Improving energy efficiency for networked applications

Jeff Mogul -- Jeff.Mogul@hp.com
HP Labs, Palo Alto

If energy efficiency matters to us, and networked applications matter to us, doesn't the energy efficiency of networked applications also matter to us?

My goal: An end-to-end view

- Understand where the energy gets used
  - At all components in an end-to-end application
- Look for places to improve efficiency
  - Focus on Barroso's "Proportionality Principle"
- Think about benchmarking and measurements

Caveats

- I’m not really an expert
- Most of this isn’t my own work
  - I’ve been a little sloppy about citations
- Don’t trust my numbers
  - You’ll even see some contradictions in this talk
- I will pose questions, but not always answer them

Motivation

Why energy efficiency matters

Per recent EPA study, as of 2006, US data centers:

- Consumed ca. 61 billion KWh of energy
- Accounted for ca. 1.5% of US total electricity use
- At a cost of $4.5 billion
- Doubling time for DC energy use: about 5 years

Not all data center applications are networked, but:

- Many are
- The trend is towards networking in every application

Why energy efficiency matters

Another cut at the big picture

- 5000 data centers
  - @ 2MW each for computing
  - + 2MW each for cooling
  - = 20GW world-wide
- 4,000,000,000 handhelds + PCs (more or less)
  - @ average of about 12W each (?)
  - = 48GW world-wide
- Don’t ignore the end-hosts!

What does this mean in terms of CO_2?

- Perhaps 175 million metric tons/year
- Could save 47M metric tons of CO2 by 2011
  - Best case, per 2006 EPA study
- Global warming results from an excess of ca. 20 billion metric tons/year
  - IT as a whole: less than 1% of the problem
  - Vs. "all fossil fuels" = ca. 23 bmt/year
  - But solution may depend on lots of small improvements

How much of all this computing-related energy goes into networking?

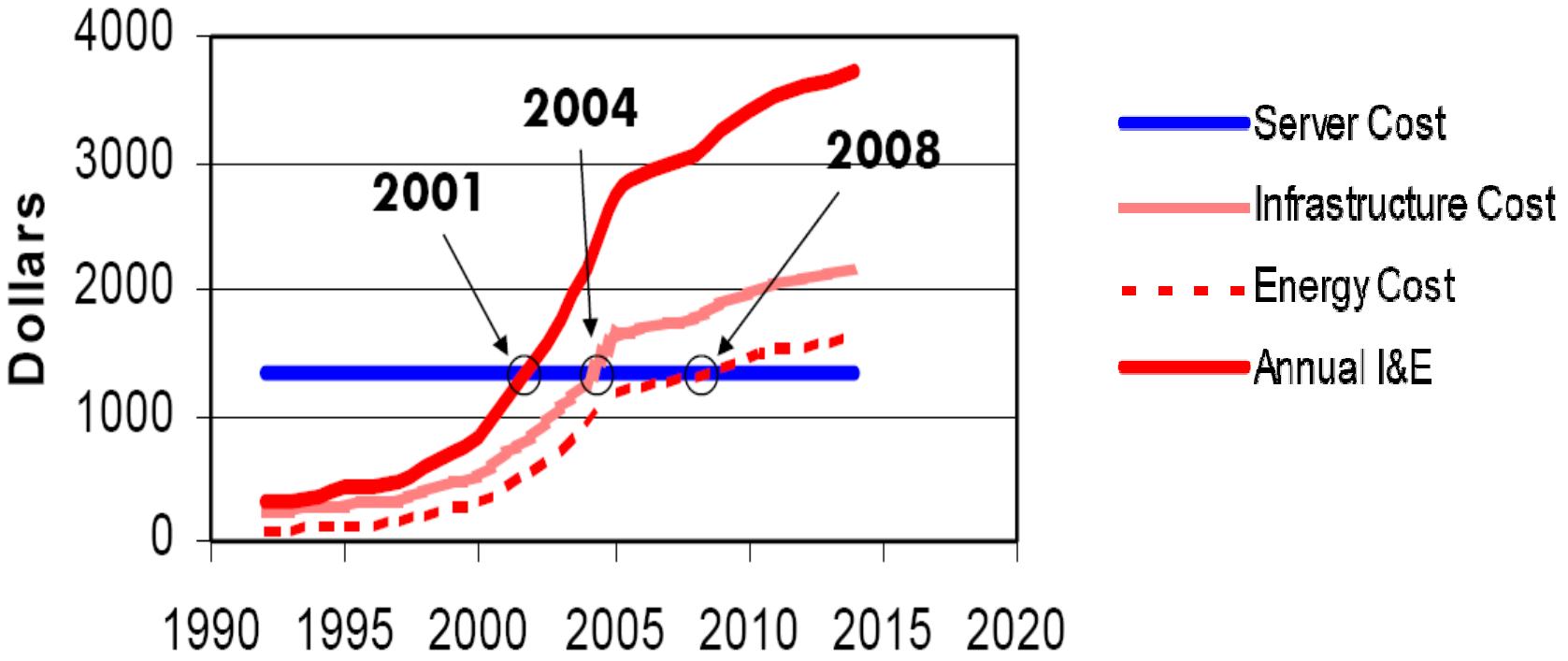
Estimates from various sources via Nordman + Christensen [2005]:

- The numbers are old – mostly year = 2000:
- "Big IT" – all electronics: 200 TWh/year
- "Little IT" – offices/telecom/data centers: 97 TWh/year

- Possibly "networking": 44%
  - PCs: 31 TWh/year (perhaps 46 TWh/year in 2005?)
  - Servers: 12 TWh/year
- Definitely "networking": 18%
  - NICs: 5.3 TWh/year
  - Switches/Hubs/Routers: 6.05 TWh/year
  - Telecom: 6.1 TWh/year

Ultimately, perhaps half of this is "networking related"

Energy: cost matters, too



- Infrastructure & Energy Costs 2X price of a 1U server
  - In the future, the I&E costs will be even higher

Source: Malone, C.G., Belady, C., 2007, 'Metrics and an Infrastructure Model to Evaluate Data Center Efficiency,' Proceedings of IPACK 2007, Vancouver, BC. IT & cooling power & electricity cost of $0.1/kW-hr

Some principles

Basic principles

- Barroso's Proportionality Principle
  - Energy use should be proportional to system activity
- Design for efficiency vs. peak performance
  - A little more peak performance isn't worth a big efficiency loss
- Get things done as fast as possible
  - A system that is waiting is a system that is wasting energy

Barroso's Proportionality Principle

See "The Case for Energy-Proportional Computing," Luiz André Barroso and Urs Hölzle, IEEE Computer, Dec. 2007

