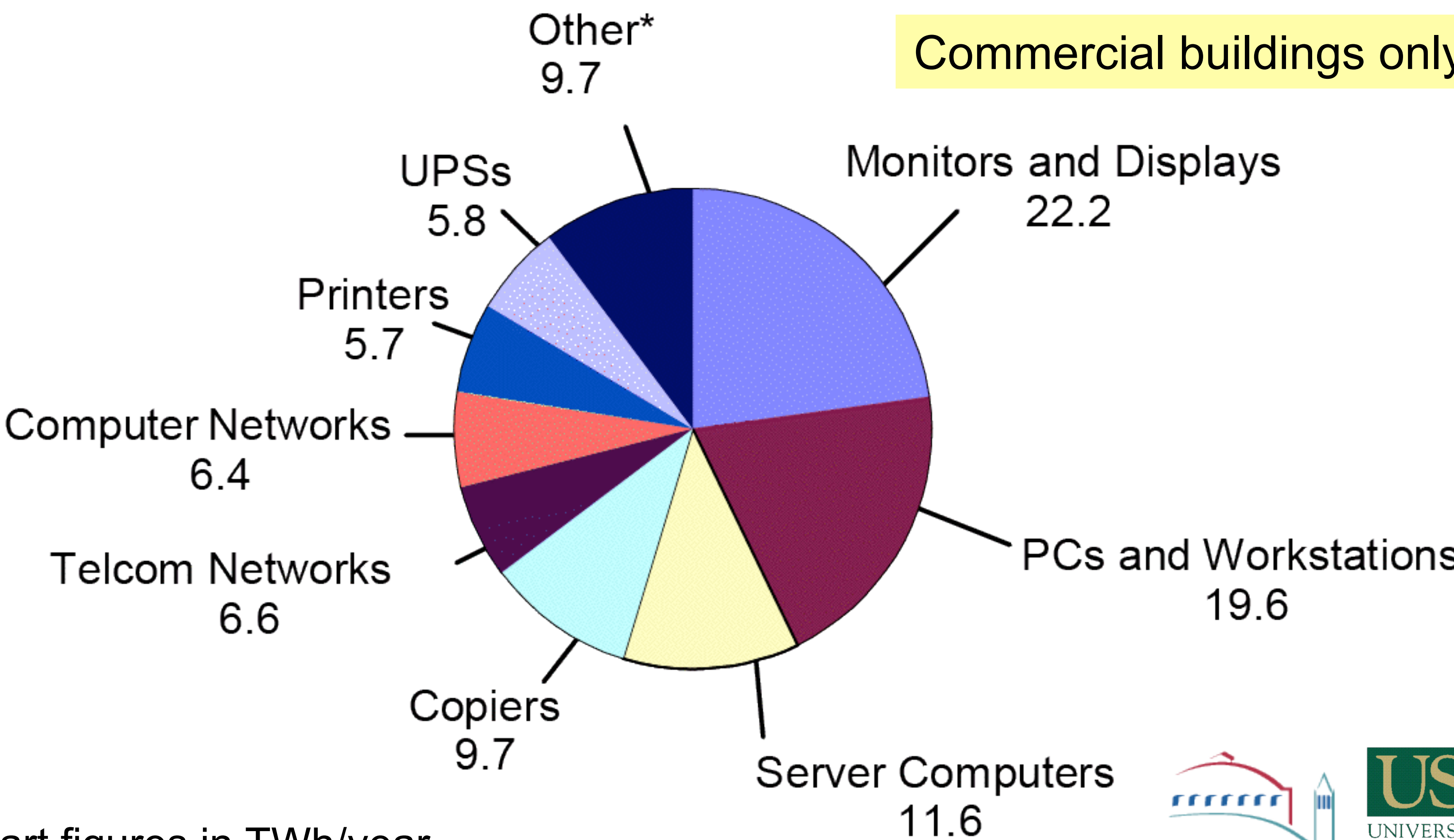
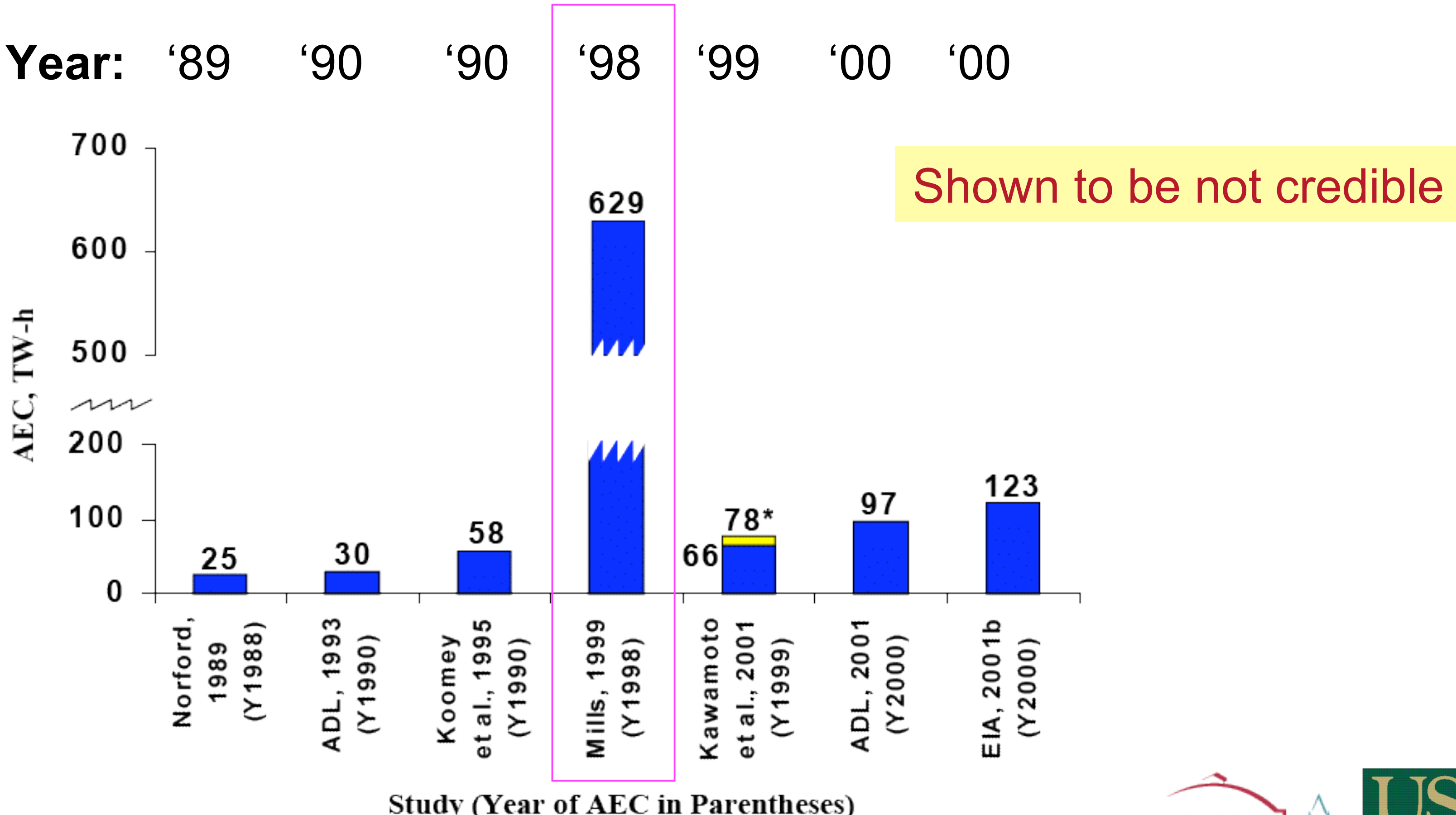


Chart figures in TWh/year

Current IT Energy Use: Huber / Mills “Analysis”

- **1999: Forbes, *Dig more coal -- the PCs are coming***
  - **Claim**: “Internet” electricity 8% in 1998 and growing to 50% over 10 years



Huber/Mills compared to other studies

PC energy use

❖ PCs

- Computing box only — not including displays
- PCs: **31 TWh/year** (2000)

  ➔ **$2.4 billion/year**

- Servers: 12 TWh/year (2002)
- PC energy use could be **46 TWh/year** by now and is rising steadily

  ➔ **$3.7 billion/year**

BERKELEY LAB

USF UNIVERSITY OF SOUTH FLORIDA

PC energy use: 24/7 PC example

- **Bruce's home PC and display***

	On	Sleep	Off
Computer	**57.5 W**	**7.5 W**	**6.0 W**
Display	**17**	**2**	**2**

  - Display can power manage – On 20 hours/week; Sleep 148
  - Computer can't **(and stay on network)** – On 168 hours/week

- **Annual consumption**
  - 540 kWh/year
  - ~$70/year

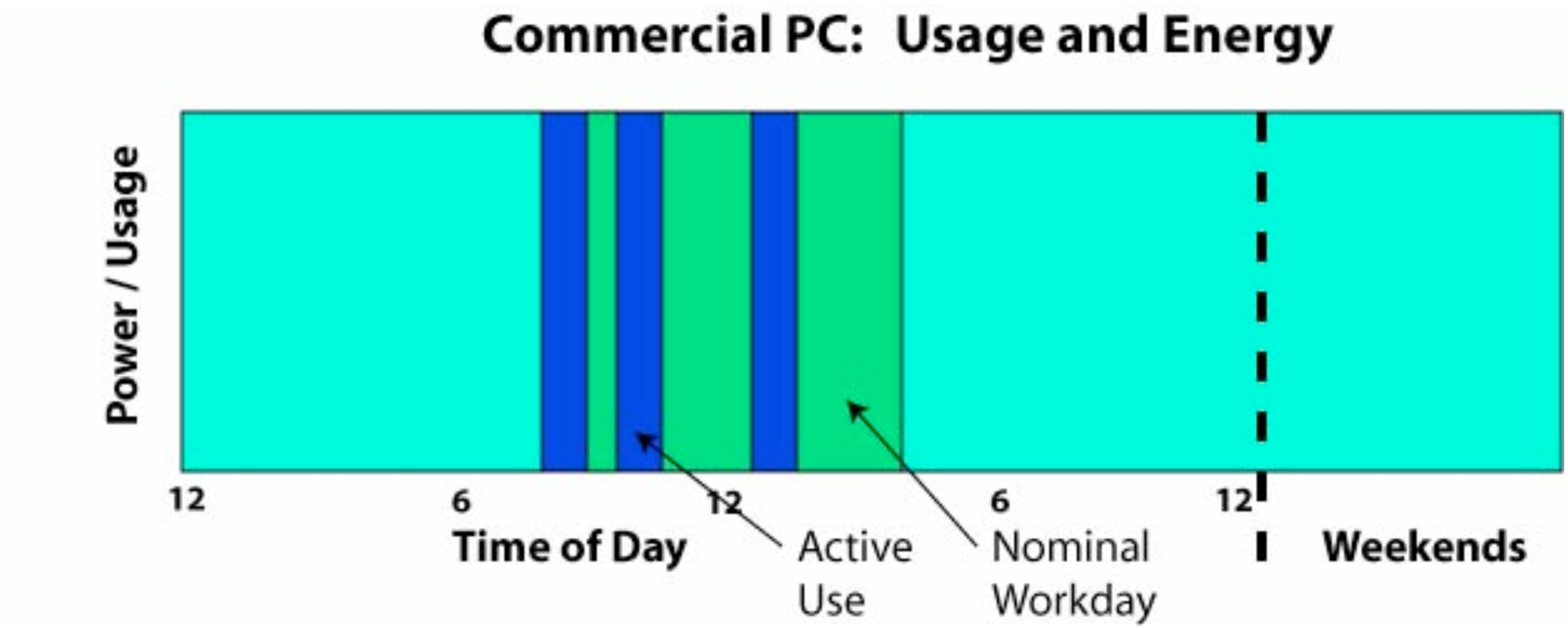
16% of current annual electricity bill

* Bruce doesn't leave the PC on 24/7

BERKELEY LAB

USF UNIVERSITY OF SOUTH FLORIDA

PC energy use: How PCs use energy



- ❖ **Active use is a small part of week**
  - Energy use is not closely related to activity
- ❖ **Most commercial PCs are on continuously**
  - Increasingly true for residential PCs
  - Most of time, highly powered but doing little or no work

Savings opportunity!

