Doing Nothing Well – the essence of Green computing

2nd International Green Computing Conference

David E. Culler
University of California, Berkeley
July 26, 2011
Energy …

**Energy Consumption in the United States 1949 - 2005**

- **Avoided Supply = 70 Quads in 2005**
- **If E/GDP had dropped 0.4% per year**
- **Actual (E/GDP drops 2.1% per year)**
- **New Physical Supply = 25 Q**
- **70 Quads per year saved or avoided corresponds to 1 Billion cars off the road**

- **1949**
- **1973**
- **2005**

- **$1.7 Trillion**
- **$1.6**

- **1990**
- **50%**
- **80%**

- **2030**
- **2050**

Source: LBNL
We’ve been ‘doing something’ very well
with more and more to do it with …
in more and more ways …

Bell’s Law: new computer class per 10 years

Computers Per Person

1:10^6

1:10^3

1:1

10^3:1

years

1:10^6 Mainframe

1:10^3 Mini

1:1 Workstation

10^3:1 PC

years

Laptop

PDA

Cell

Mote!
for more and more people ...
and times they are a changing

Global temperature change (relative to pre-industrial era)

<table>
<thead>
<tr>
<th>Temperature</th>
<th>Food</th>
<th>Water</th>
<th>Ecosystems</th>
<th>Weather</th>
<th>Feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td>0°C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Today</td>
</tr>
<tr>
<td>1°C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2°C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3°C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4°C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5°C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Food**: Crop yields fall
- **Water**: Glaciers melt, Water shortages, Rising seas
- **Ecosystems**: Reefs damaged, Species extinction
- **Weather**: Storms, droughts, fires, heat waves
- **Feedback**: Abrupt climate change

Global temperature changes and their impacts on various aspects of the environment. The chart illustrates the changes that occur at different temperature levels, starting from 0°C (current climate) to 5°C (projected future temperatures). The impacts include changes in food, water, ecosystems, weather, and feedback mechanisms that accelerate climate change. The chart highlights the urgency of addressing climate change, with current conditions leading to significant environmental disruptions and future scenarios that could be catastrophic without intervention.

Graph inset: A graph showing the relationship between global temperature and carbon dioxide levels, indicating a correlation between rising temperatures and increased CO2 emissions.
Information Technology

- Moderate Part of the problem (2-4%)

- Big part of the solution

820m tons CO₂

2007 Worldwide IT carbon footprint: 2% = 830 m tons CO₂
Comparable to the global aviation industry

Expected to grow to 4% by 2020

360m tons CO₂

260m tons CO₂

Total emissions: 1.43bn tonnes CO₂ equivalent
‘Doing Nothing’ well
The Industrial Age Grid

Baseline + Dispatchable Tiers

Oblivious Loads

Generation ➔ Transmission ➔ Distribution ➔ Demand
Towards an “Aware” Energy Infrastructure

Baseline + Dispatchable Tiers

Non-Dispatchable Sources

Communication

Transmission

Generation

Distribution

Demand

Aware Interactive Loads

Oblivious Loads

LoCal
Limits to Renewable Penetration

- Variability, Intermittency of Supply
- Visibility into Availability of Supply
- Ability of Loads to Adapt
- Algorithms and Techniques for Reactive Load Adaptation

- Capability of the Infrastructure to maintain the match
A Cooperative Grid

- Availability
- Pricing
- Planning

Source

Intelligent Power Switch

energy subnet

Load IPS

Monitor, Model, Mitigate
- Deep instrumentation
- Waste elimination
- Efficient Operation
- Shifting, Scheduling, Adaptation

- Forecasting
- Tracking
- Market
Where to Start?

Buildings

- 72% of electrical consumption (US),
- 40-50% of total consumption,
- 42% of GHG footprint
- US commercial building consumption doubled 1980-2000, 1.5x more by 2025 [NREL]

Where Coal is used

Prime target of opportunity for renewable supplies
Our Building

- Lighting
- Servers
- PDUs
- CRACs

HVAC
IT and Plug Load
PDUs, CRACs
Servers

Annual Consumption

- kWh vs. kWh/sq ft
- National average
- LSA
- Cory
- Stanley
- McConne
- Soda
- Koshland

Soda Hall Power Consumption: 494 KW

Custom period: 2008-01-18 22:00 - 2009-01-24 22:45
Zoom: 1D 1'V 2'V MAX

Green Comp 2011
Doing nothing poorly
Our Other Building

- MCL equip
- MCL infra
- MCL vac
- servers
- Whole Bldg
- Central vent
- office HVAC
- DOP HVAC
- inst Lab 199 HVAC
- Lighting
- Parking Lot
- 7/26/11
- Green Comp 2011
Guess this Load …

Refrigerator
Guess this load…

HP desktop Workstation

Duty Cycle

Idle Power

Active Power
Guess this Load …

Dell Desktop + monitor

Green Comp 2011  7/26/11
Guess this Load … Laptop + monitor

Monitor “unplugged” screensaver

Windows “turns off” monitor
Guess this Load …

HP B&W laser printer
Guess this Load …

HP color laser printer
Guess this Load …

Coffee maker
Guess this Load ...