- Server systems mostly between 10% and 50% utilized

Energy use should be proportional to system activity

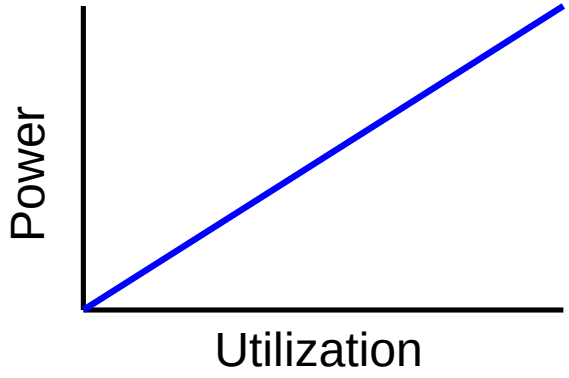
- Ideally, $P_{idle} = 0$
  - This may be an unreachable ideal, in most cases
  - Implies a wider dynamic power range than for current systems – these generally have $P_{idle} >> 0$

Why proportionality helps

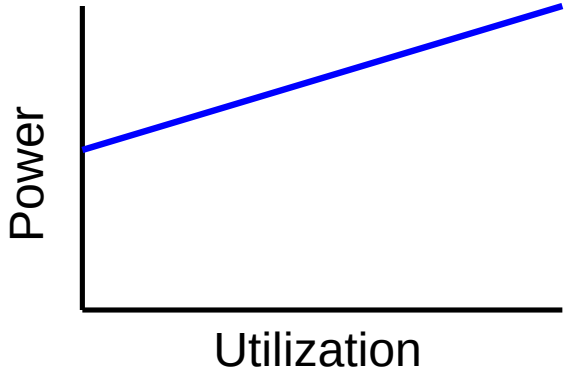
Non-proportional systems force us to actively manage for energy efficiency – this is a bug

- E.g., moving load around so we can shut off servers
- Explicit power state transitions can add too much latency, especially in server applications
- Managing things requires information we might not have
- Managing things can lead to mistakes

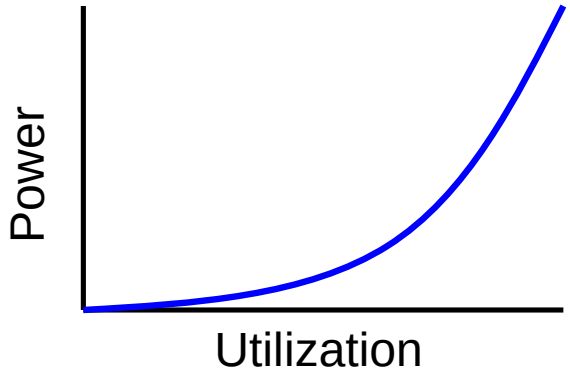
Examples of power vs. utilization curves



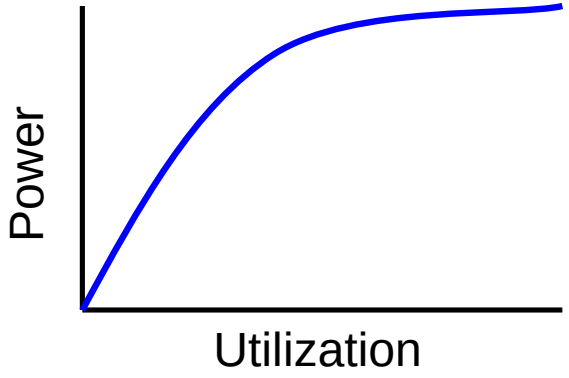
Proportional, $P_{idle} = 0$



Proportional, $P_{idle} >> 0$

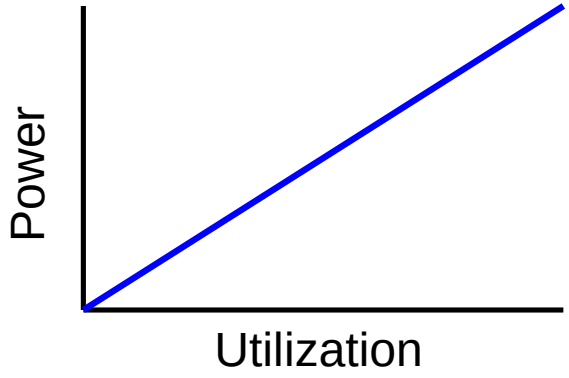


Could be better than simply linear, but makes management harder

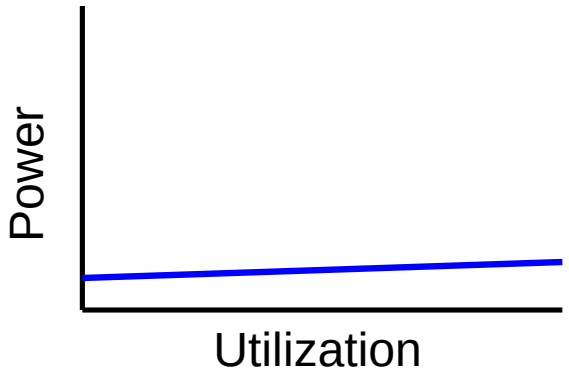


Not proportional, OK only extremes of utilization

But don’t neglect basic energy efficiency



Proportional, $P_{idle} = 0$



Not proportional, but generally much better unless system is often idle.

Intrinsic vs. managed proportionality

- Managed proportionality: a higher-level or external manager needs to configure a component to align energy and load
  - E.g., choose a power state or power-down a component
- Intrinsic proportionality: each component naturally consumes energy proportional to load
  - E.g., variable-speed fan with temperature sensor
  - Generally this is more robust
- Managed proportionality at one level can look like intrinsic proportionality from a higher level

Efficiency vs. peak performance

Researchers and marketers tend to focus on peak performance:

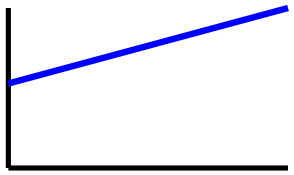
- Easiest way to make quantified comparisons
- But: leads to diminishing returns – small increase in peak performance costs big increase in energy
  - Violates proportionality principle

Barroso and Hölzle observe:

- Mobile devices generally either “very idle” or “very peak”, so designers tend to underemphasize energy efficiency at mid-level utilizations

