Energy Efficient Computing: From Bits to Buildings

Horst D. Simon
Lawrence Berkeley National Laboratory
and EECS Dept., UC Berkeley
hdsimon@lbl.gov

The Salishan Conference on High-Speed Computing
April 29, 2009
Acknowledgements

A large number of individuals have contributed to energy efficiency in computing at Berkeley Lab, UC Berkeley, and to this presentation:

David Bailey (CRD), Michael Banda (CRD), Michael Bennett (ITD), Shoaib Kamil (CRD), Jonathan Koomey (Stanford), Randy Katz (EECS), Tsu Jae King (EECS), Chuck McParland (CRD), Bruce Nordman (EETD), Lenny Oliker (CRD), Ekow Otoo (CRD), Vern Paxson (UCB/ICSI/CRD), Doron Rotem (CRD), Dale Sartor (EETD), John Shalf (NERSC), Erich Strohmaier (CRD), Bill Tschudi (EETD), Howard Walter (NERSC), Michael Wehner (CRD), Kathy Yelick (NERSC/CRD) … and many others

Almost all Berkeley resources about energy efficiency are available at http://www.lbl.gov/CS/html/energy%20efficient%20computing.html
Energy “Spaghetti” Chart
Power has become an industry-wide issue for computing

Two interrelated issues:

- Building and infrastructure problem -- continued increase in demand for computing ("buildings")

- Computer technology problem -- no more power density scaling ("bits")
Why does saving energy matter?
Energy Consumption in the United States 1949 - 2005

Source: Art Rosenfeld, California Energy Commission,
http://www.energy.ca.gov/commission/commissioners/rosenfeld_docs/index.html
An Honest Question?

Does the HPC community really care about reducing the carbon footprint?

NO!
HPC Interests

• Energy efficiency in computer rooms
  – Spend more resources on computing than on infrastructure

• Energy efficient technology
  – Maintain performance growth and get things done that could not be done before
Khazzoom-Brookes Postulate

• Energy efficiency at the micro-level leads to higher energy consumption at the macro-level
  – cheaper energy increases use
  – increased energy efficiency leads to economic growth
  – increased efficiency in one bottleneck resource increases use of companion technologies

• HPC follows Khazzoom-Brookes
Energy and IT

• “Big IT” – all electronics
  – PCs / etc., consumer electronics, telephony
    • Residential, commercial, industrial
  – More than 200 TWh/year
  – $16 billion/year
    • Based on .08$/KWh
  – Nearly 150 million tons of CO₂ per year
    • Roughly equivalent to 30 million cars!

Numbers represent U.S. only

One central baseload power plant (about 7 TWh/yr)
... and IT electricity use is increasing

Available at: http://www.koomey.com/publications.html

Sources: IDC data for installed base, shipments, and most popular models, and manufacturer data on power use for individual server models. Total expenditures assume US industrial electricity prices (2006 dollars)

0.8% of estimated 2005 world electricity sales, $7.2B/year

1.2% of 2005 U.S. electricity sales, $2.7B/year

Cooling and auxiliary equipment
High-end servers
Mid-range servers
Volume servers

U.S.
World
Worldwide IT Carbon Footprint

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
2020 IT Carbon Footprint

“SMART 2020: Enabling the Low Carbon Economy in the Information Age”, The Climate Group

Fig. 2.3 The global footprint by subsector

Datacenters: Owned by single entity interested in reducing opex

billion tons CO₂
Power has become an industry-wide issue for computing

Two interrelated issues:

- **Building and infrastructure problem** -- continued increase in demand for computing ("buildings")
- **Computer technology problem** -- no more power density scaling ("bits")
Absolute Power Levels
The Problem

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

(Modified with permission)
Top Challenges to Clusters

- Facility issues power, cooling
- System management capability
- Complexity of parallel algorithms
- Application availability/maturity
- Complexity of purchase and deployment
- Interconnect latency
- Facility issues, space, density
- Supported data storage mechanisms
- Interconnect bandwidth
- I/O performance
- 3rd-party software costs
- Interconnect complexity
- Facility issues noise

n = 96
Responses

• Cloud
• Containerized data centers
• Large scale data “factories”
• Increased emphasis on computer room and building efficiency
Containerized Datacenter
Mechanical-Electrical Design

Inside Project Blackbox, racks of up to 38 servers apiece generate tremendous heat. A panel of fans in front of each rack forces warm exhaust air through a heat exchanger, which cools the air for the next rack (detail), and so on in a continuous loop.