BERKELEY LAB

USF UNIVERSITY OF SOUTH FLORIDA

PC energy use: Factors

Many figures here are not well known,
but conclusions do not rely on precision

- **Annual PC energy consumption is a function of**
  - **Power** levels — in each major operating mode
  - **Usage** patterns — % of year by mode

→ **Unit annual energy use**

  - The **stock** of PCs

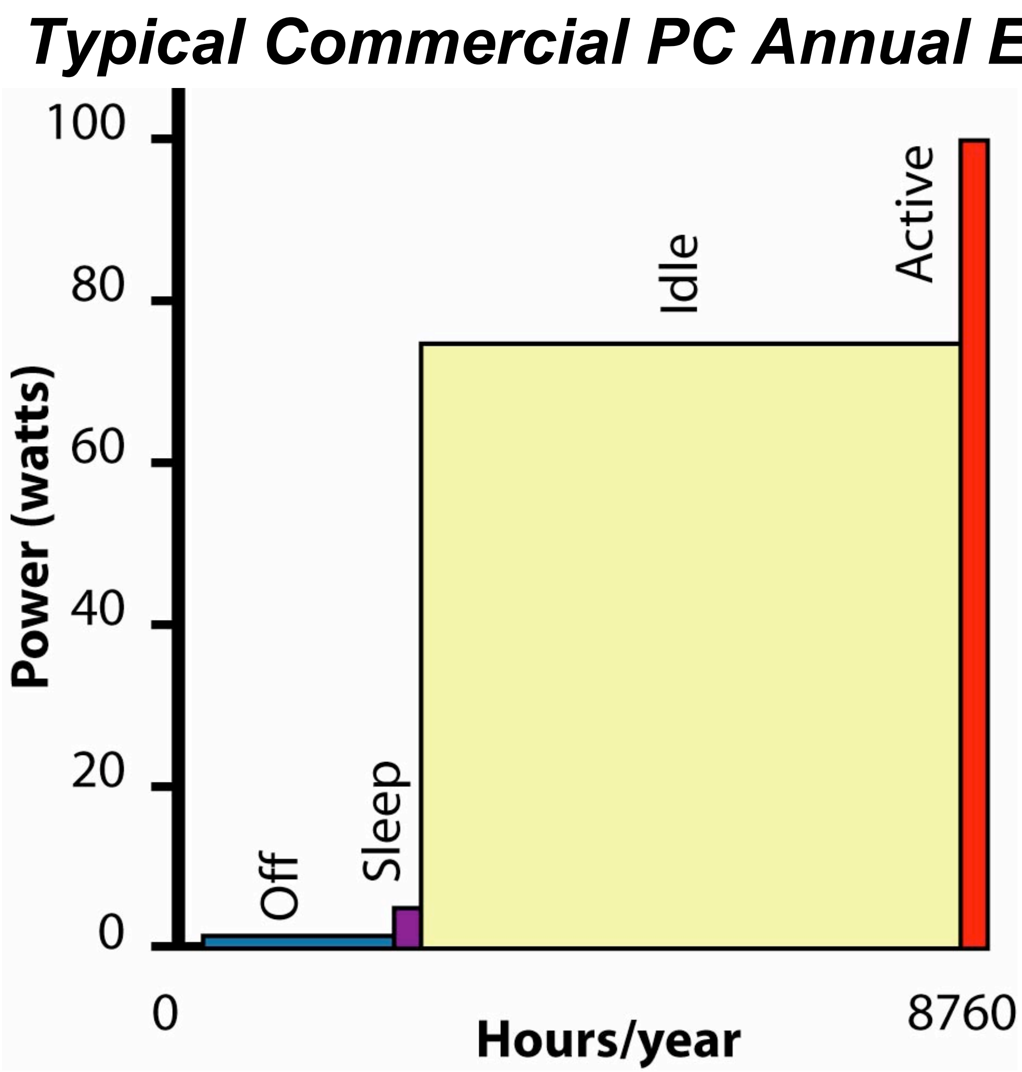
→ **National energy use**

- **All factors vary with**
  - Residential vs. commercial
  - Now vs. future
  - Desktop vs. notebook

BERKELEY LAB

USF UNIVERSITY OF SOUTH FLORIDA

PC energy use: Structure



$P_{on} >> P_{sleep}$

$P_{sleep} \cong P_{off}$

Consumption is driven by on-times, not by usage

PC energy use: Numbers

Power levels

- 70 W in On (notebooks 20); 5 W in Sleep; 2 W in Off

Usage

	Portion of Stock "Continuous On"	% Sleeping
Commercial	About 2/3 (2003)	6%
Residential	~20% (2001) and rising*	~10% ?

- Most home PCs in homes with >1 PC
- Home broadband penetration rising (~50%)

➔ > 50% on 24/7

Stock

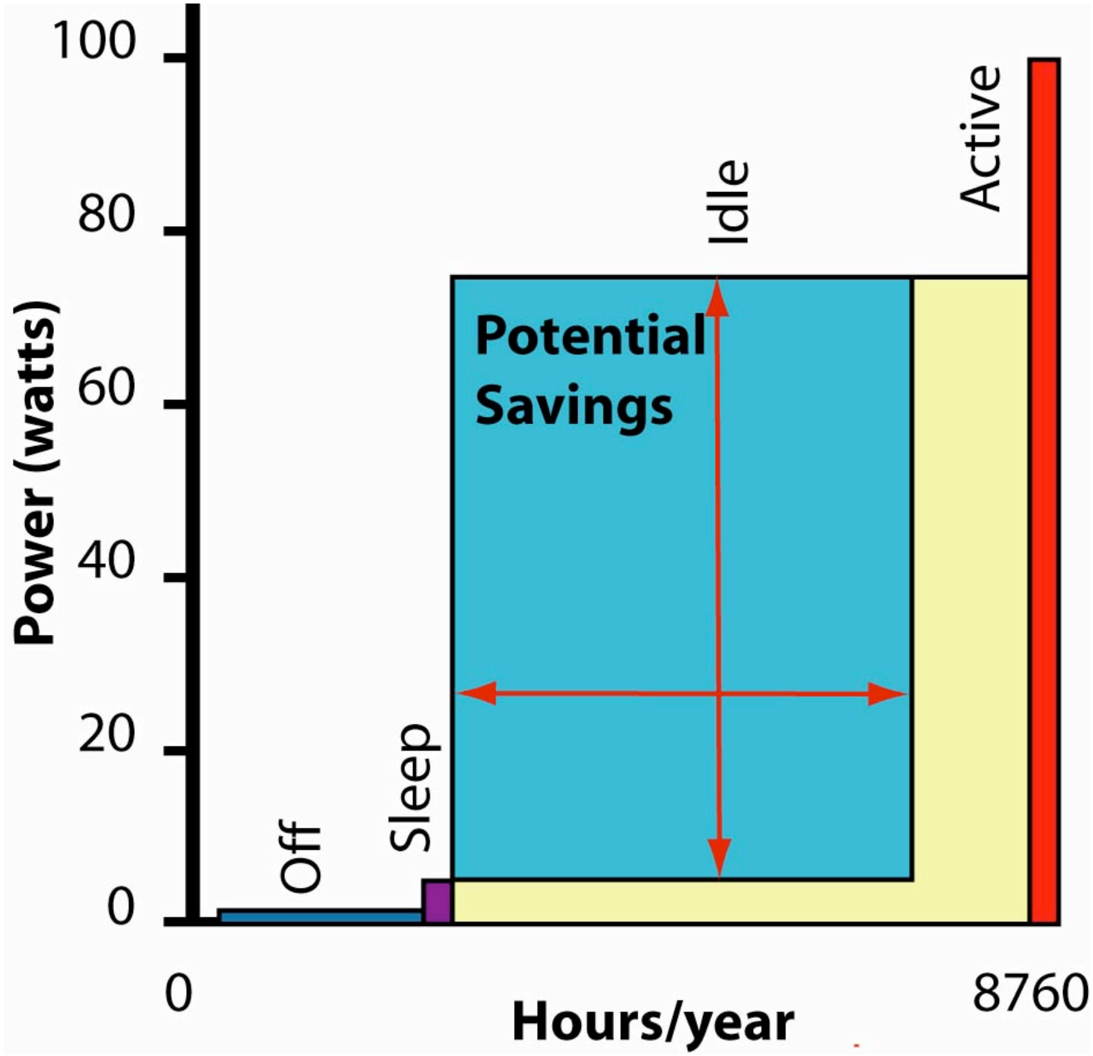
- Roughly 100 million each residential and commercial

➔ 46 TWh/year

* Half of these on 40-167 hours/week

BERKELEY LAB USF UNIVERSITY OF SOUTH FLORIDA

PC energy use: “Waste” / Savings opportunity



Most of time when idle, could be asleep;
PC savings potential is most of current consumption

EPA Energy Star program

- **1992 — Began with PC and monitor power mgmt.**
  - Capability to PM; sleep/off levels
- **1999 — Reduced power levels; addressed network connectivity**
- **Current specification revision process**
  - Power supply efficiency
  - Limits on system “idle” power
  - Network connectivity in Sleep
- **Could play a key role in reducing energy use from networks**

Network equipment energy use

❖ **At SIGCOMM 2003…**

Greening of the Internet

Maruti Gupta
Department of Computer Science
Portland State University
Portland, OR 97207
mgupta@cs.pdx.edu

Suresh Singh
Department of Computer Science
Portland State University
Portland, OR 97207
singh@cs.pdx.edu

ABSTRACT

In this paper we examine the somewhat controversial subject of energy consumption of networking devices in the Internet, motivated by data collected by the U.S. Department of Commerce. We discuss the impact on network protocols of saving energy by putting network interfaces and other router & switch components to sleep. Using sample packet

Device	*Approximate Number Deployed*	*Total AEC* TW-h
Hubs	93.5 Million	1.6 TW-h
LAN Switch	95,000	3.2 TW-h
WAN Switch	50,000	0.15 TW-h
Router	3,257	1.1 TW-h
Total		6.05 TW-h

pp. 19-26

Network equipment energy use continued

- **Switches, Hubs, Routers** (commercial sector only)
  - 6.05 TWh/year — 2000 [Singh] ➔ ~$500 million/year
- **Telecom equipment** (mobile, local, long distance, PBX)
  - 6.1 TWh/year — 2000 [Roth] ➔ ~$500 million/year
- **NICs alone — Quick Estimate**
  - 300 million products with NICs; NIC at both ends
  - 1 W per NIC; Continuous use
  - ➔ 600 MW NIC power; ➔ 5.3 TWh/year

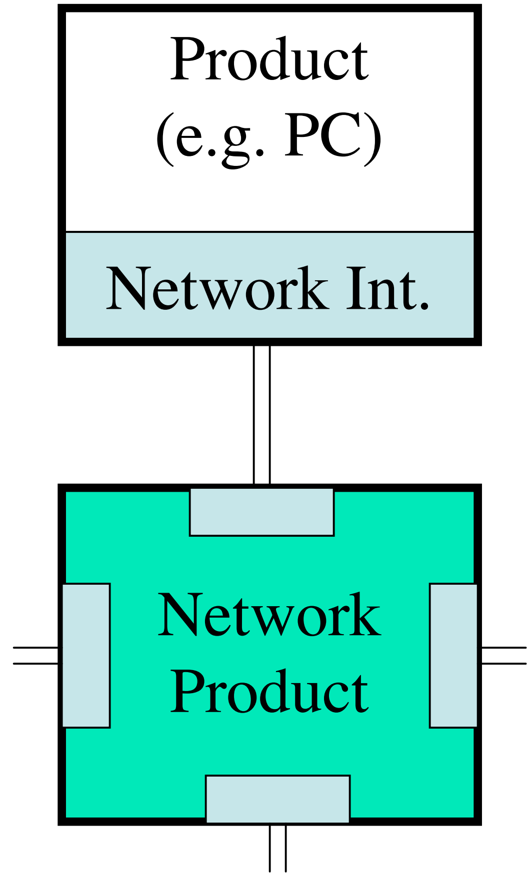
➔ > $400 million/year

BERKELEY LAB

USF UNIVERSITY OF SOUTH FLORIDA

Network direct and induced energy use

- **Network Direct**
  - NICs
  - Network Products
    - Switches, Routers, Broadband Modems, Wireless Access Points, …
- **Network Induced**
  - Increment for higher power state of devices needed to maintain network connectivity (usually On instead of Sleep or Off)
  - Common causes:
    - Can't maintain needed connectivity
    - Too cumbersome to set up or use



IT from an energy perspective

- **IT in general, and PCs in particular**
  - Consume a lot of power
  - Consumption is increasing
  - Many inefficiencies that can be removed (savings opportunities)
  - Networks increase consumption — direct and induced
- **Energy for "traditional" uses is declining**
  - Heating, cooling, lighting, appliances
- **Electronics and Miscellaneous are rising**
  - Absolute and % of total
  - Only now getting attention from energy community

Needs attention from the networking community!

Overview of power management

- Welcome to Part #2

In this part… an overview of power management, wake on LAN, and current technology directions.

Power and energy

- **Some quick definitions…**
  - Power is *W* = *V* x *A*
    - For DC this is correct, for AC we have a power factor
  - Energy is *Wh* = *Power* x *Time*
- **Consumed energy produces useful work and heat**
  - Silicon has an operational heat limit – too hot and it fails
  - Generated heat must be removed via cooling
    - Cooling is needed within the PC and also within the room
- **For mobile devices, energy use is a critical constraint**
  - Battery lifetime is limited

Power and energy continued

❖ **In a clocked CMOS chip…**

- Power is (to a first order) ACV^2f
  - ***A*** is activity factor and ***C*** is capacitance
  - Power is proportional to the square of voltage
- V is linear with f
  - We can scale frequency (and voltage) to reduce power
  - Power (***P***) is thus proportional to the cube of frequency

$$\boxed{P = P_{fixed} + c*f^3}$$

Where P_{fixed} is the fixed power (not frequency dependent) and ***c*** is a constant (which comes from ***A*** and ***C*** above)