Get things done as fast as possible

- Most systems work like this:
  - $P_{idle} >> 0$

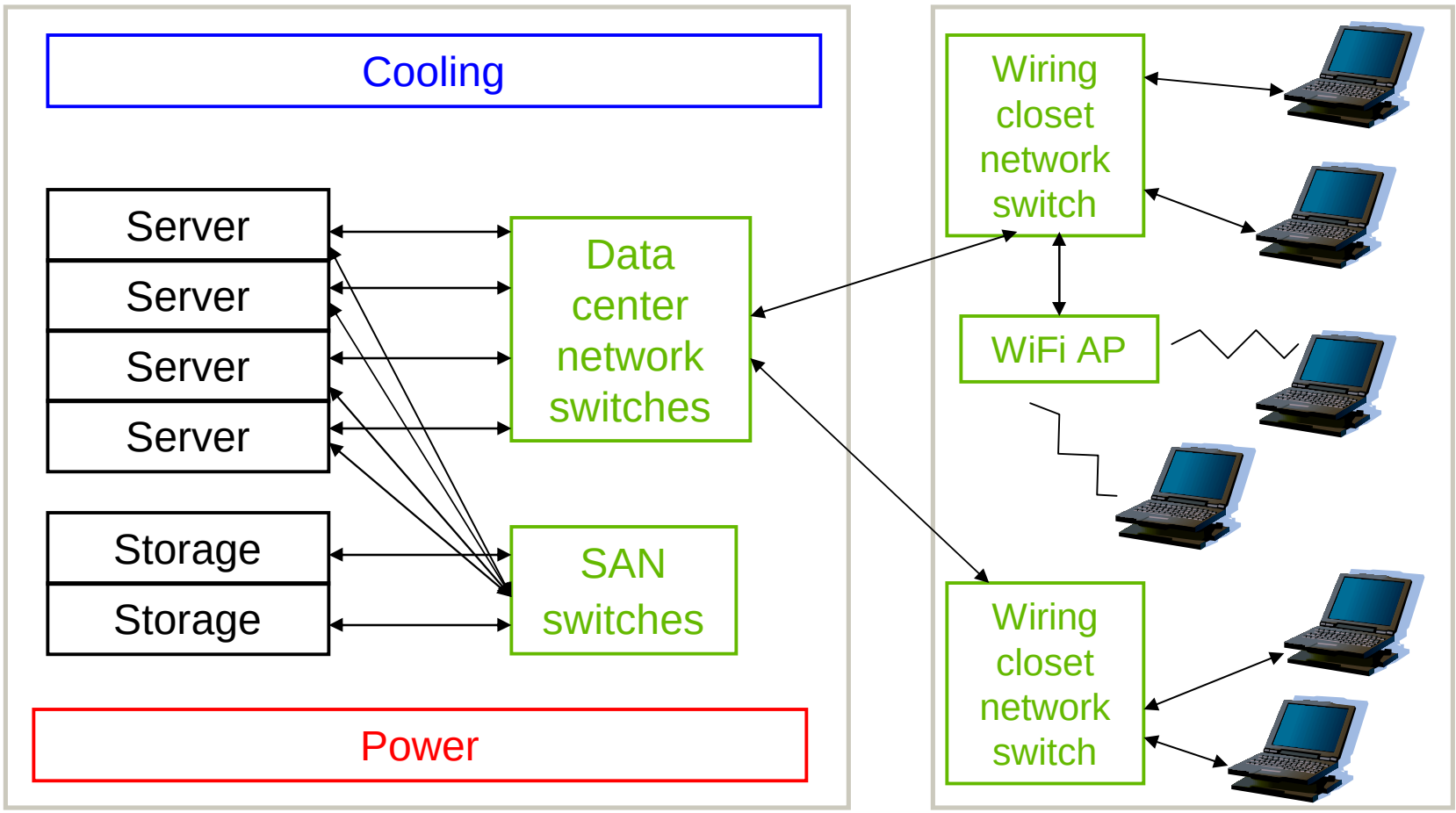


- Leaving components idle while they wait for something to happen wastes energy
- If efficiency is defined in terms of tasks completed per unit time per joule, then generally it's optimized by not wasting any time
  - "Watching a movie" is not this kind of task

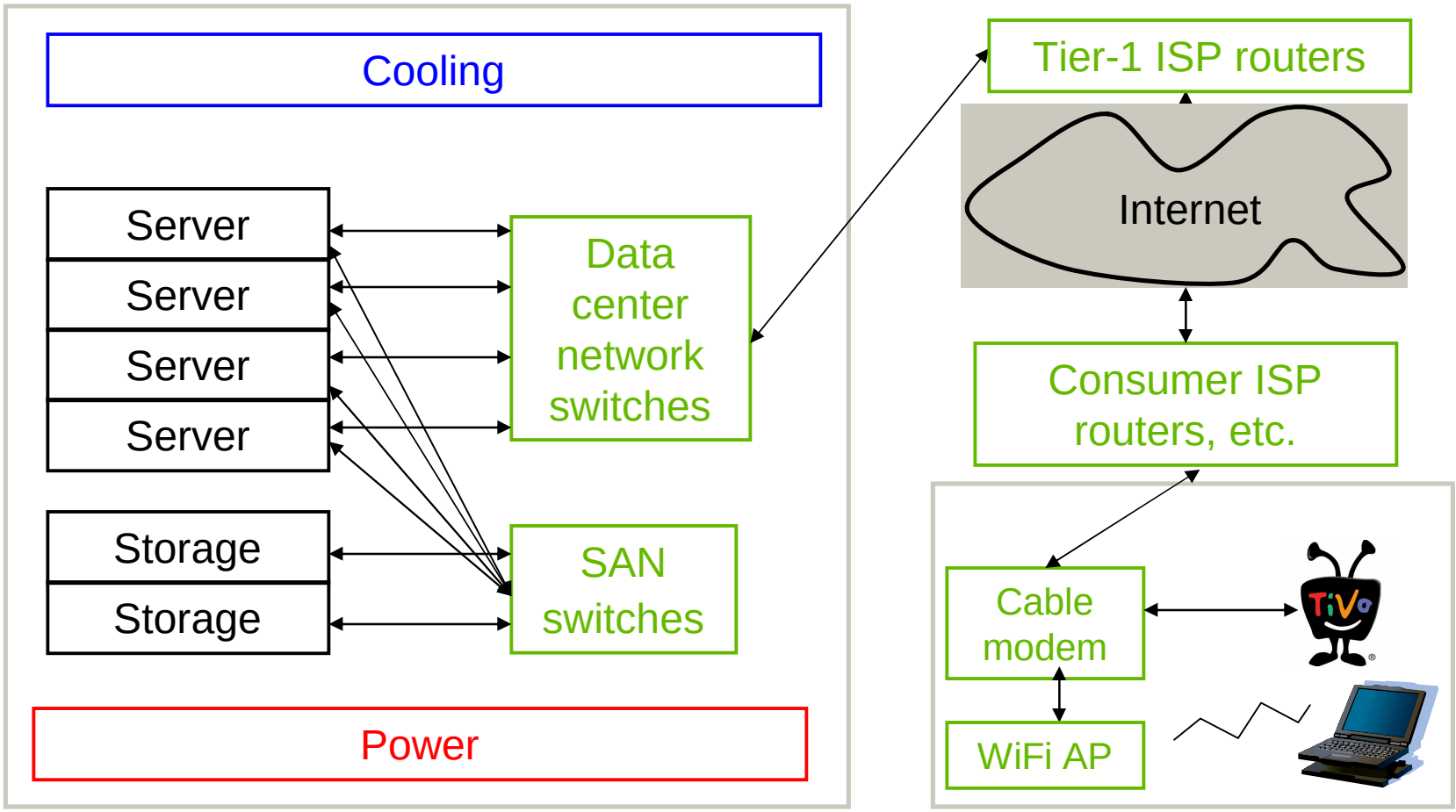
Where is the energy consumed?