Polycom Phone
Doing Nothing
Brawny or Scrawny

Requests

LBS

Power

Max Request Rate
Active Power (W)
J/resp

Nehalem Xeon X5550
Atom 330
Atom Z530
Atom N450
ARM OMAP3530

Requests

Power

Green Comp 2011 7/26/11

28
# Something or Nothing

<table>
<thead>
<tr>
<th>Platform</th>
<th>Max Request Rate</th>
<th>J/resp</th>
<th>Active Power (W)</th>
<th>Idle Power (W)</th>
<th>Idle / Active</th>
<th>Power in S3</th>
<th>Sleep / Active</th>
<th>Wake Time from S3</th>
<th>Wake Time from Off</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nehalem Xeon X5550</td>
<td>310</td>
<td>0.6</td>
<td>248</td>
<td>149</td>
<td>60%</td>
<td>-</td>
<td>NA</td>
<td>-</td>
<td>60</td>
</tr>
<tr>
<td>Atom 330</td>
<td>28</td>
<td>1</td>
<td>28</td>
<td>22</td>
<td>79%</td>
<td>1</td>
<td>4%</td>
<td>2.4</td>
<td>40</td>
</tr>
<tr>
<td>Atom Z530</td>
<td>15</td>
<td>0.6</td>
<td>9.5</td>
<td>7.7</td>
<td>81%</td>
<td>1.7</td>
<td>18%</td>
<td>0.8</td>
<td>35</td>
</tr>
<tr>
<td>Atom N450</td>
<td>15</td>
<td>0.9</td>
<td>12.8</td>
<td>9.7</td>
<td>76%</td>
<td>1.9</td>
<td>15%</td>
<td>5.2</td>
<td>45</td>
</tr>
<tr>
<td>ARM OMAP3530</td>
<td>3</td>
<td>1.1</td>
<td>3.4</td>
<td>3.4</td>
<td>100%</td>
<td>-</td>
<td>NA</td>
<td>-</td>
<td>33</td>
</tr>
</tbody>
</table>
Where does the Power go?

- The Glue, not the processor
Reduced threshold devices tend to lower the optimal voltage; however, as seen in Section 3.3, at thresholds below 0.2V, power dissipation due to the subthreshold current will soon start to dominate and limit further power improvement.
DVFS really

Dynamic Frequency-Voltage Scaling

Power (W)

Performance (ips)

Clock Frequency (MHz)

0

50

100

150

200

250

1E+10

2E+10

5E+09

1E+10

1.2E+10

1.4E+10

1000

1500

2000

2500

3000

3500

Xeon Core i7 extreme

Power vs Performance

Power (W)

Power Density (nw per ips)

Performance (Inst per sec)

Green Comp 2011 7/26/11
Remember …

- Energy = Power x Time
## Something or Nothing

<table>
<thead>
<tr>
<th>Platform</th>
<th>Max Request Rate</th>
<th>J/resp</th>
<th>Active Power (W)</th>
<th>Idle Power (W)</th>
<th>Idle / Active</th>
<th>Power in S3</th>
<th>Sleep / Active</th>
<th>Wake Time from S3</th>
<th>Wake Time from Off</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nehalem Xeon X5550</td>
<td>310</td>
<td>0.6</td>
<td>248</td>
<td>149</td>
<td>60%</td>
<td>-</td>
<td>NA</td>
<td>-</td>
<td>60</td>
</tr>
<tr>
<td>Atom 330</td>
<td>28</td>
<td>1</td>
<td>28</td>
<td>22</td>
<td>79%</td>
<td>1</td>
<td>4%</td>
<td>2.4</td>
<td>40</td>
</tr>
<tr>
<td>Atom Z530</td>
<td>15</td>
<td>0.6</td>
<td>9.5</td>
<td>7.7</td>
<td>81%</td>
<td>1.7</td>
<td>18%</td>
<td>0.8</td>
<td>35</td>
</tr>
<tr>
<td>Atom N450</td>
<td>15</td>
<td>0.9</td>
<td>12.8</td>
<td>9.7</td>
<td>76%</td>
<td>1.9</td>
<td>15%</td>
<td>5.2</td>
<td>45</td>
</tr>
<tr>
<td>ARM OMAP3530</td>
<td>3</td>
<td>1.1</td>
<td>3.4</td>
<td>3.4</td>
<td>100%</td>
<td>-</td>
<td>NA</td>
<td>-</td>
<td>33</td>
</tr>
</tbody>
</table>
Engineering 101