BERKELEY LAB

USF UNIVERSITY OF SOUTH FLORIDA

Power and performance

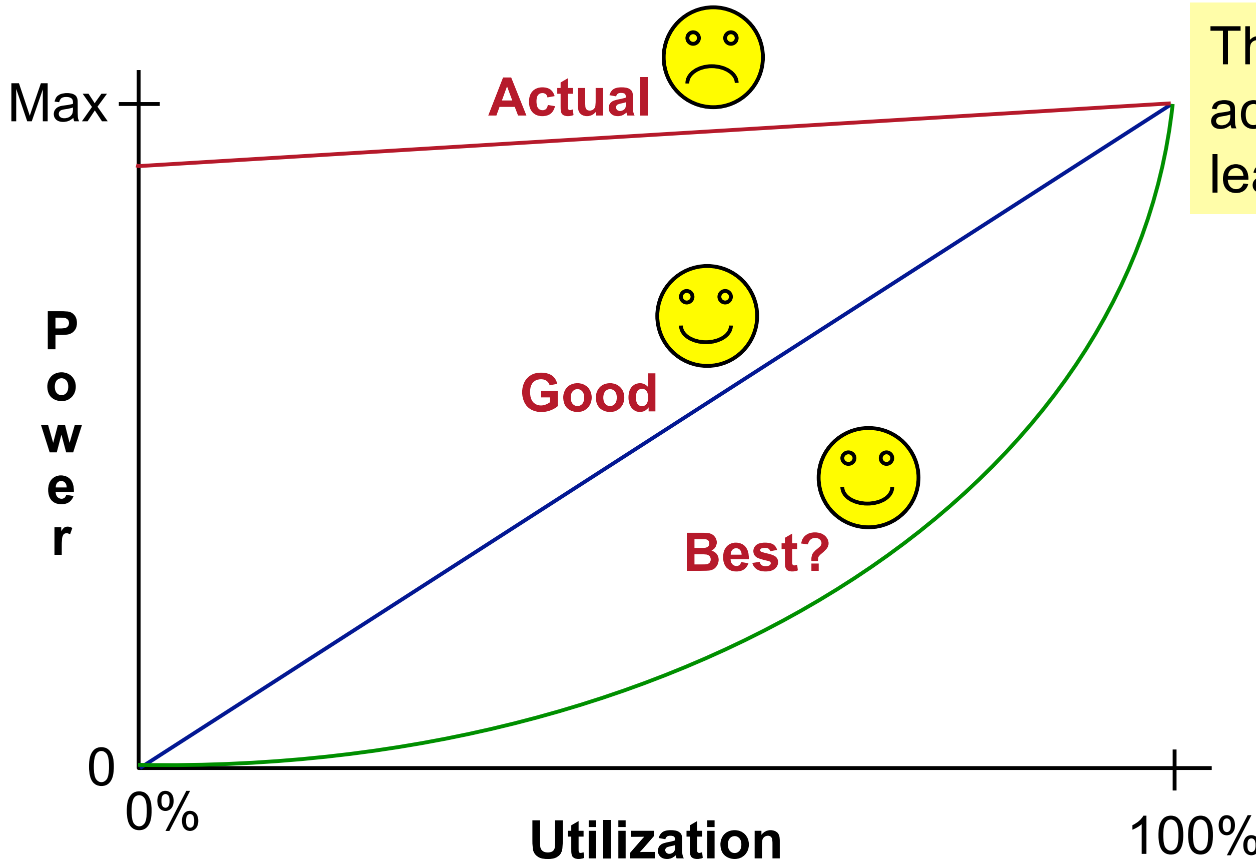
- **Key performance metrics for IT services…**
  - *Response time* for a request
  - *Throughput* of jobs

  Mean and 99 percentile
- **We have a trade-off…**
  - Reducing power use may increase response time
  - Trade-off is in energy used versus performance

A response time faster than "fast enough" is wasteful

Power and utilization

- **Power use should be proportional to utilization**
  - But it rarely is!



The goal is to achieve at least linear

Basic principles of power management

- **To save energy we can:**
  - Use more efficient chips and components
  - Better power manage components and systems

- **To power manage we have three methods:**

Do less work (processing, transmission)
- Transmitting is very expensive in wireless

Slow down
- Process no faster than needed (be deadline driven)

Turn-off "stuff" not being used
- Within a chip (e.g., floating point unit)
- Within a component (e.g., disk drive)
- Within a system (e.g., server in a cluster)

BERKELEY LAB

USF
UNIVERSITY OF SOUTH FLORIDA

Basic principles of power management <u>continued</u>

Time scales of idle periods

- Nanoseconds – processor instructions
- Microseconds – interpacket
- Milliseconds – interpacket and interburst
- Seconds – flows (e.g., TCP connections)
- Hours – system use

Basic principles of power management continued

❖ **The key challenges for power management are:**

- Predicting, controlling, and making the best use of idle times
- Increasing the predictability of idle times
- Creating added idle time by bunching and/or eliminating processing and transmission

Power management in PCs

- **PCs support power management**
  - For conserving batteries in mobile systems
  - For energy conservation (EPA Energy Star compliance)

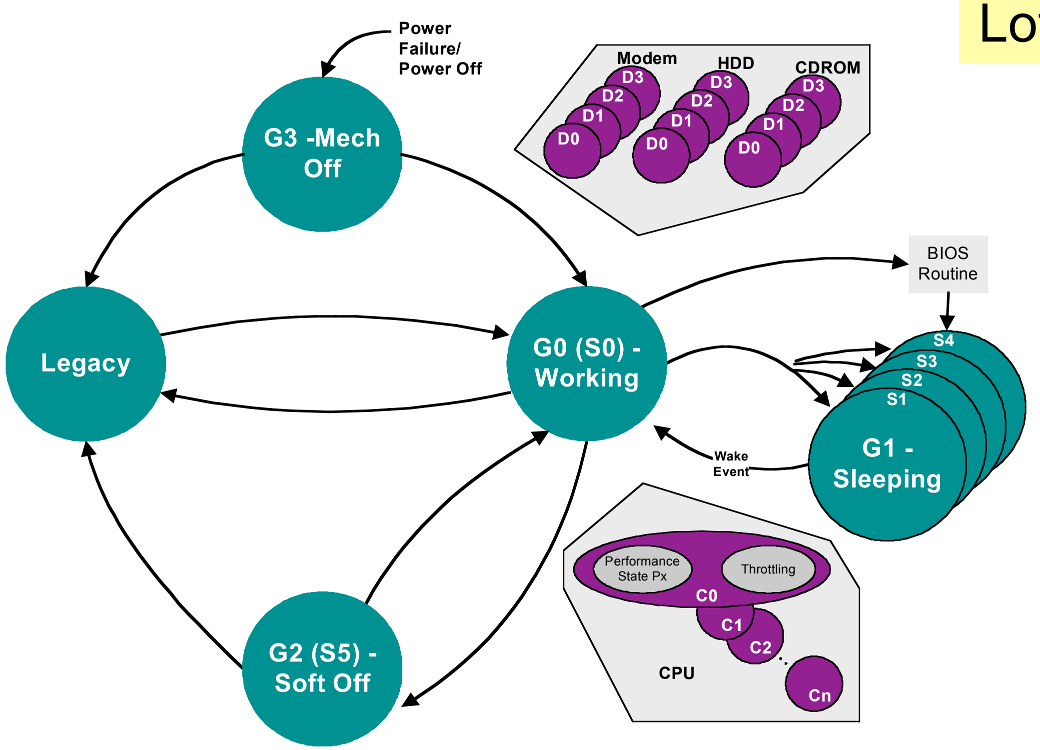
- **How it works …**
  - Use an inactivity timer to power down
  - Power down monitor, disks, and eventually the entire system
    - Sleep (Windows *Standby)* and *Hibernate*
  - Resume where left-off on detection of activity
    - Mouse wiggle or key stroke to wake-up

Power management in PCs continued

❖ **Advanced Configuration and Power Interface (ACPI)**

- ACPI interface is built-in to operating systems
  - An application can "veto" any power down

Lots of states!



* From page 27 of ACPI Specification (Rev 3.0, September 2, 2004)

Power management in PCs continued

- **Wake events**
  - User mouse wiggle or keystroke
  - Real time clock alarm
  - Modem “wake on ring”
  - LAN “wake on LAN” (WOL)
  - LAN packet pattern match

Time to wake-up is less of an issue than it used to be

BERKELEY LAB

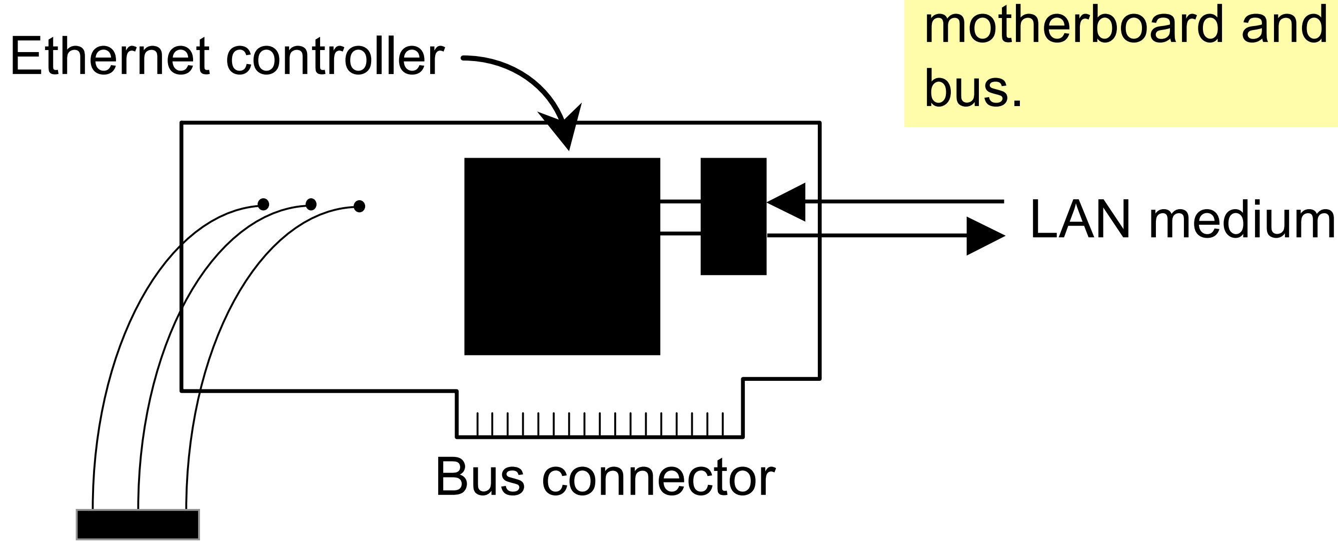
USF UNIVERSITY OF SOUTH FLORIDA

Wake on LAN

- **Wake on LAN (WOL)**
  - A special MAC frame that a NIC recognizes

    (MAC address repeated 16 times in data field)
    - Developed in mid 1990's
    - Called Magic Packet (by AMD)
    - Intended or remote administration of PCs

All this is now on the motherboard and PCI bus.



Cable and connector for auxiliary power and wake-up interrupt lines

Wake on LAN <u>continued</u>

❖ **WOL has shortcomings…**

- ✖ Must know the MAC address of remote PC
- ✖ Cannot route to remote PC due to last hop router timing-out and discarding ARP cache entry
- ✖ Existing applications and protocols do not support WOL
  - For example, TCP connection starts with a SYN

WOL implemented in most Ethernet and some WiFi NICs

Directed packet wake-up

- **A better WOL**
  - Wake on interesting packets and pattern matching*

4.3.2.1 “Interesting” Packet Event

In the power-down state, the 82559 is capable of recognizing “interesting” packets. The 82559 supports pre-defined and programmable packets that can be defined as any of the following:

- ARP Packets (with Multiple IP addresses)
- Direct Packets (with or without type qualification)
- Magic Packet*
- Neighbor Discovery Multicast Address Packet (‘ARP’ in IPv6 environment)
- NetBIOS over TCP/IP (NBT) Query Packet (under IPv4)
- Internetwork Package Exchange* (IPX) Diagnostic Packet
- TCO Packet

This allows the 82559 to handle various packet types. In general, the 82559 supports programmable filtering of any packet in the first 128 bytes.

Datasheet 31

* From page 31 of Intel 82559 Fast Ethernet Controller datasheet (Rev 2.4)

Directed packet wake-up <u>continued</u>

❖ Directed packet wake-up has shortcomings…

- **Wake-up on unnecessary or trivial requests**
  - "Wake on Junk"
- **Not wake-up when need to**
- **Needs to be configured**

A pattern match is "unintelligent" — no concept of state

BERKELEY LAB

USF UNIVERSITY OF SOUTH FLORIDA

Current research and development

- **There are current efforts to reduce energy use in …**
  - Power distribution
  - Processors
  - Wireless LANs
  - Supercomputers
  - Data centers
  - Corporate PCs (central control)
  - Displays
  - **LAN switches**
  - **NICs**
  - **Universal Plug and Play (UPnP) protocols**
  - **ADSL2**

Reducing energy in LAN switches

- **Over 6 TWh/year used by LAN switches and routers**

About $500 million/year

  - Turning switch core off during interpacket times
    - Keep buffers powered-up to not lose packets
    - Prediction (of idle period) triggers power-down
    - Arriving packets into buffer trigger wake-up
  - NSF funded work at Portland State University (Singh et al.)

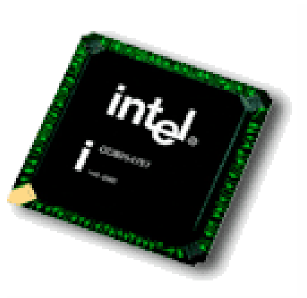
Interesting idea, more work needs to be done

BERKELEY LAB

USF UNIVERSITY OF SOUTH FLORIDA

Reducing energy in NICs

❖ **NICs are implemented with multiple power states**
- D0, D1, D2, and D3 per ACPI



Typical notebook NIC

Intel 82541PI Gigabit Ethernet Controller*
- 1 W at 1 Gb/sec operation
- Smart power down
  - Turns-off PHY if no signal on link
- Power save mode
  - Drops link rate to 10 Mb/sec if PC on battery

* From Intel 82641PI product information web site (2005)

Reducing energy in UPnP

❖ UPnP may become widespread in homes

- UPnP uses distributed discovery (SSDP)
  - Every device must periodically send and receive packets
- UPnP Forum developing a standard for a proxy
  - Single proxy per UPnP network
  - Proxy sends and receives on behalf of sleeping devices
  - Due out in summer 2006
- Developed and tested a similar UPnP proxy at USF
  - Available at *http://www.csee.usf.edu/~christen/upnp/main.html*

The UPnP proxy is protocol specific

BERKELEY LAB

USF UNIVERSITY OF SOUTH FLORIDA

Reducing energy in ADSL2

- **ADSL2 is a last mile "to the home" technology**
  - 30 million DSL subscribers worldwide
- **ADSL2 is G.992.3, G.922.4, and G.992.5 from ITU**
  - Standardized in 2002
- **ADSL2 supports power management capabilities**
  - Link states L0 = full link data rate
  - Link state L2 = reduced link data rate
  - Link state L3 = link is off