… and where could we consume less?

A schematic diagram: Local-area application



A schematic diagram: Wide-area application



Data center

Grandma's house

hp invent

Places to look for energy inefficiencies

The data center:

- Power provisioning
- Cooling
- Servers (CPU, RAM, NICs, storage)
- Network switches and routers

The long-haul network:

- POPs as instances of data centers
- Long-distance transmission of bits

The other end of the network:

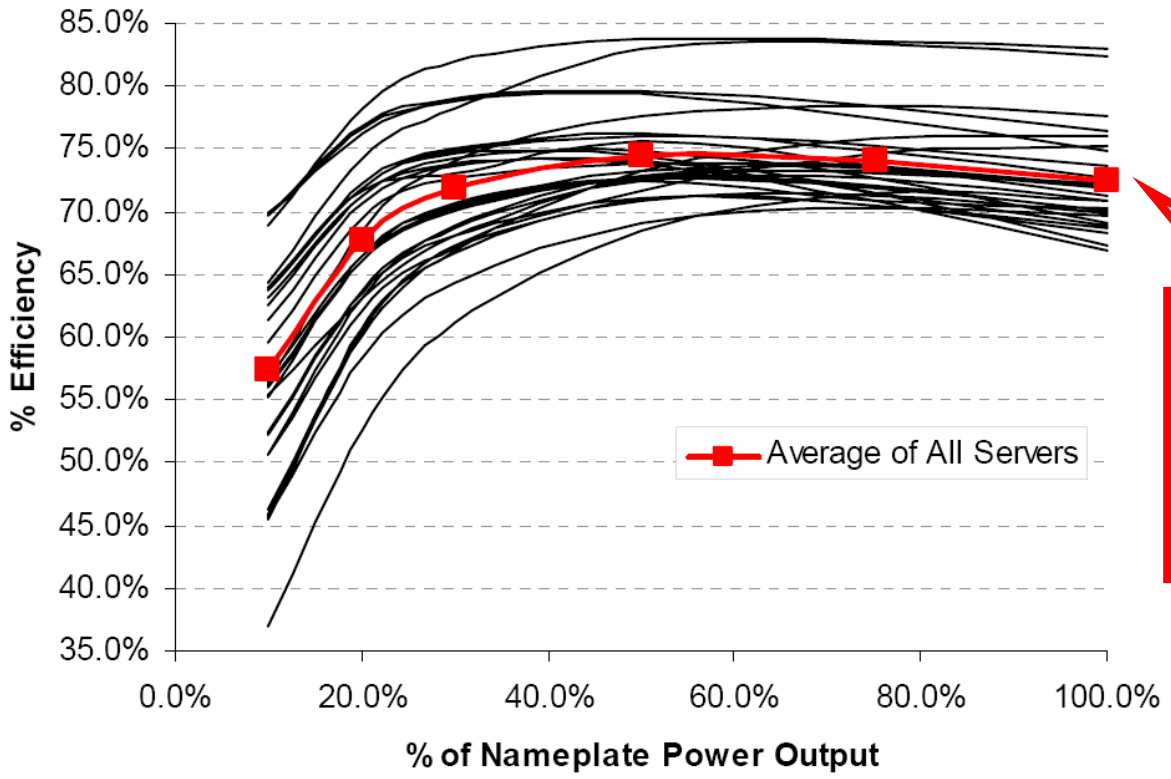
- Home-user energy costs
- Could also be another data center

Data Center power provisioning

- Data centers must be provisioned for peak power
  - Otherwise, circuit breakers trip
  - Usually a firm upper limit on available watts from power co.
- Power delivery infrastructure loses ca. 10%
  - Power line + distribution losses
- Data centers need backup power – more inefficiencies
  - UPS: 70%--95% efficient – but generally, smaller ones less efficient
  - Flywheels: 93%--98% efficient (but low efficiency near idle)
  - These systems are always in the loop
- Diesel generators: 30%--55% efficient
  - Energy input is diesel fuel
  - Not normally running, though

Power supply efficiencies

- Typically 55%--75% efficient; getting to 85%--95%
  - Efficiency declines at lower loads (non-proportional)
  - N+1 / N+2 redundancy ca. ½ as efficient on average



Some efficiency decline at peak loads, too

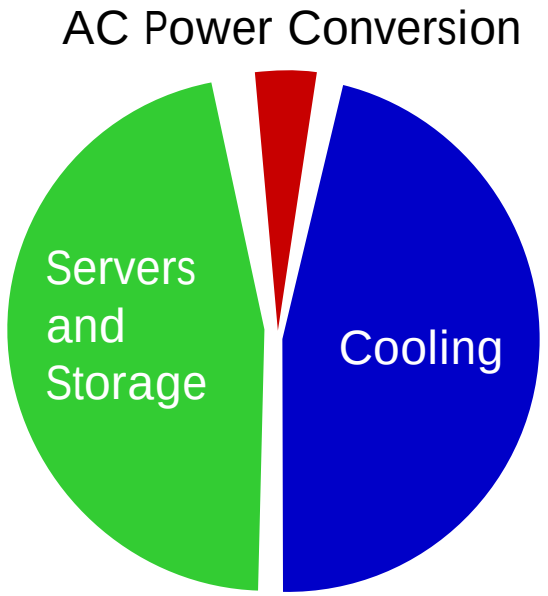
From: M. Ton, B. Fortenberry (2005, LBNL), "High-Performance Buildings: Data Center Server Power Supplies"

hp invent

Data center cooling

- Traditional data center cooling is inefficient
  - Rules of thumb are often wrong
  - Simplistic feedback loops force over-cooling