- Design, Plan, Size, and Test at **Full Load**
  - Performance measured at full Load
- Add headroom and safety margin
- Operate at **Partial Load**
Power Proportionality

![Power Proportionality Graph](image)

Consumption vs. Productivity
Power Proportionality

Productivity

Consumption

Power Proportionality (idle = 50%)

Load

0% 20% 40% 60% 80% 100%

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

Efficiency

Power

Green Comp 2011  7/26/11
Doing less better and better

- http://www.spec.org/power_ssj2008
Opportunity in Doing Nothing Well?
A Day in the life of Wikipedia

- $P_{\text{cap}} = 70,000 \sim 1.5 \times P_{\text{max}}$
- $P_{\text{max}} = 46,591$
- $P_{\text{ave}} = 22,939$
The Opportunity

\[
\frac{\text{Energy non-PP}}{\text{Energy ideal PP}} = C \times \frac{\text{Peak}}{\text{Ave}}
\]
MapReduce Processing @ …

![Graph showing the number of jobs over time.](image-url)
Facebook data analytics

![Graph showing bandwidth usage over time with peak, waste factor, and average levels indicated.]
NERSC
112 node (576 core) cluster
~1 month trace, 128,915 jobs

35% avg. utilization

Time

11/21 00:00 11/28 00:00 12/05 00:00 12/12 00:00 12/19 00:00

Jobs
Storage

Enterprise NFS traces

MSR file server traces
Engineering 101 102

- Design, Plan, Size, and Test at **Full Load**
  - Performance measured at full Load
- Add headroom and safety margin
- Operate at **Partial Load**

- Repeat, designing for efficiency at **Typical Load**
PP Systems of non-PP pieces

Computational “Spinning Reserve”

Request Rate

Load Distribution
Scheduling
Power Management

Power

IPS

7/26/11

Green Comp 2011
Power/Performance Sweet Spot

![Graph showing Power/Performance Sweet Spot](image)

- **Green Comp 2011**
- **7/23/11**
Prediction and Penalties

![Graph showing prediction and penalties](image)

- **RMS Error (normalized to mean arrival rate)**
  - EWA ($\alpha = 0.85$)
  - EWA ($\alpha = 0.95$)
  - Last Arrival
  - MWA ($\alpha = 4$)
  - MWA ($\alpha = 15$)

- **Efficiency (Joules per request)**
  - 1 hr interval
  - 10 min interval
  - 5 min interval
  - 1 min interval
  - 1 sec interval

- **Violations (% of total requests)**

- **Prediction Time (s)**

Green Comp 2011  7/26/11
Power-Proportional Buildings?

Stanley Hall: Office + BioScience - 13 NMRs

Min = 72% of Max
Power-Proportional Buildings?

LeConte Hall: Office

Min = 31% of Max

202 KW

62 KW
Power Proportional Buildings?

50 Ton Chiller

200 Ton Chiller

10 months

2 months

Scott McNally Bldg Manager

Green Comp 2011 7/26/11
Power Proportional Buildings?

Observations on a Campus in Beijing

- Part-time
- Part-space
- Natural venting
- Light

Average: 30
Average: 64
Average: 82

Key reason:
Part time
Part space
Open windows

Yi Jiang
Building Energy Research Centre
Tsinghua University P R China

Case study: AC in Residential buildings

Measured energy consumption
in every units of a residential
building in Beijing, 2006, split unit

Centre AC for residential buildings in Beijing: 19.8 kWh/m².a
Towards Supply-Following Loads

SUPPLIES:
- provide power
- communicate renewable availability, price

LOADS:
- adapt demand
- communicate forecast

Internet

Grid

- electricity
- information
Growing proportion of renewables leads to higher price volatility. October 2008 to March 2010: >90 hours with negative prices; highest price reached: +€500/MWh, lowest -€500/MWh

Source: EEX spot prices.
... and @ CA

Energy Price

- Real-Time Market Price (5 min)
- Day-Ahead Market Price (1 hr)
- Now

Avg(24hr): 72.93
Avg(24hr): 45.53
Renewable Integration

Non-dispatchable, variable supply

Pacheco wind farm

Power proportional, grid-aware loads

Scientific computing cluster

NREL Western Wind and Solar Integration Study Dataset
http://wind.nrel.gov/Web_nrel/
Wind power over 48 hours from a wind farm in Monterrey County, California.