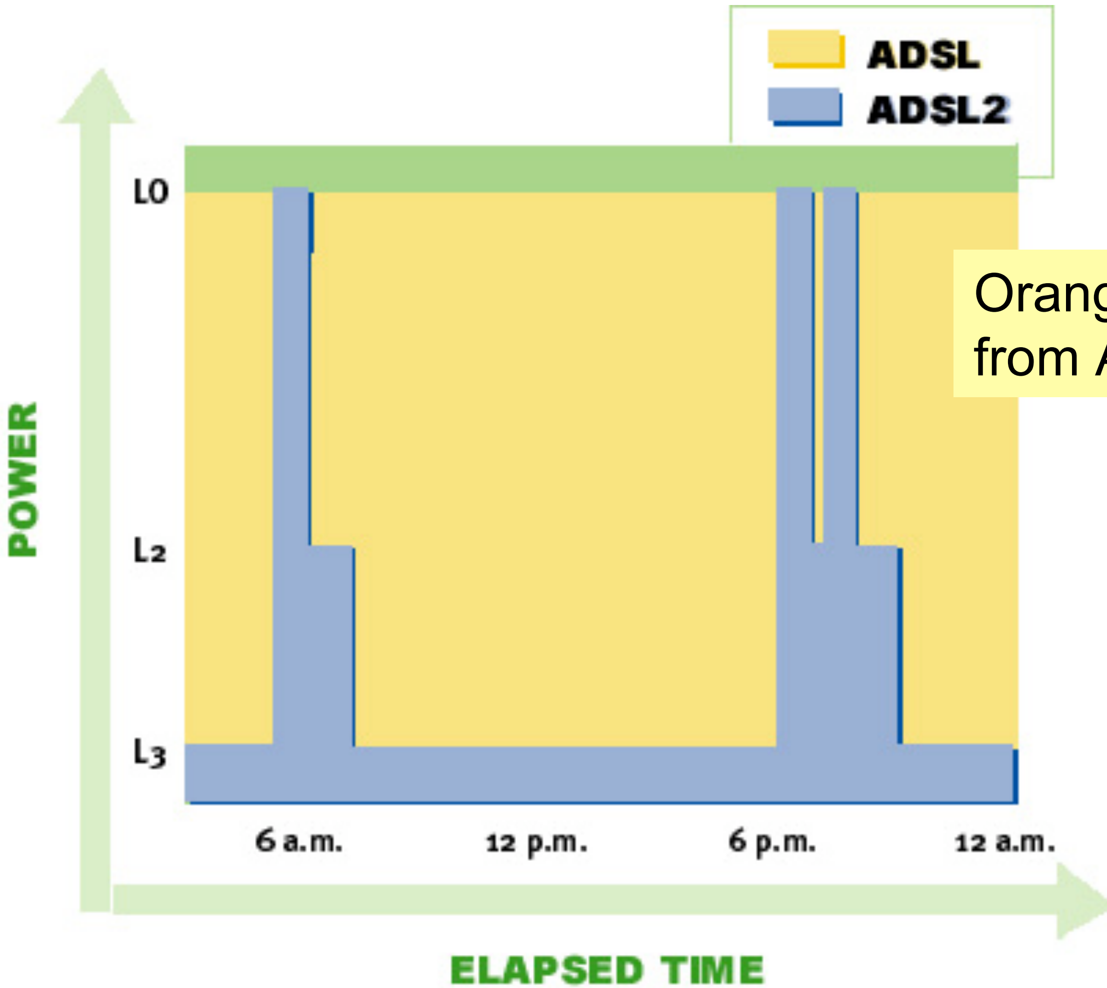
Symbol based handshake

How might this apply to Ethernet?

Reducing energy in ADSL2 continued

❖ **ADSL2 energy savings…**

This is utilization based control



Orange region is savings from ADSL2 versus ADSL

* From M. Tzannes, "ADSL2 Helps Slash Power in Broadband Designs," *CommDesign.com*, January 30, 2003.

Reducing network-induced energy use

- **Welcome to Part #3**

In this part… the "sleep-friendly" PC – its motivation, requirements, design, and next steps.

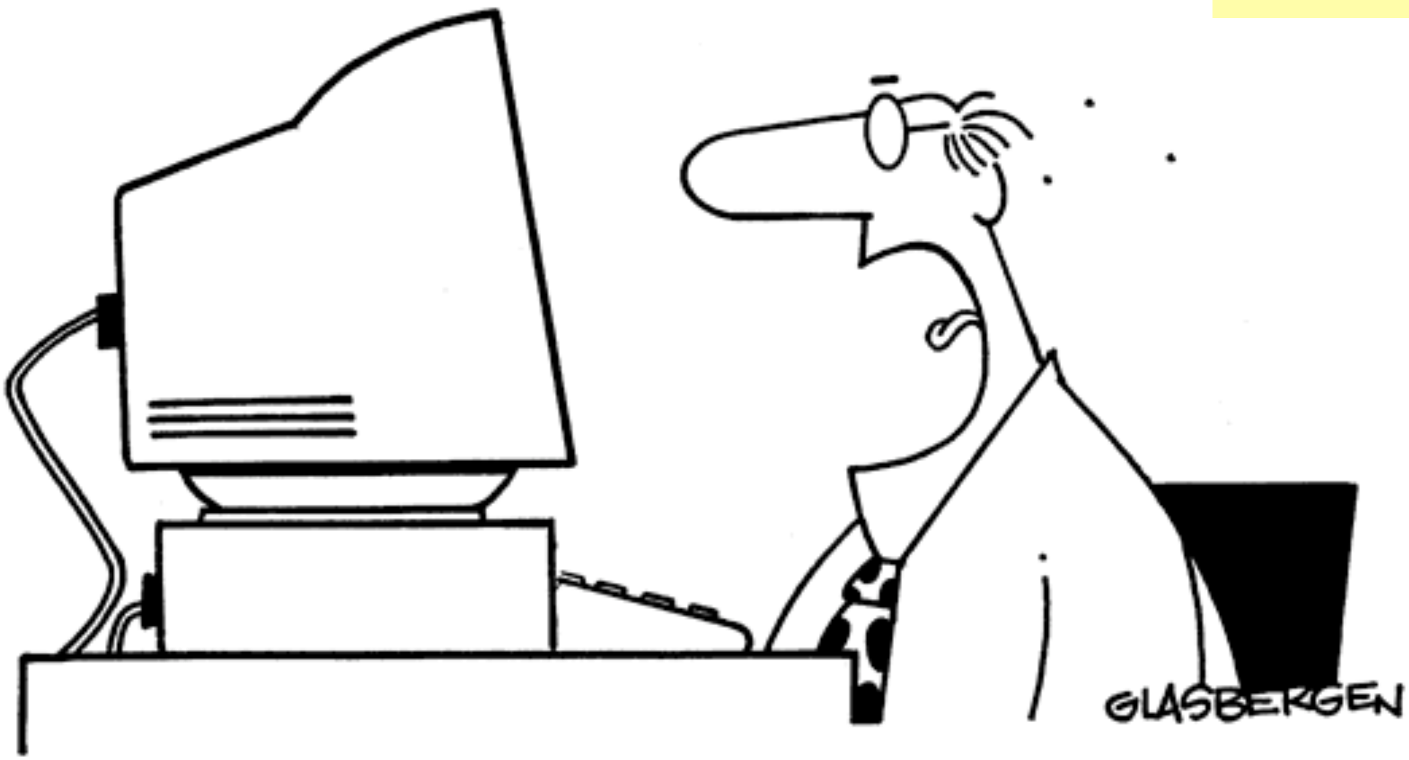
➡ Goal is to reduce network *induced* energy use

Disabling of power management

- **Why is power management disabled in most PCs?**
- **Why are many PCs fully powered-on "all the time"?**
  - Historically this was for reasons of poor performance
    - Crash on power-up, excess delay on power-up, etc.
  - Today increasing for network-related reasons

Increasing number of applications are network-centric

This is not a cartoon



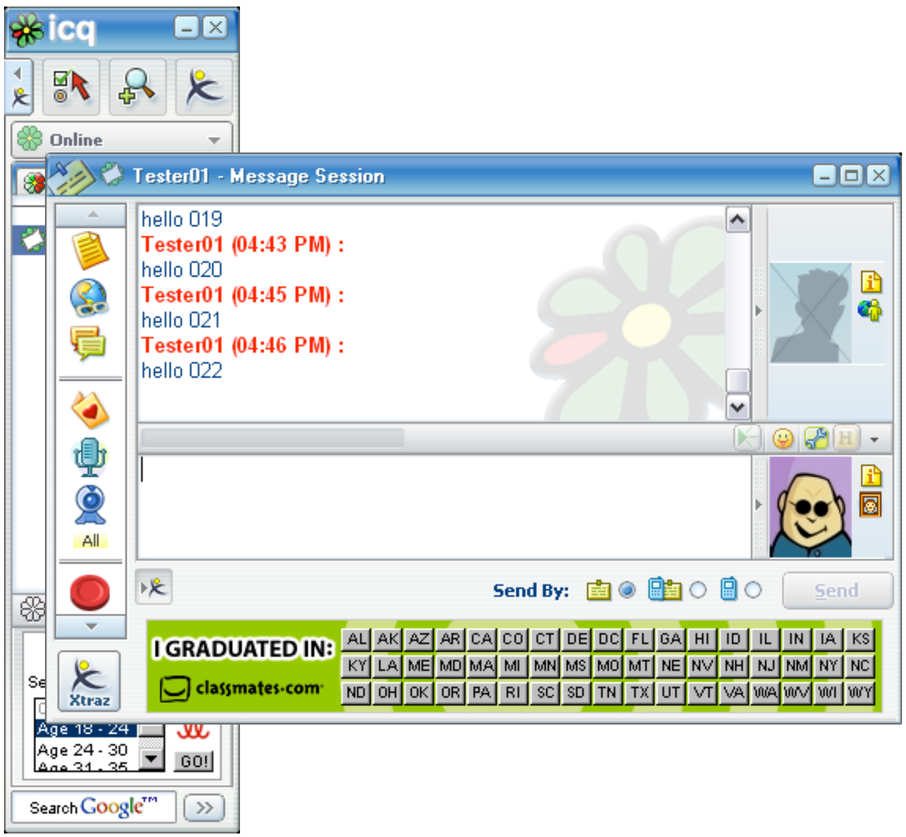
"I don't have to wait for my television to boot up! I don't have to wait for my stereo to boot up! I don't have to wait for my microwave to boot up!"

Disabling for protocols

- **Some protocols require a PC to be fully powered-up**
- **Some examples…**
  - ARP packets – *must respond*
    - If no response then a PC becomes "unreachable"
  - TCP SYN packets – *must respond*
    - If no response then an application is "unreachable"
  - IGMP query packets – *must respond*
    - If no response then multicast to a PC is lost
  - DHCP lease request – *must generate*
    - If no lease request then a PC will lose its IP address

Connections are everywhere

- **Permanent connections are becoming common**
  - At TCP level – “keep alive” messages are exchanged
  - At app. level – app. “status” messages are exchanged
    - Must respond at either level or connection can be dropped



Dropped connection returns user to log-in screen (and messages lost!)

PC goes to sleep

Disabling for applications

- **Some applications require a PC to be fully powered-up**
  - Permanent TCP connections are common
- **Some examples…**
  - Remote access for maintenance
  - Remote access for GoToMyPC or Remote Desktop
  - File access on a remote network drive
  - P2P file sharing
  - Some VPN
  - Some IM and chat applications
- **Some applications disable sleep**
  - No way to know power status of a remote PC
  - No way to guarantee wake-up of a remote PC

A traffic study

- **We traced packets arriving to an idle PC at USF (2005)**
  - Received 296,387 packets in 12 hours and 40 minutes

This is 6 pkts/sec

Protocol	% in trace
ARP	52.5 %
UPnP	16.5
Bridge Hello	7.8
Cisco Discovery	6.9
NetBIOS Datagram	4.4
NetBIOS Name Service	3.6
Banyan System	1.8
OSPF	1.6
DHCP	1.2
IP Multicast	1.0

Remaining 2.7% and less than 1% each we found RIP, SMB, BOOTP, NTP, ICMP, DEC, X display, and many others

BERKELEY LAB

USF
UNIVERSITY OF SOUTH FLORIDA

Another reason for disabling power management?

A traffic study continued

- **Four categories of packets were identified:**

1) Ignore
   - Packets intended for other computers

2) Require a simple response
   - e.g., ARP and ICMP ping

3) Require a simple response and a state update
   - e.g., some NetBIOS datagrams

4) Require a response and application activity
   - e.g., TCP SYN

Majority

Wake event

- **Fifth category would be**
  - "originated by protocol or application" (e.g., DHCP lease)

A sleep-friendly PC

What capabilities would a sleep-friendly PC need?