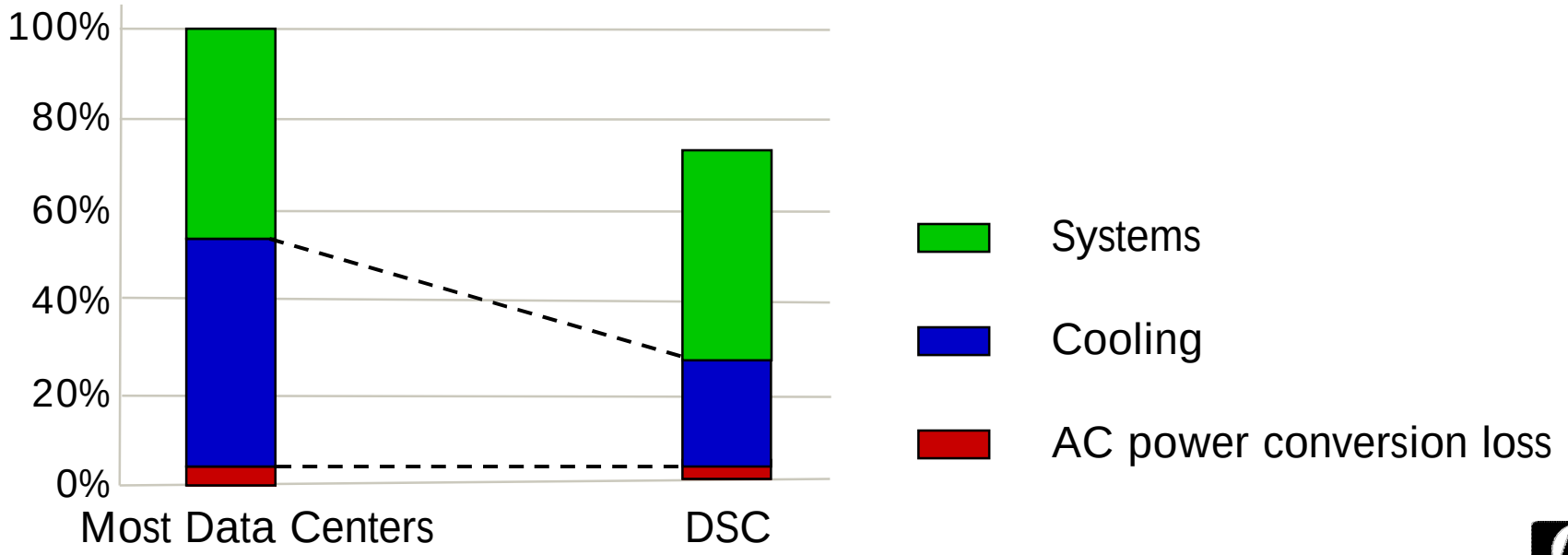
- Cooling represents about **50%** of an "average" data center's power consumption
- Most data centers are over-provisioned w.r.t. cooling

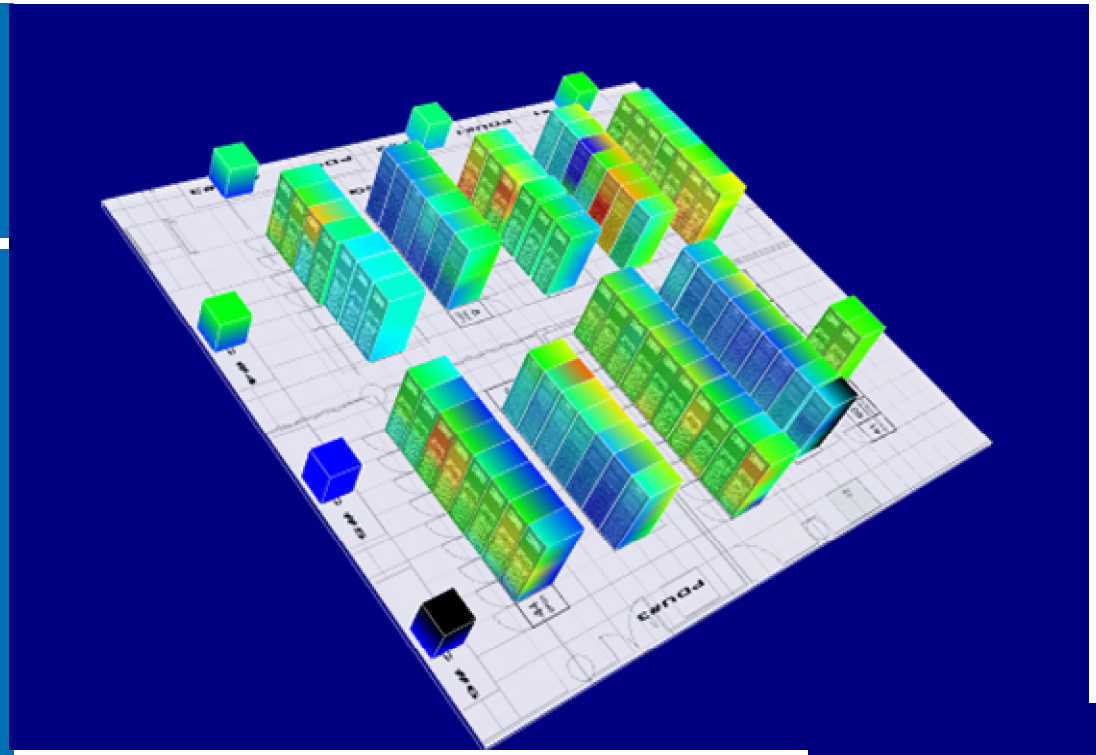


Belady, C., Malone, C., "Data Center Power Projection to 2014", 2006 ITHERM, San Diego, CA (June 2006)

Cooling can be much more efficient: "Dynamic Smart Cooling"

- Use good models for heat flow
- Use many more temperature sensors (sensor network!)
- Use fine-grained control over airflow
- Use reactive control
  - Cooling is proportional to computing energy load