Variation in wind power for month long intervals at multiple wind farms.
Supply-Following Computational Loads

Background Processing (shiftable)

QoS (fidelity & latency)

Controllable Storage

Requests
Availability
Forecasts

Power
Energy-Agile Cluster Architecture
Grid-Aware Interactive Web Service

- **example goal:** maintain delivery cost despite Elect. price spikes
- **method:** distill content based on price & let PP workload mgr govern demand
Slack in Compute Load

slack = max run time – job duration
Cluster: NERSC Franklin
Average duration: 98 min
Average slack: 68 min

Cluster: EECS PSI
Average duration: 55 min
Average slack: 17 hours
Grid-Aware Scheduler

Run immediately, grid-oblivious scheduler

Greedy, grid-aware scheduler
Low Renewable Penetration

Most of the available wind is used
All algorithms perform comparably

17.68% wind energy
20.63% wind energy
Wind Scale Impact

- Max Wind Energy
- Greedy
- Foresight
- No Scheduling

Percent Cluster Energy from Wind (%) vs. Wind Scale (E Wind / E Cluster)

Green Comp 2011  7/26/11
Gusts and Bursts

63.43% wind energy

81.55% wind energy
Wind Characteristics

Coefficient of variation in wind energy availability ($\sigma/\mu$)

Grid Dependence (%)

- Grid Oblivious Scheduling
- Grid Aware Scheduling

Cluster Energy (MWh)

- Wind Energy
- Grid Energy

Non-PP
- PP Grid Oblivious
- PP Grid Aware

No Wind
- 100%
- 100%

Med Wind
- 69%
- 54%
- 30%

High Wind
- 60%
- 47%
- 19%

Green Comp 2011 7/26/11
Job Characteristics

![Graph showing Grid Dependence in percentage against Avg. Slack in hours. The graph compares Grid Oblivious and Grid Aware systems.](attachment:graph.png)
Scheduling vs Energy Storage

Grid Dependence (%)

Storage

- Grid Oblivious Scheduling
- Grid Aware Scheduling
Data Storage?
Why Replication?

- Why Replication?
  - data set size (B)
  - bandwidth requirements (B/s)

- NAS / NFS
  - Web farms (Wikipedia)
  - Data analytics, DFS (Hadoop)

- Disk arrays
  - Web farms (Wikipedia)
  - Data analytics, DFS (Hadoop)
Wikipedia 1-day trace

- Replicated thrice
- Partitioned over 5 disks
MSR-Cambridge Web/SQL server 7
day trace

- Not replicated
- 4 volumes
- 10 disks each
Sleepy Hierarchy

REQUESTS

CACHE
(<1 ns)

MEMORY DRAM
(~100 ns)

ACTIVE DISK
(~10 ms)

IN-ACTIVE DISK
(~5 sec)

\[ T_{avg} = T_{cache} + P_{miss,cache} \cdot T_{memory} + P_{miss,cache} \cdot P_{miss,memory} \cdot T_{disk} + P_{miss,cache} \cdot P_{miss,memory} \cdot P_{miss,active-disk} \cdot T_{inactive-disk} \]

Penalty : 100

Penalty : 1000

\[ P_{miss,cache} \]

\[ P_{miss,memory} \]
Techniques

1. Put disks to sleep after a loiter time, and wake-up on demand
2. Caching
3. Prefetching disk blocks
4. Dynamic object creation/Intelligent layout

\[ T_{avg} \approx T_{active-disk} + P_{miss,active-disk} \cdot T_{non-active-disk} \]
Active-disk misses
Stages of Energy Effectiveness

- **Waste Not**
  - Do Nothing Well !!!

- **Power Proportionality**
  - Peak Performance: Power => Safety
  - Optimize Partial Load - from nothing to peak

- **Sculpting**
  - Identify the energy *slack* and utilize it

- **Negotiated Grid / Load / Human Interaction**
  - Plan, Forecast, Negotiate, Manage
Why we don’t do it today…

- ... didn’t buy ‘em to turn ‘em off
- ... only a small fraction of the cost of the data center is the utility bill
- ... rather raise the utilization by co-locating instances
- ... in rare cases may see some performance degradation
- ... might not wake up reliably (wake-on-lan)
- ... a sleeping machine might die on wake-up
- ... don’t get paid to serve to do nothing.
"Perhaps the most important and pervasive principle of computer design is to make the common case fast: In making a design trade-off, favor the frequent case over the infrequent case."

– John L. Hennessy and David A. Patterson, Computer Architecture – a Quantitative Approach
Thanks