- No changes to existing protocols
  - Only minimal changes to applications
- No change in user experience
- Maintain network presence with little or no wake-up of PC
- Generate routine packets as needed
- Reliably and robustly wake-up PC when needed
- Not wake-up PC when not needed
- Provide for exposing power state to network

A sleep-friendly PC continued

- **Key capabilities**

1) Ignore
   - Ignore and discard packets that require no action

2) Proxy
   - Respond to trivial requests without need to wake-up PC

3) Wake-up
   - Wake-up PC for valid, non-trivial requests

4) Handle TCP connections
   - Prevent permanent TCP connections from being dropped

Proxying

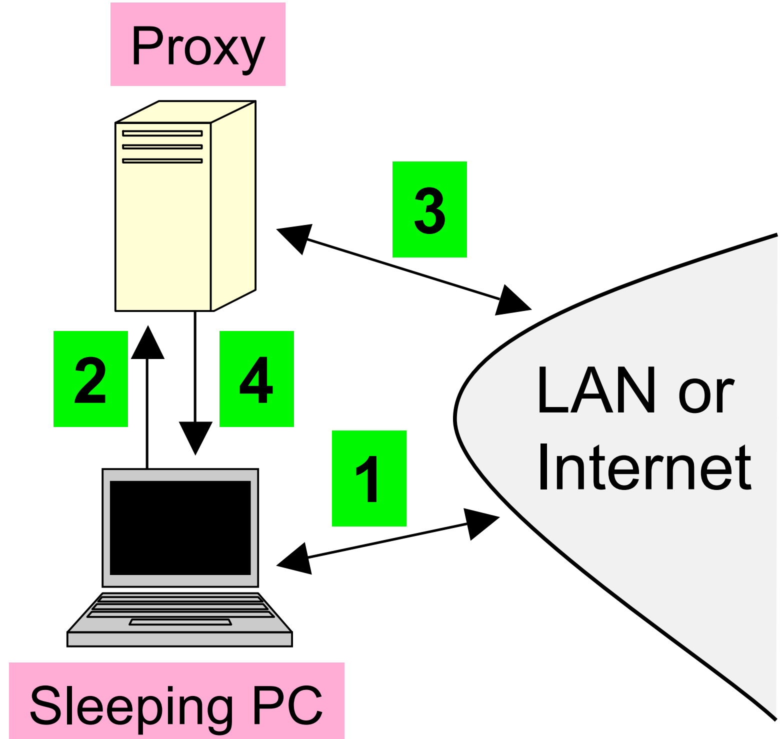
❖ **Flow for proxying…**

1 **PC awake; becomes idle**

2 **PC transfers network presence to proxy on going to sleep**

3 **Proxy responds to routine network traffic for sleeping PC**

4 **Proxy wakes up PC as needed**



Proxy can be <u>internal</u> (NIC) or <u>external</u> (in other PC, switch or router, wireless base station, or dedicated device)

Wake-up

- **Is a better wake-up needed?**
- **We may need:**
  - A more stateful (or intelligent) wake-up decision
  - Wake-up as an application semantic
    - Applications may have standard wake-up templates
    - Current wake-up packet pattern is established by the OS

BERKELEY LAB

USF UNIVERSITY OF SOUTH FLORIDA

Handling TCP connections

- **How to handle permanent TCP connections?**
- **We may need:**
  - TCP connections that are “split” within a PC
    - NIC can answer for keep-alive while PC is sleeping
  - Wake-up for TCP keep-alive messages
  - Applications to not use permanent TCP connections
    - Possibly could only connect when actively sending/receiving data

Energy aware applications

Should it be "Green application" in addition to "Green PC"?

- **Can applications increase the enabling of power management?**
- **We may need:**
  - Applications that maintain state to drop TCP connections
  - Applications that are power aware in entirely new ways

Options for a Sleep Friendly PC

- **Four possible options…**

1) ~~Selective wake-up NICs~~
    - ~~Such as WOL or direct packet wake-up~~

2) Proxy internal to a NIC
    - We call this a SmartNIC (and includes wake-up)

3) Central proxy in a switch, access point, etc.
    - Build on UPnP proxy idea

4) Very low power fully-operational mode of PC
    - OS and processor active, but operate slowly

SmartNIC is most promising, (3) and (4) can have a role

BERKELEY LAB

USF UNIVERSITY OF SOUTH FLORIDA

SmartNIC concept

Can we add capability to a NIC such that a PC can remain in a low-power sleep state more than it can today?

- **A SmartNIC contains**
  - Proxy capability (*new*)
  - Wake-up capability (*as today and improved*)
  - Ability to advertise power state (*new*)

- **When a PC is powered-down the SmartNIC…**
  - Remains powered-up
  - “Covers” or “proxies” for the PC
  - Wakes-up the PC only when needed
  - Communicates power state as needed

BERKELEY LAB

USF UNIVERSITY OF SOUTH FLORIDA

SmartNIC requirements

❖ **Need to better understand what is needed**

- Categorize network traffic
  - No response needed
  - Trivial response needed
  - Non-trivial response needed
  - Routine packet generation

How much time to respond?
When can we lose "first one"?

- Understand application and OS state changes
  - Incoming packets that cause a state change
  - Outgoing packets that cause a state change
- Understand likely needs of future devices and applications
  - Wireless, mobile, etc.
- Assess security implications

BERKELEY LAB

USF UNIVERSITY OF SOUTH FLORIDA

SmartNIC requirements continued

❖ **SmartNIC must be able to…**

- Have some knowledge of protocol state
  - For example, DHCP leasing
- Have some knowledge of application state
  - For example, listening TCP ports
- Receive, store, process, and send packets
  - Execute some subset of the IP protocol stack

Also appeals to "green" consumers

Adding a few dollars cost to the NIC may save many tens of dollars of electricity costs per PC per year.

Reducing network direct energy use

- Welcome to Part #4

In this part… a discussion of how to reduce direct energy use with adaptive link rate.

➡ Goal is to reduce network *direct* energy use

Power management of a link

Can we power manage an Ethernet link and NICs?

- **Can we trade-off performance and energy?**
  - High data rate = high performance (low delay)
  - Low data rate = low performance (high delay)
- **If idle or low utilization, <u>do not need high data rate</u>**
  - Can we switch link data rate?
  - How fast can we switch link data rates?
  - What policies do we use to switch data rates?

Low utilization periods

- ***Low utilization*** **is time periods with "few" packets**

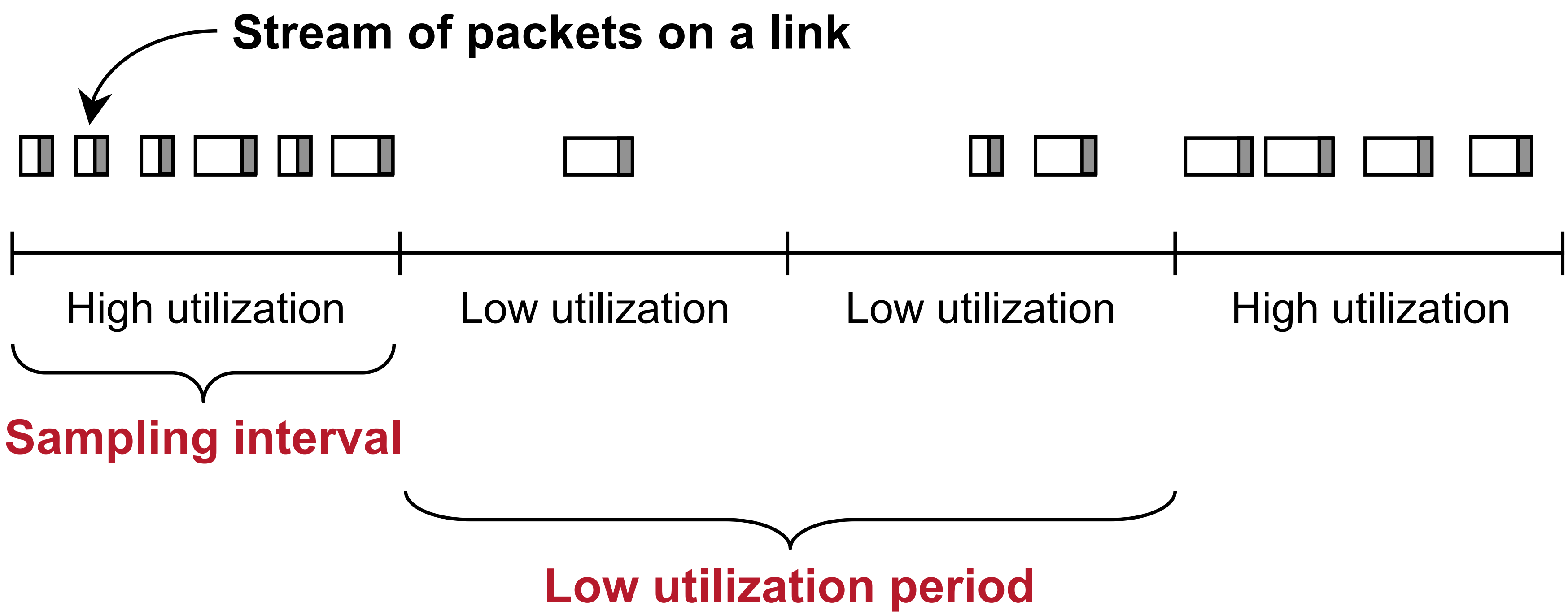
- **We measure low utilization as**
  - Less than 5% utilization (in bits/sec) in a 1 millisec sample

Low utilization period **= count of successive low samples**

Possibly can partially power down for idle periods and switch link to lower data rate for low utilization periods.

Low utilization periods continued

- **Low utilization in a stream of packets**
  - Packets are variable in length (64 to 1500 bytes)



Power measurements

How much power use is direct from the network?

- **We study power consumption due to Ethernet links**
- **We measure…**
  - Cisco Catalyst 2970 LAN switch
  - Intel Pro 1000/MT NIC
- **We study the specifications for…**
  - Intel 82547GI/82547EI Gigabit Ethernet Controller (NIC)
  - Chelsio N210 10GbE Server Adapter (NIC)

Power measurements continued

- **Power use measurement***
  - Catalyst 2970 24-port LAN switch

Active configured links

Measured at wall socket (AC)

# ports	10 Mb/sec	100 Mb/sec	1000 Mb/sec
0	69.1 W	69.1 W	69.1 W
2	70.2	70.1	72.9
4	71.1	70.0	76.7
6	71.6	71.1	80.2
8	71.9	71.9	83.7

10 and 100 Mb/sec are about the same

At 1000 Mb/sec it is about 1.8 W added per active link

* By Chamara Gunaratne from University of South Florida (August 2004)

Power measurements continued

- **Power use measurements***
  - For Intel Pro 1000/MT NIC

Measured at PCI bus (DC)

Idle Link (*no activity*)			
Rate (Mb/s)	Current (mA)	Voltage (V)	Power (W)
1000	770	5.08	3.91
100	224	5.11	1.14
10	130	5.11	.664

Active Link (*file transfer*)			
Rate (Mb/s)	Current (mA)	Voltage (V)	Power (W)
1000	768	5.08	3.90
100	224	5.11	1.14
10	124	5.11	.633

Difference between 1000 and 10 Mb/sec is about 3.2 W

No significant difference between idle and active link

* By Brian Letzen from University of Florida (February 2005)

Power measurements continued

- **Power use specifications for 1 Gb/sec***
  - For Intel 82547GI/82547EI Gigabit Ethernet Controller

Typical PC NIC

	D0a							
	Unplugged No Link		@10 Mbps		@100Mbps		@ 1000 Mbps	
	Typ Icc (mA)[a]	Max Icc (mA)[b]	Typ Icc (mA)[a]	Max Icc (mA)[b]	Typ Icc (mA)[a]	Max Icc (mA)[b]	Typ Icc (mA)[a]	Max Icc (mA)[b]
1.8V	21	25	10	95	115	120	320	325
1.2V	55	65	145	160	160	170	440	485
Total Device Power	135		150		435		1.1W	1.2W

a. Typical conditions: operating temperature (TA) … at full duplex, and PCI 33 MHz system interfac…
b. Maximum conditions: minimum operating temp… tinuous network traffic at full duplex, and PCI 33 MHz system interface.

Difference between 1000 and 10 Mb/sec is about 1 W

* From page 15 of Intel 82547GI/82547EI datasheet (Rev 2.1, November 2004)

Power measurements continued

- **Power use specifications for 10 Gb/sec*** Server NIC
  - For Chelsio N210 10GbE Server Adapter
    - Fiber link (previous NICs were copper)

Physical and Environmental

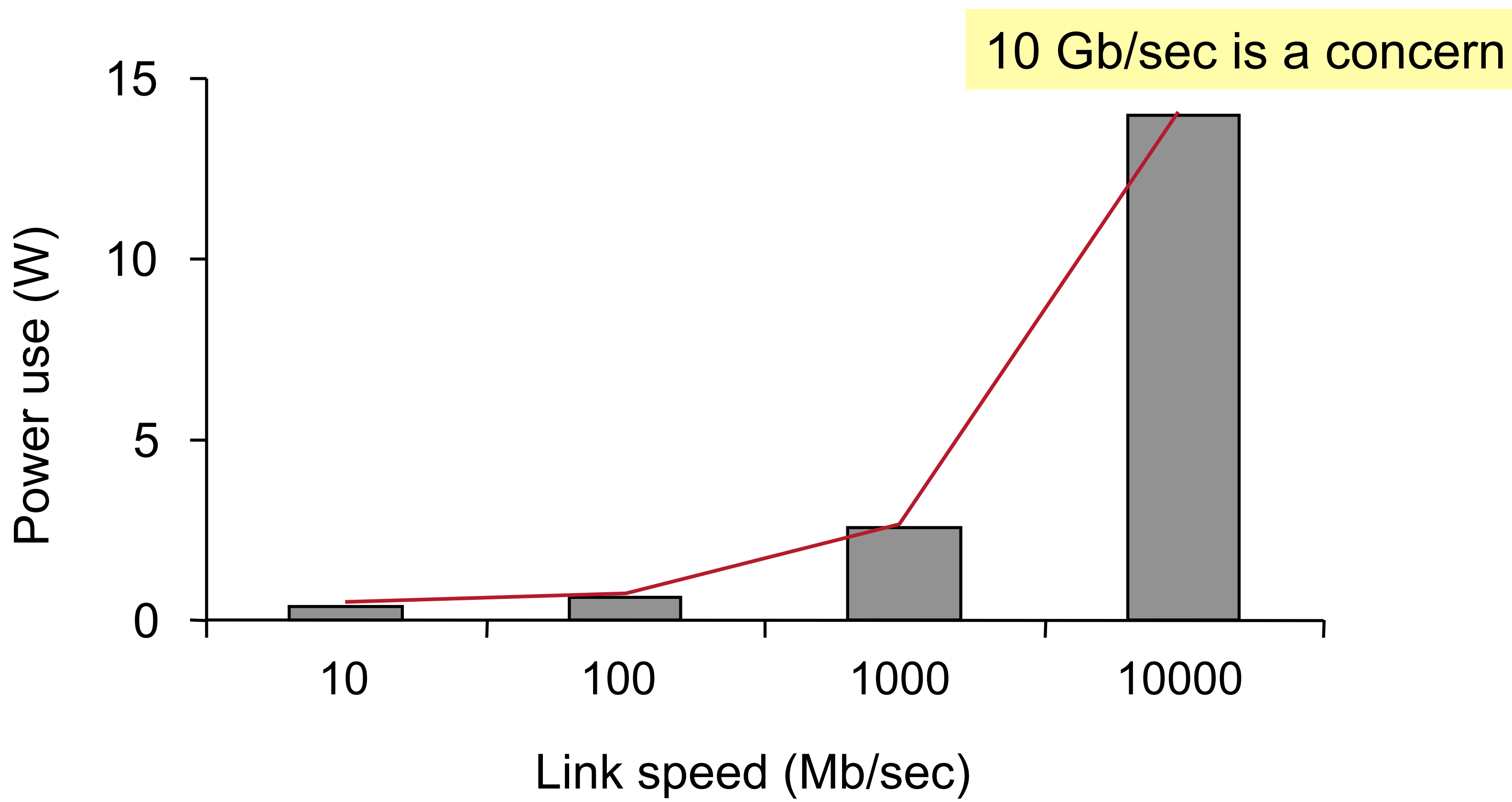
- Length: 6.6 in.
- Height: 3.8 in.
- Operating Temperature 0 to 65 deg C
- Operating Humidity: 5 to 95%
- Typical Power Consumed: 14 Watts