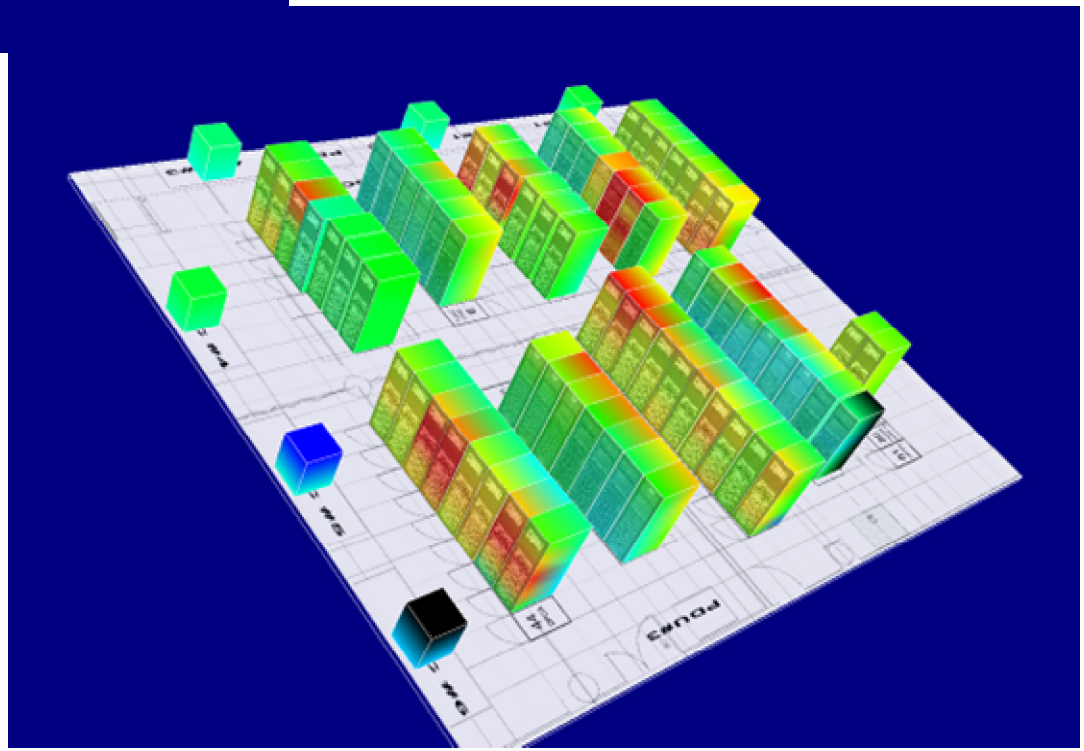




HP Labs Data Center

3D Real-Time Imaging
of Measured
Environmental Data

Conventional Mode



Dynamic Smart Cooling Mode

- 35% Energy Savings
- Improved reliability
- Improved AC infrastructure utility

Server hardware

Peak energy goes into:

- CPUs: 28% to 51%
- RAM: 19% to 31%
- Disks: 1% to 4% (more for a storage server!)
- PCI slots: 7% to 20% (when fully used)
- NICs: 2% to 6% (when all ports used)
- Fans, Power supply, other stuff: 7% to 22%

Based on plausible server configurations

From Table 3-3, Kevin Leigh, “Design and analysis of Network and IO consolidations in a General-Purpose Infrastructure”, PhD dissertation, University of Houston, 2007

Peak-to-idle power ratios for servers

Barroso and Hölzle observe:

- "even an energy-efficient server consumes ca. half its full power" when idle
- Most servers run 20% -- 30% utilized, so at less than half of their best-case energy efficiency

Where does the idle power go?

- CPUs: 1/10 to 1/3 of peak power
- DRAM: ½ of peak power
- Disks: ¾ of peak power
- NICs: little or no decline from peak?
- Other stuff: not clear

How can we improve server power efficiency?

- CPUs
  - Lots of work going on here – no more low-hanging fruit?
  - Some create proportionality
    - e.g., clock-gating; voltage/frequency scaling
- RAM
  - E.g., avoid refreshing "dead" rows in DRAM
    - Proportional to RAM use if fragmentation is limited
    - OS support could help?
- Disks
  - Can use Flash in specific (niche) applications
    - Could increase chances for disk spin-down?

What about NIC power efficiency?

NICs don't take a lot of the power now (2%--6%)

- But mostly non-proportional (?)
- Future NICs may take more power (TOE, RDMA)
  - 10GB copper Ethernet might be as much as 14W/NIC
  - Vendors claim 5W/NIC is feasible

What can be done?

- NIC CPU: should improve analogous to GP CPU
- Physical interface: improves w/shift to new PHY
- Later I'll describe "Energy-efficient Ethernet" work

How the Operating System can help

- Usual power-management stuff
  - Maybe this can be improved, but it looks iffy
- "Tickless kernel"
  - Idle system doesn't wake up CPU on every clock tick
  - Available in Linux 2.6.21
  - Incremental power savings might be small
  - Except: when multiplexing lots of nearly-idle VMs on one physical machine?

How multi-core can help

Obtain proportionality by shutting down some cores under low load

- Can respond relatively fast to load increases

Asymmetric multi-core:

- Some complex cores with high peak performance
- Some simple cores with low power consumption
- Shift application threads based on their phases
  - See Kumar et al. in MICRO-36, 2003
- Or: use simple cores just for OS functions
  - Complex cores mostly wasted (space, watts) on an OS

Application Software

- Load-spreading / load-balancing:
  - good for peak performance, robustness against failure
  - Bad for energy proportionality
  - Depends on time scale of proportionality
    - Over scales >= minutes, load-balancers could manage energy
- Are DHTs bad?
  - DHTs spread load randomly, by design
  - All nodes must be always-on … or can we do better?
- botnets – they are an “application”!
  - how much energy do these waste?

Switches and Routers

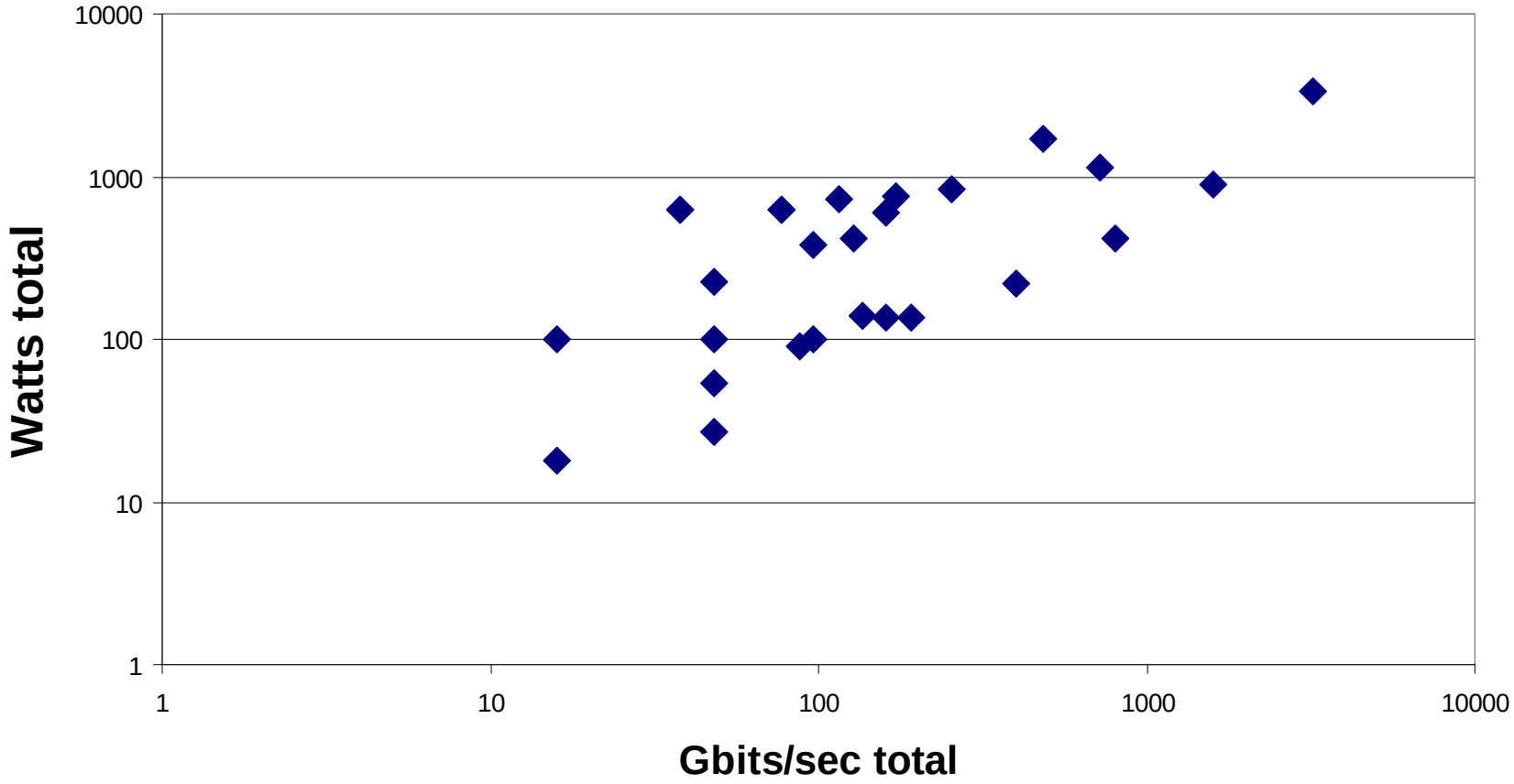
One estimate: 15% of data center compute power

- Not including firewalls
- Blade servers may reduce this somewhat

Switch power is non-proportional:

- Switches rarely offer low-power modes
- Usually over-provisioned (for hot-spots + failures)

How much power do LAN switches consume?
(Watts vs. Gbits/sec total for random switches)



Note: log-log scale

invent

Notes on preceeding plot

- "Name-plate" power from several manufacturers
  - But some make it harder to pick useful numbers
- Watts/port might be a better metric
  - But definitely varies by port type – see next slide
- Not weighted by number of ports in service!
  - In reality, lots more ports on lower-speed switches
- Other caveats:
  - Redundant power supplies will affect results
  - Biased towards systems with easy-to-understand spec sheets

Data transmission: within the data center

- Data centers are full of cables
  - Large cables block airflow, make cooling inefficient
- Driving data (high speed, low noise) takes energy
  - Copper 10GBase-T: ca. 5W—11W "per end"
  - Fiber @ 10Gbit/s: ca 0.5W per end (e.g., Finisar)
  - Typical InfiniBand (IB): 1W – 2W per end
  - IB w/active ends + low-loss: 0.25W/end (e.g., Gore)
  - (Within blade chassis: 10—20 mW/Gbit/sec)
- Multiply by several ports per server
  - Don't forget the other end of each link!
- And not proportional to load – ports always on

Data-center-wide approaches to improving network power efficiency

- Problem: always-on switches, low utilization
  - Could consolidate server load in subset of racks, then turn off unnecessary routers
  - Could add redundant low-power / low-throughput switches, then power-on switches appropriate to load
    - Ideas from Partha Ranganathan, Puneet Sharma, Sujata Banerjee
- Switch / server reconfigurations may take minutes or longer
  - Can we do anything at finer time scales?

Where in the network infrastructure does the energy go?

Data from 2000 NTIS study (after Gupta & Singh, 2003):

Device	# deployed	Annual 10^{12} WH
Hubs	93,500,000	1.6
LAN switches	95,000	3.2
WAN switches	50,000	0.15
Routers	3,257	1.1
Total		6.05

0.07% of US total for 2000

Most of this energy goes to LANs

- But a few WAN routers take a large chunk
- Table doesn't account for cooling, backup power

How much energy does it cost to send a byte over a long distance?

Gupta & Singh ("Greening of the Internet", SIGCOMM 2003) estimate:

- 0.128 to 0.225 Joules/byte
- Based on 20 – 35 x 10^{15} bytes total in 2000

Gupta & Singh suggest:

- Re-routing during low activity
- Sleeping router components when no packets
  - Their AS: inter-arrival time > 200 ms for 91% of packets

IEEE Energy-Efficient Ethernet (EEE) study group

- Formed in late 2006
- Basic approach: change PHY based on activity
  - High-power 10 GB/s PHY during peak activity
  - Low-power 1GB/s PHY during low activity
    - Could go even slower, but energy savings less significant
  - Might save 80% of NIC energy?
  - Resynchronization may take ca. 3--4 msec?
    - Longer for up-shift to 10GB? – this isn't easy!
  - Control policy decides when to change speed
    - Both ends of link must cooperate to avoid thrashing
  - Details at: http://www.ieee802.org/3/eee_study/

Home users of the Internet

- Networked applications drive most(?) of home computing
  - Email, Web, IM, games, video, MP3s, etc.
- Increasing use of home networking leads to an increase in always-on systems
- "36 million home offices in the US" (IDC)
  - Not clear what fraction are used for full 8-hour days

Some data from "Extreme energy makeover: Home office edition" by Robert Mitchell, Computerworld, Nov. 8 2007

Home-network power consumption ("Typical" of modern equipment)

- Possibly proportional:
  - Desktop: 50W – 90W on, ca. 10W standby
  - Gaming PC: 300W or more
  - My laptop: 17W – 35W on, 1W sleep
  - LCD monitor: 15W – 30W on, 1W – 2W standby
  - Networked printer: 65W on, 2W standby
- Always-on:
  - WiFi router: 5W
  - Cable modem: 5W
  - Backup storage: 17W (or more – data is hard to find)
  - TiVo: 28W
- And those blobby AC adapters really suck … energy
  - 30% to 60% efficient, per EPA
  - ENERGY STAR adapters are at least 30% better than old average

Client systems at the office

How many people leave their desktops running all night for networked functions such as:

- Backups
- Patch scans and updates
- Virus checks
- etc.?