10 Gb/sec is 10x power consumption of 1 Gb/sec?

* From Chelsio N210 product brief (Rev 2.1, November 2004)

Power measurements continued

- **Summary of power measurements**
  - Bar graph showing averages of all measurements



g00.xls

Adaptive link rate (ALR)

- **Automatic link speed switching***
  - For 82547GI/82547EI Gigabit Ethernet Controller

Typical PC NIC

- Automatic link speed switching from 1000Mb/s down to 10 or 100Mb/s in standby
- Low power in standby states
- Supports power-down states without software assistance

Drops link speed to 10 Mb/sec when PC enters low-power state

Motivates dropping link data rate if low utilization

* From Intel 82547GI/82547EI product information (82547gi.htm)

Adaptive link rate (ALR) <u>continued</u>

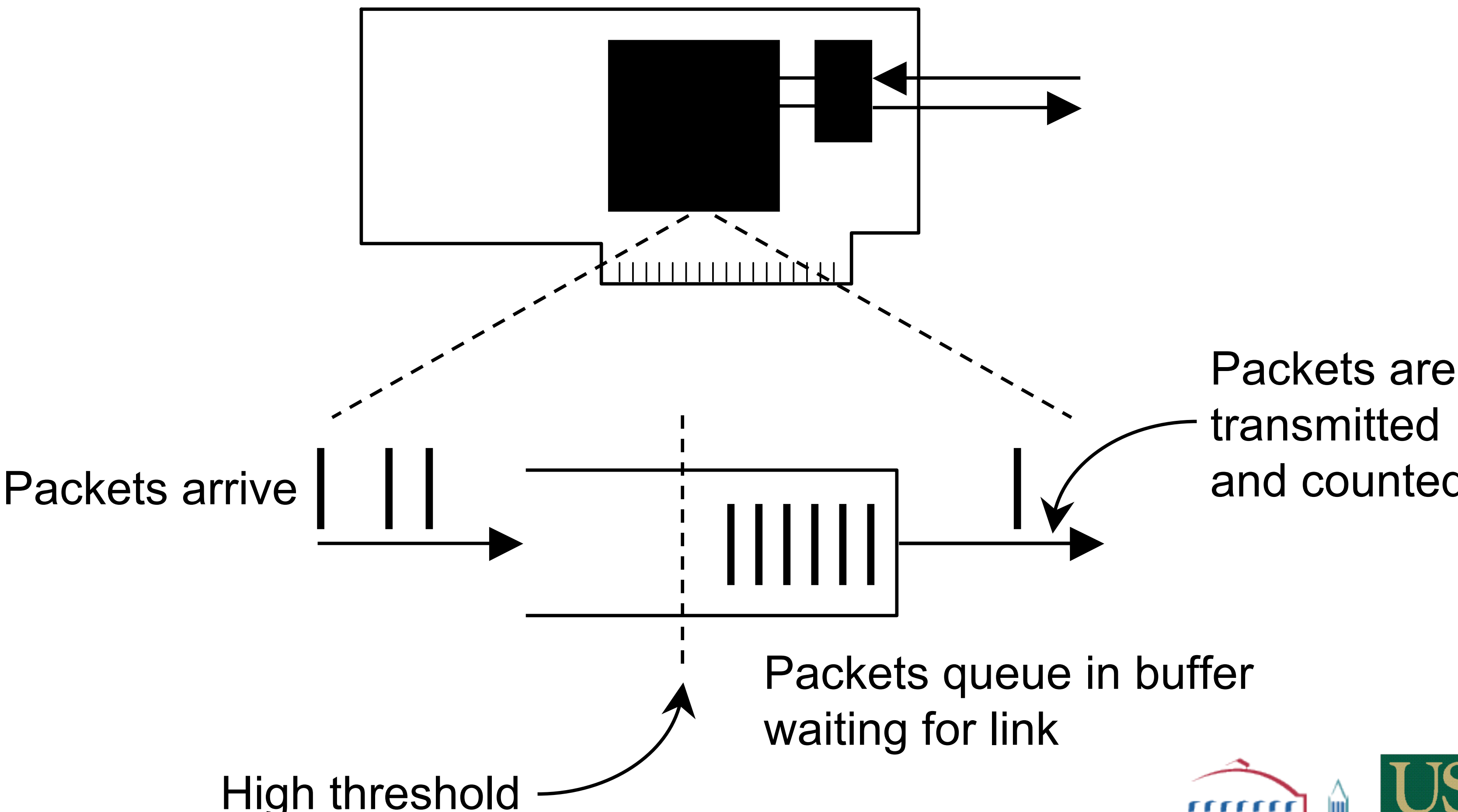
Independent of PC power management

<u>Goal:</u> Save energy by matching link data rate to utilization

- **Change (or adapt) data rate in response to utilization**
  - Use 10 or 100 Mb/sec during low utilization periods
  - Use 1 or 10 Gb/sec during high utilization periods
- **Need new *mechanism***
  - Current auto-negotiation is not suitable (too slow)
    - Designed for set-up (e.g., boot-up time), not routine use
- **Need *policies* for use of mechanism**
  - *Reactive policy* possible if can switch link rates "quickly"
  - *Predictive policy* is needed otherwise

Policies for ALR

- **Can use queue length and utilization (*reactive policy*)**
  - In a NIC (within PC or a LAN switch)



Policies for ALR continued

- For *reactive policy* two new processes execute
  - Check for threshold crossing
  - Check for utilization is low

Executes on an arriving packet…

```
if (link rate is low)
  if (buffer exceeds threshold)
    wait for current packet transmission to finish
    handshake for high link rate
transmit the next queued packet
```

Executing at all times…

```
if (link rate is high)
  if (utilization is low)
    wait for current packet transmission to finish
    handshake for low link rate
transmit the next queued packet
```

Traffic characterization

How much time is there for power management?

- **We collect and characterize traffic "in the wild"**
- **We are interested in understanding…**
  - Low utilization periods
- **We are also interested in understanding…**
  - Idle periods

Traffic characterization continued

- **Traffic collection at University of South Florida (USF)**
  - Three traces from dormitory LAN (3000+ users) in mid-2004
    - USF #1 – The busiest user
    - USF #2 – 10th busiest user
    - USF #3 – Typical user

- **Traffic collection details**
  - All are 100 Mb/sec Ethernet links
  - USF traces are 30 minutes captured with Ethereal

Traffic characterization continued

❖ Summary of the traces continued

Utilization is low

Trace	Total busy time	Total idle time	Total low util time	Utilization at 100 Mb/sec
USF #1	75 s	1759 s	1415 s	4.11 %
USF #2	47	1771	1571	2.63
USF #3	0.55	1801	1799	0.03

BERKELEY LAB

USF UNIVERSITY OF SOUTH FLORIDA

Traffic characterization continued

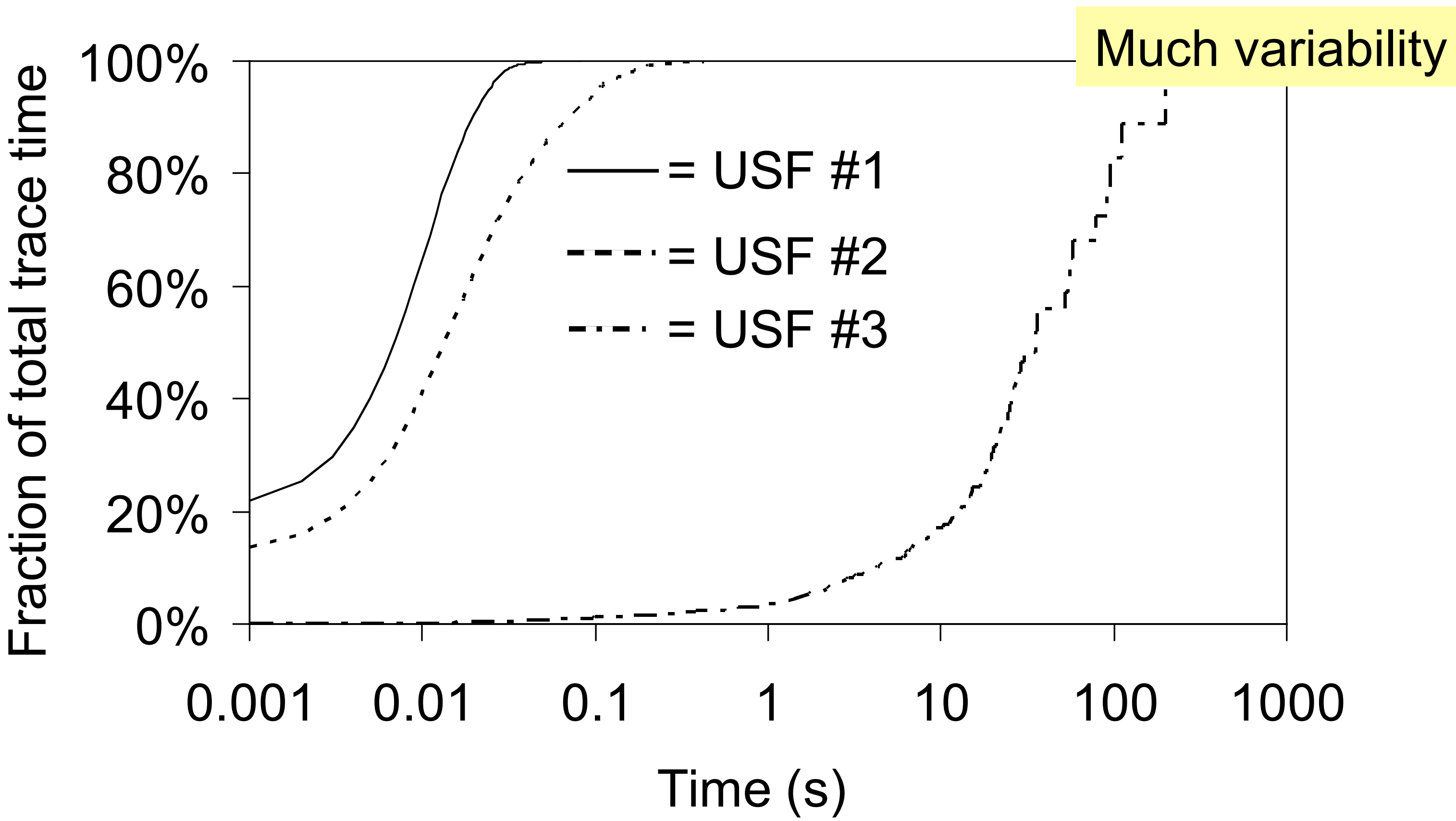
❖ Summary of the traces continued

Large variability

Trace	Mean low util period	CoV of low util period	Mean idle period	CoV of idle period
USF #1	0.0060 s	0.91	0.0011 s	1.79
USF #2	0.0094	1.50	0.0020	2.21
USF #3	1.0892	7.22	0.1100	13.95

Traffic characterization continued

- **Fraction of low utilization periods for USF traffic**
  - For USF #1 and #2, most low utilization less than 100ms



g04.xls

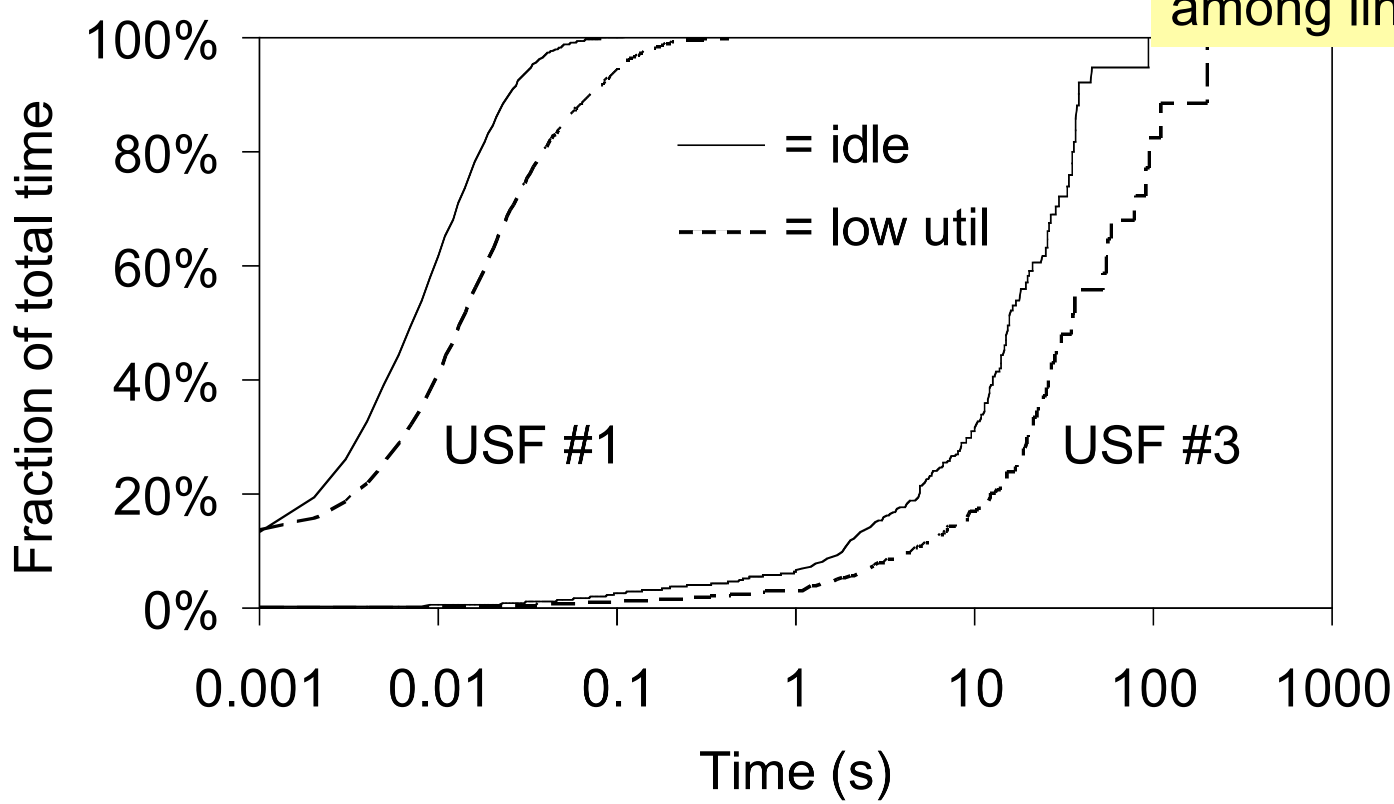
BERKELEY LAB

USF UNIVERSITY OF SOUTH FLORIDA

Traffic characterization <u>continued</u>

- **Idle and low utilization periods together**
  - Example of busiest (USF #1) and typical (USF #3)