Networking + virtual machines could help:

- VM migrates to server before desktop turns off
- Maintenance function wakes up VM as needed
- See CMU's "Internet Suspend/Resume", Stanford's "Collective"

“Sleep-friendly PCs”

Unpublished(?) work by Bruce Nordman (LBNL) and Ken Christensen (U. South Florida)

http://www.csee.usf.edu/~christen/energy/main.html

Goal: sleep as much as possible without losing “network presence”

Approach:

- Reliably wake up PC exactly when needed
- Expose PC’s power state to rest of network
- Proxy functions run in a “SmartNIC”
  - low power, protocol-aware for background traffic

Keeping things in context

Don’t get too focused on computers

Plasma TVs probably consume more than home PCs

- E.g., 50” TV = 450W
- Hours/day that TV is on in average U.S. home:
  - 6 hours, 47 minutes

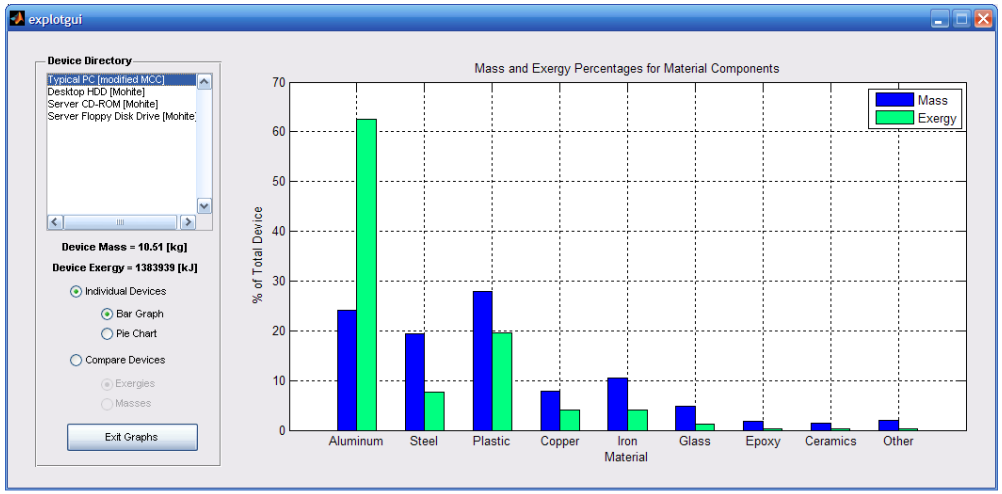
Some data from “Extreme energy makeover: Home office edition” by Robert Mitchell, Computerworld, Nov. 8 2007

Your home office probably has several of these

Ca. 75 W

Ca. 20 W

And don't forget the energy required to build the hardware!



Prediction of "exergy" (available energy) destroyed during raw material extraction (better metric than mass)

- nearly 400 kWh of exergy required to extract materials for example laptop
- Aluminum has highest exergy loss, even though more mass in plastic
- As comparison, a 100-W system used for 6 hrs/day will consume about 440 kWh over a 2-yr period

Reference: Hannemann, Carey – UC Berkeley, 2007

What next?

“More research is needed”

In the next few slides, I’ll discuss:

- Benchmarking energy efficiency
- Accounting for costs of sustainable energy
- What kinds of measurements we need
- Things I left out, for lack of time

Benchmarking networking energy efficiency

Progress is seldom made without good benchmarks:

- Easy-to-understand results, even if oversimplified
- Agreed upon by majority of vendors + users
- Will need good metric of “useful work”
- Will need anti-cheating rules

Other application energy-efficiency benchmarks:

- JouleSort: Records sorted / Joule
  - Rivoire, Shah, Ranganathan, Kozyrakis, SIGMOD ’07
  - Paper shows way to optimize system hardware for energy efficiency
- SPECpower: server-side Java, various usage levels

How to design a benchmark of networking energy efficiency?

Define an application-level "goodput" metric

- E.g., Fully-rendered pages; VoIP sessions
- May require some QoS level (e.g., dropouts/sec)

Define how to measure energy use

- What energy counts against "networking"?
- How "end-to-end" could this actually be?

Think really hard about the "anti-cheating" rules

- Is caching/prefetching "cheating"?
- Is lousy QoS cheating?

How would IT change if electricity costs reflected full sustainability?

Simply "carbon-neutral" isn't good enough; even zero-carbon generators can have unsustainable externalities (e.g., big hydro-electric)

- Wide range of "social cost" estimates: $5--$150/ton of C
- Electricity cost estimates: 3% to 100% more per KWh

Thought experiments: if electricity costs doubled --

- Would PC buyers be willing to pay the costs?
- What HW changes would designers make?
- How would corp. IT managers change their ways?
- Would this drive better broadband bandwidths?

Measurement problems

It's very hard to get power vs. load info:

- Vendors typically specify power-provisioning #s
- Little info on power consumed by manageable components

I had a hard time getting WAN power numbers:

- ISPs are secretive
- Some of it must go into telco equipment and lines

Most systems don't provide "power introspection"

- Makes automated self-management harder

Things I left out of this talk

- Encryption
  - Can take a lot of energy
- Energy analysis of cell phones + handhelds
  - There are a lot of them
  - People have already done a lot of energy optimization
- Sensor nets
  - All sorts of good energy-related research
- Protocol (re-)designs for energy efficiency
  - Probably plenty of room for improvement

Summary

People are taking energy efficiency seriously

- Networking might seem like just a sliver, but not if you account for the complete end-to-end path
- The next few years should see some real improvements in how networking uses energy

Proportionality matters, because …

- Most computers are mostly idle
- It simplifies energy management
- It's a useful way to think about design goals

We still don't really understand how to measure this stuff

- We really need better metrics + benchmarks

Thanks to the HP people who helped me

- Eric Anderson
- Sujata Banerjee
- Hans Boehm
- Cullen Bash
- Christopher Hoover
- Kevin Leigh
- Moray McClaren
- Partha Ranganathan
- Puneet Sharma
- Amip Shah
- Mehul Shah

We have openings for summer interns

- Try www.hpl.hp.com/jobs in January
- Or contact me if that doesn't seem to work

Questions?