Extreme variability among links



Energy and performance metrics

- **Need performance metrics that include energy**

- **Define**
  - ***E*** is energy consumed with no power management enabled
  - $\boldsymbol{E_s}$ is energy consumed with power management enabled
  - $\boldsymbol{D_{bound}}$ is target mean delay bound
  - $\boldsymbol{D_s}$ is mean delay with power management enabled

- **Singh et al. energy savings metric (α)**

$$\boxed{\alpha = E / E_s}$$

- **Our green energy-performance metric (γ)**

$$\gamma = \begin{cases} (E / E_s)(D_{bound} / D_s) & \text{if } D_s > D_{bound} \\ (E / E_s) & \text{if } D_s < D_{bound} \end{cases}$$

Simulation evaluation of ALR

- **Need to study performance of reactive policy**

- **Simulate a NIC (or switch port) buffer**
  - A single server queue
  - Packet arrivals are from traces
  - Packet service is 10 Mb/sec or 100 Mb/sec

- **Key control variables**
  - Target delay threshold (D_{bound})
  - Time to switch between data rates
  - Energy used at 10 Mb/sec
  - Energy used at 100 Mb/sec

- **Response variables**
  - Delay (mean and 99%)
  - Green metric

Results should be representative for 1 Gb/sec case

BERKELEY LAB

USF UNIVERSITY OF SOUTH FLORIDA

Simulation evaluation of ALR continued

- **Experiment to evaluate effect of time to switch rates**

- **Control variable settings:**
  - Queue threshold = minimum of 10 pkts or number of packets that can arrive in a switching time at 5% utilization
  - Utilization measurement period = 100 milliseconds
    - Sampling interval = 0.01 millisecond
  - Time to switch data rate ranging from 0 to 50 milliseconds
  - Energy used at 10 Mb/sec = 4.0 W
  - Energy used at 100 Mb/sec = 1.5 W
  - D_{bound} = 5 milliseconds

- **Response variables collected:**
  - Mean and 99% packet delay (from queueing)
  - Green metric (γ)

Simulation evaluation of ALR continued

- **Cases for simulation experiment**
  - 100-Mbps link rate (no power management
  - 10-Mbps link rate (no power management)
  - ALR case (power management)
- **For each case we collect**
  - Mean and 99% delay
  - CoV of delay
  - Metrics α and γ

Simulation evaluation of ALR continued

❖ Results for USF traces with no ALR

- For fixed 10 or 100 Mb/sec link speed

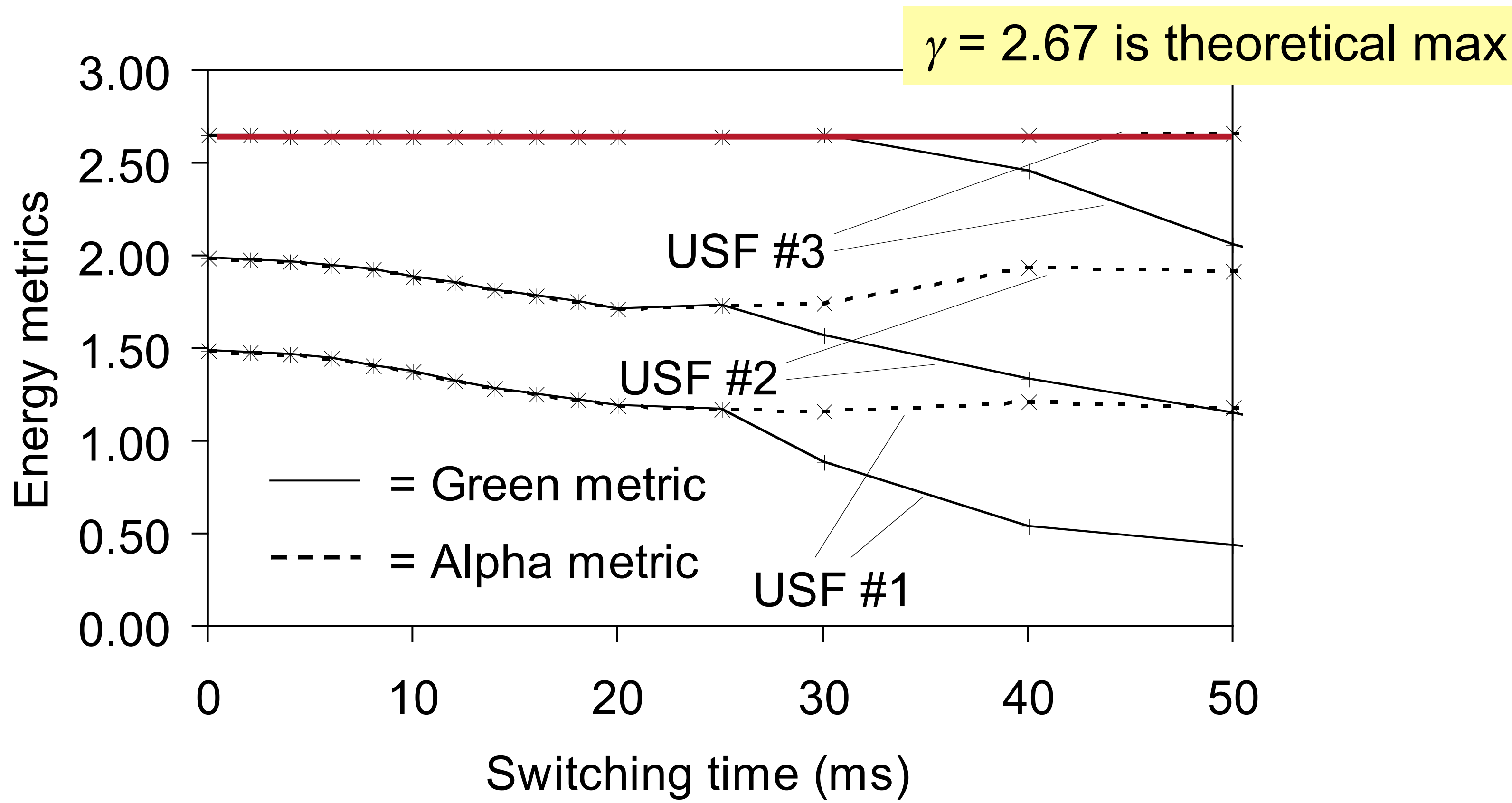
Trace	Mean delay	CoV of delay	99% delay
USF #1	7.60 ms	2.03	77.46 ms
USF #2	3.95	2.62	60.07
USF #3	196.30	1.68	919.24
USF #1	0.09	1.16	0.46
USF #2	0.08	0.93	0.29
USF #3	0.05	1.37	0.26

10 Mb/sec

100 Mb/sec

Simulation evaluation of ALR continued

❖ **Results for energy metrics for USF traces**

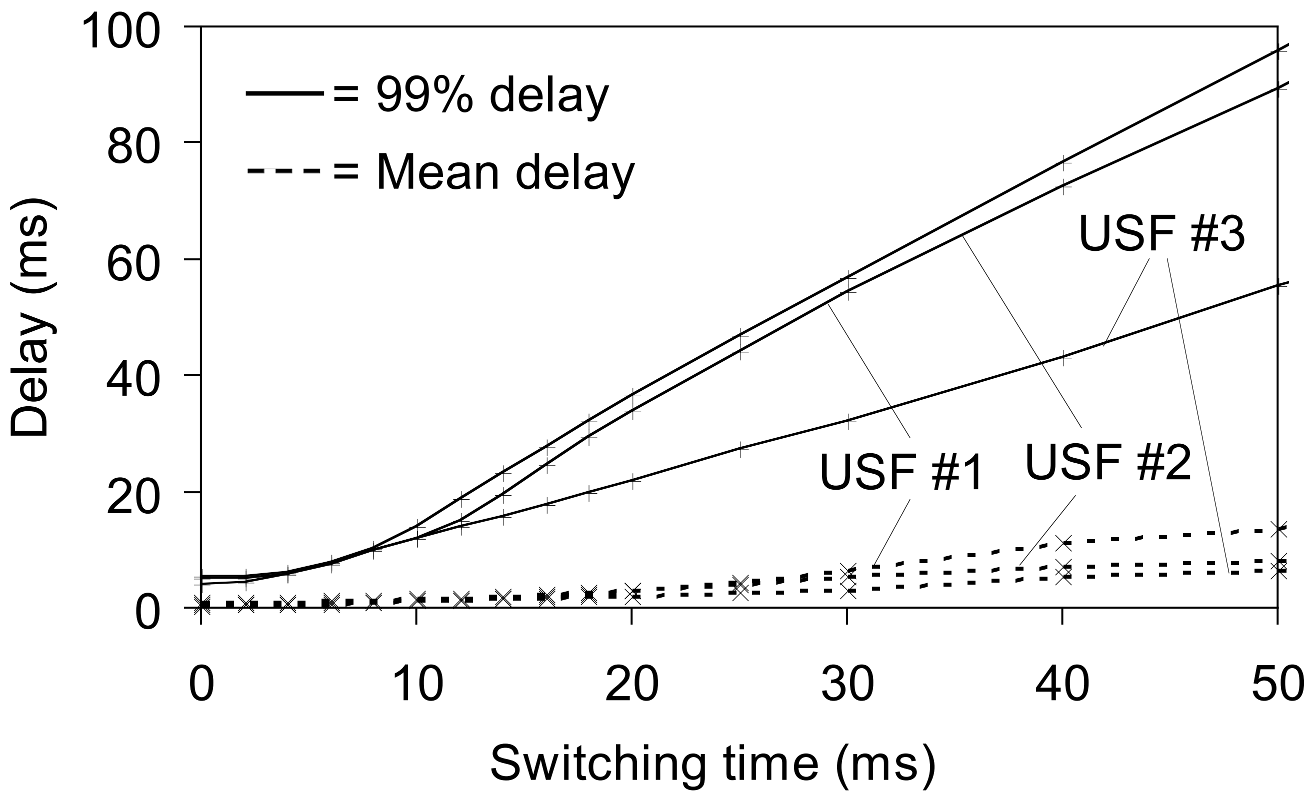


g21.xls

UNIVERSITY OF SOUTH FLORIDA

Simulation evaluation of ALR continued

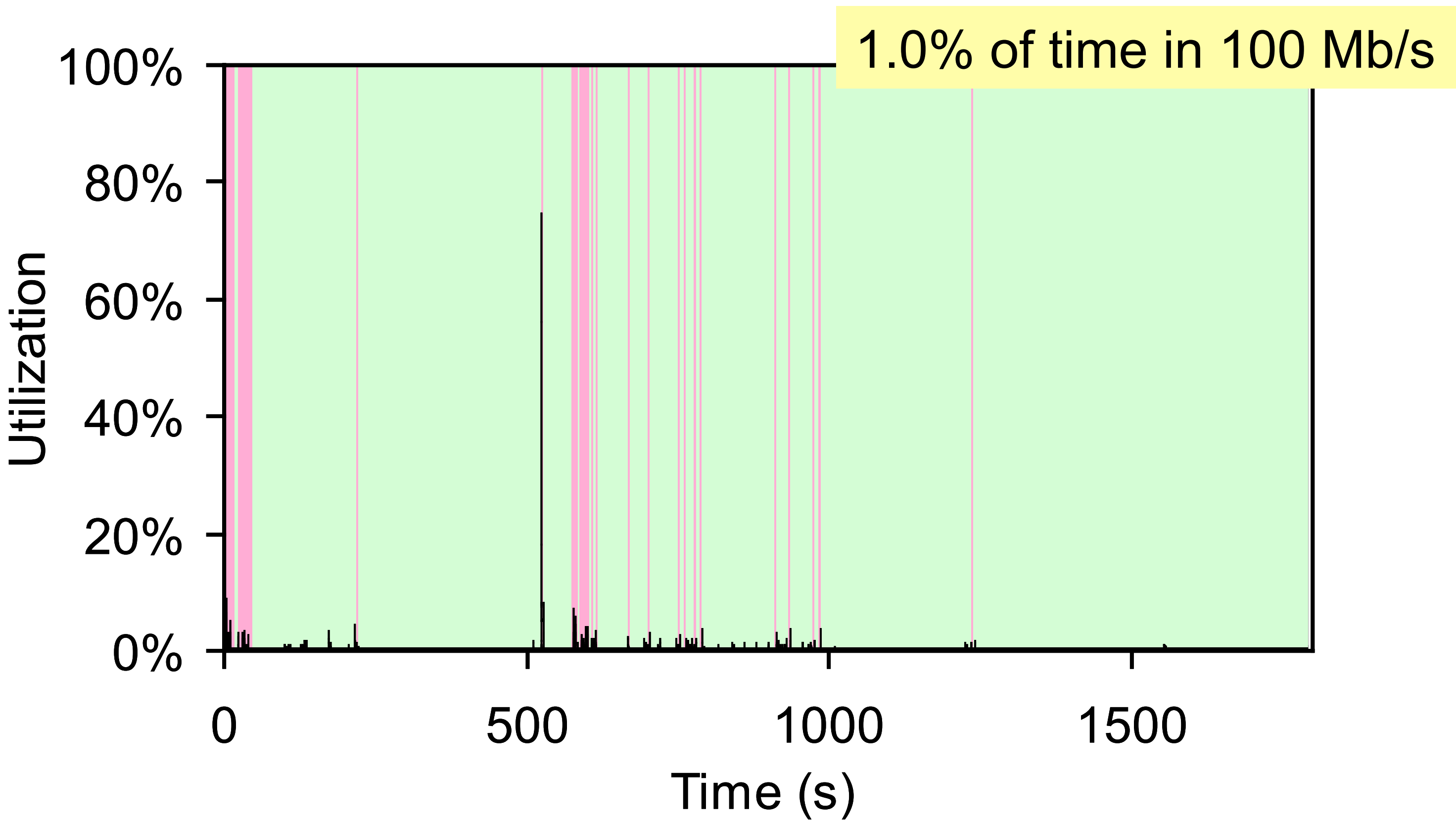
❖ **Results for delay for USF traces**



g23.xls

Simulation evaluation of ALR continued

- **Utilization and link speed graphic**
  - Sample USF trace (USF #1)



g32.xls

Simulation evaluation of ALR continued

Discussion of results…

- Great variation in length of low utilization periods
- Can achieve energy savings and low delay for all traces
- Expect that these results will hold for 1 Gb/sec
- Need to consider energy cost of transition between rates

As with ADSL2, may be very important for MetroEthernet

BERKELEY LAB

USF UNIVERSITY OF SOUTH FLORIDA

Potential Energy Savings

- Welcome to Part #5

In this part… energy savings calculations for the SmartNIC and Ethernet Adaptive Link Rate.

Savings Estimates

- All factors — stock, power levels, usage — not well known and changing

- Conclusions rely on magnitude of savings
  - Not on precise figures

- Assumptions
  - 100 million commercial PCs
  - 100 million residential PCs

  half desktops
  half notebooks

  - Today's power levels

  - Usage patterns — rising # of PCs left on continuously

SmartNIC savings

- **First, consider one Continuous-on PC**
  - 40 hours/week in-use
  - 128 hours/week asleep (was fully-on before SmartNIC)

- **Unit Savings**

	Desktop / Notebook
Annual Electricity kWh/year	470 / 100
Annual Electricity $	$37 / $8
4-year lifetime $	**$150 / $32**

SmartNIC Savings <u>continued</u>

- **Stock-wide Savings**
  - Use unit savings for half of stock

➔ 28 TWh/year; $2.3 billion/year

- **EPA/Energy Star estimate**

If all power managed, US would annually save 25 billion kWh, equivalent to:

- Saving $1.8 billion
- Lighting over 20 million homes annually (all the homes in NY and CA combined)
- Preventing 18 million tons of carbon dioxide (emissions of over 3 million cars)

SmartNIC Savings <u>continued</u>

❖ **Stock-wide average savings**

- Desktop: \$75;   Notebook: \$16
- “Budget” for retail cost of SmartNIC hardware
  - Except for notebooks — SmartNIC adds to functionality

❖ **If SmartNIC adds \$5 to system cost, average payback time:**

- Desktop: About 3 months
- Notebook: 15 months

Highly Cost-effective.

BERKELEY LAB

USF UNIVERSITY OF SOUTH FLORIDA

Adaptive Link Rate savings

- **“Success” rate: Should be nearly 100%**
  - At least once the stock of network equipment turns over
  - Does not rely on system sleep status
- **Average on- or asleep-time of whole stock almost 70%**
  - Take 80% of this as low-traffic time
    - ➔ 55% potential reduced data rate time
- **High data rate**
  - 1Gb/s - 80% of commercial; 20% of residential (50% average)
  - 100Mb/s - 10% commercial; 70% residential (40% average)

BERKELEY LAB

USF UNIVERSITY OF SOUTH FLORIDA

Adaptive Link Rate savings continued

- **Per unit savings** (counts both ends of link)
  - 1Gb/s - 10 kWh/year $3.20 lifetime
  - 100 Mb/s - 3 kWh/year $0.96 lifetime

- **Cost-effectiveness**
  - Hardware cost should be minimal or zero; modest design cost
    - ➔ Very short payback times

- **Stock-wide savings**
  - 1.24 TWh/year

➔ **$100 million/year**

Summary and next steps

- Welcome to Part #6

In this part… we summarize the key points and discuss the next steps needed to energy savings.

IT equipment uses a lot of energy

- All electronics about $16 billion/year of electricity
- PCs about $3.7 billion/year
- … and both growing …

Networks induce energy use

- **Many products must stay in a higher power state than otherwise needed to maintain connectivity**
  - 802 networks
  - USB (some implementations)
  - TV set-top boxes (many)
  - and more…
- **Network applications increase on-times**
- **… and growing …**

BERKELEY LAB

USF UNIVERSITY OF SOUTH FLORIDA

Networks directly use energy

- Network interfaces and network products
- Combined about $1 billion/year
- … and growing …

Large savings potential

❖ SmartNIC

- Now: $2.2 billion/year
- Future savings growing
  - More PCs
  - More non-PC products with network connections
  - Longer on-times
  - Growing difference between On and Sleep power
- Savings highly cost-effective

❖ Adaptive Link Rate

- Now: $100 million/year
- Future savings growing
  - More products with network interfaces
  - Higher speeds lead to (much) greater base power level

IETF for sleep friendly systems

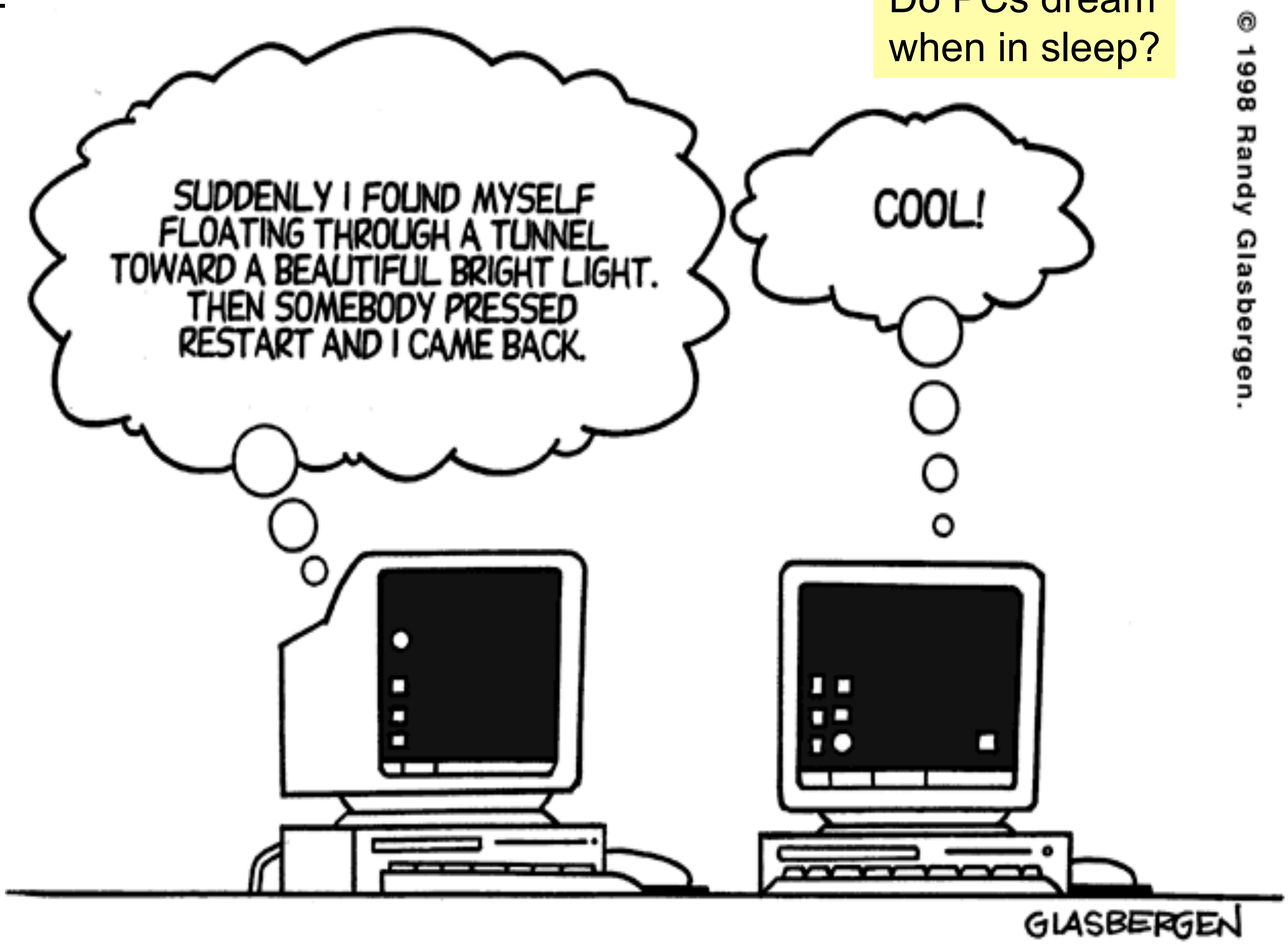
- **IETF** (or similar organization) **should:**
  - Create a study group on the topic
  - Define generic proxy functionality (internal and external)
  - Define data exchange standards between OS and NIC
  - Create guidelines for sleep-friendly software
- **Implementation**
  - Energy Star could help educate consumers, transform markets

IEEE 802.3 for adaptive link rate

- **Form study group**
  - 1G NICs
  - 10G NICs (copper and fiber)
  - Assess implications for wireless (or different study group)

- **Implementation**
  - Roll capability into all NIC products

Do PCs dream when in sleep?



Questions / Comments

Bruce Nordman
Energy Analysis
Lawrence Berkeley National Laboratory
Berkeley, CA 94720
bnordman@lbl.gov

Ken Christensen
Computer Science and Engineering
University of South Florida
Tampa, FL 33620
christen@cse.usf.edu

BACKUP SLIDES

Reducing energy in power distribution

- **Power distribution is the first point of inefficiency**
  - UPS causes loss
    - Use of UPS is increasing
  - Type of power supply matters
    - Switching versus series regulated
  - Number of power supplies matter
    - More efficient may be one DC supply per rack
    - Power over Ethernet may improve efficiency in this way

Substantial savings still possible in the “analog” realm

Reducing energy in processors

- **Processor is the main energy consumer in a PC**

Graphics unit may be main energy user in a game unit.

  - Within a chip can turn-off and/or scale clock to components
    - Nanosecond time scale
    - Use predictive strategies
  - AMD PowerNow, Intel PowerStep, and Transmeta LongRun
  - "… delivering just enough performance to satisfy the workload at hand."
    - Transmeta LongRun brochure

Processor level has no "view" of long time scale events

BERKELEY LAB USF UNIVERSITY OF SOUTH FLORIDA

Reducing energy in wireless networks

- **Wireless networks can be mobile and ad hoc**
  - Very expensive to transmit (wireless is non-directional)
    - Processing and storage require much less power
  - New routing protocols
  - New data distribution methods
  - New approaches for data fusion

From sensor network research community

Does not apply to existing Internet protocols

Reducing energy in supercomputers

- **Energy use is the limiting factor in supercomputers**
  - "If current trends continue, future petaflop systems will require 100 megawatts of power…"
    - Cameron et al. at USC (2005)
  - 100 MW is $8000 per hour!
    - This does not include cooling costs!
  - Current work is in characterizing program execution
    - Goal is smarter program scheduling

Does not apply to "ordinary" desktop applications

BERKELEY LAB

USF UNIVERSITY OF SOUTH FLORIDA

Reducing energy in data centers

- **Energy use is a major cost component in data centers**
  - Cooling is 25% of operating cost
  - Data centers use clusters of mirrored Web servers
  - Exploring ways to power on/off servers as a function of request rate
    - Keeping response time below a threshold is the goal
  - NSF funded work at several universities

Does not apply to "ordinary" desktop applications

Reducing energy in corporate PCs

Central control of Windows power management

- Use a centralized management PC to control Windows power management settings in desktop PCs
  - Lock-out users from disabling power management
- At night use “aggressive” power management settings
  - Short delay to sleep and possibly even turn-off PCs
- During the day use “lite” power management settings
  - Long delays to sleep and no use of off
- Verdiem Surveyor and other products

Does not address root problems and not useful for residential PCs

BERKELEY LAB

USF UNIVERSITY OF SOUTH FLORIDA

Reducing energy in displays

❖ **Displays are proliferating, but are not always watched**

- LCD displays require less power than CRTs, however multiple displays per desktop is becoming normal
- Can use camera to detect if person is watching display
  - Camera is an “occupancy sensor”
- “FaceOff” at Duke University to power manage a notebook
  - Dalton and Ellis

User context need to play a role in power management

SmartNIC requirements <u>continued</u>

- **Hard part is determining what is a "typical" PC**
- **Usage patterns for home and office differ**
  - Home PC…
    - P2P file sharing
    - Entertainment center controller
    - Part of a UPnP network
  - Office PC…
    - File sharing via network drives
    - Always connected to a database
    - Remote access from home or travel
    - Nightly s/w patches, virus scans, etc.

Microsoft Windows and IP protocol are in common

- Home and office blur together in notebook computers