**DESIGN SPECS**
- Dimensions: 8 x 8 x 20 feet
- Weight: 20,000 pounds
- Cooling water supply: 60 gallons per minute
- Computing capacity: 7 terabytes
- Data storage: 2 petabytes
Data Center Economic Reality (2006)

• June 2006 - Google begins building a new data center near the Columbia River on the border between Washington and Oregon
  – Because the location is “at the intersection of cheap electricity and readily accessible data networking”
    “Hiding in Plain Sight, Google Seeks More Power” by John Markoff, NYT, June 14, 2006

• Microsoft and Yahoo are building big data centers upstream in Wenatchee and Quincy, Wash.
  – To keep up with Google, which means they need cheap electricity and readily accessible data networking

Google Dalles Oregon Facility
68,680 Sq Ft Per Pod

Source: Levy and Snowhorn, Data Center Power Trends, February 18, 2008
Microsoft Quincy, Wash.
470,000 Sq Ft, 47MW!

Source: Levy and Snowhorn, Data Center Power Trends, February 18, 2008
Microsoft’s Chicago Modular Datacenter
The Million Server Datacenter

• 24000 sq. m housing 400 containers
  – Each container contains 2500 servers
  – Integrated computing, networking, power, cooling systems

• 300 MW supplied from two power substations situated on opposite sides of the datacenter

• Dual water-based cooling systems circulate cold water to containers, eliminating need for air conditioned rooms
Potential Benefits of Improved Data Center Energy Efficiency:

- 20-40% savings typically possible
- Aggressive strategies can yield better than 50% savings
- Extend life and capacity of existing data center infrastructures
- But is my center good or bad?
Energy benchmarking can allow comparison to peers and help identify best practices.

LBNL conducted studies of over 30 data centers:

- Found wide variation in performance
- Identified best practices
High Level Metric—
Data Center Infrastructure Efficiency (DCiE)
Ratio of Electricity Delivered to IT Equipment to Total Data Center Power

Average: 0.57

Higher is better
Using benchmark results to find best practices:

- Air management
- Right-sizing
- Central plant optimization
- Efficient air handling
- Liquid cooling
- Free cooling
- Humidity control
- Improve power chain
- On-site generation
- Design and M&O processes
UC’s Computational Research and Theory (CRT) Facility
Use Free Cooling:

• Water-side Economizers
  – No contamination question
  – Can be in series with chiller

• Outside-Air Economizers
  – Can be very effective (24/7 load)
  – Must consider humidity
System Design Approach:

- Air-Side Economizer (93% of hours)
- Direct Evaporative Cooling for Humidification/ pre-cooling
- Low Pressure-Drop Design (1.5” total static)

<table>
<thead>
<tr>
<th>Hours of Operation</th>
<th>Mode 1</th>
<th>Mode 2</th>
<th>Mode 3</th>
<th>Mode 4</th>
<th>Mode 5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>100% Economiser</td>
<td>2207 hrs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OA + RA</td>
<td>5957 hrs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Humidification</td>
<td></td>
<td>45 hrs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Humid + CH cooling</td>
<td></td>
<td></td>
<td>38 hrs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CH only</td>
<td></td>
<td></td>
<td></td>
<td>513 hrs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8760 hrs</td>
<td></td>
</tr>
</tbody>
</table>
Water Cooling: Four-pipe System

- Allows multiple temperature feeds at server locations through mixing of CHW & TRW
- Closed-loop treated cooling water from cooling towers (via heat exchanger)
- Chilled water from chillers
- Headers, valves and caps for modularity and future flexibility

Predicted CRT Performance

- DCIE of 0.95 based on annual energy
- DCIE of 0.88 based on peak power
Design Guidelines Are Available

- Design Guides were developed based upon the observed best practices
- Guides are available through PG&E and LBNL websites
- Self benchmarking protocol also available

http://hightech.lbl.gov/datacenters.html
Links to Get Started

DOE Website: Sign up to stay up to date on new developments
www.eere.energy.gov/datacenters

Lawrence Berkeley National Laboratory (LBNL)
http://hightech.lbl.gov/datacenters.html

LBNL Best Practices Guidelines (cooling, power, IT systems)
http://hightech.lbl.gov/datacenters-bpg.html

ASHRAE Data Center technical guidebooks
http://tc99.ashraetcs.org/

The Green Grid Association – White papers on metrics
http://www.thegreengrid.org/gg_content/

Energy Star® Program
http://www.energystar.gov/index.cfm?c=prod_development.server_efficiency

Uptime Institute white papers
www.uptimeinstitute.org
TALK TO DALE:
Join his network to share information and
Pull market towards higher efficiency products

Contact Information:
Dale Sartor, P.E.
Lawrence Berkeley National Applications Team
MS 90-3111
University of California
Berkeley, CA 94720

DASartor@LBL.gov
(510) 486-5988
http://Ateam.LBL.gov
Power consumption has become an industry-wide issue for computing.

Two interrelated issues:

- Building and infrastructure problem -- continued increase in demand for computing
- Computer technology problem -- no more power density scaling ("bits")
An Early Warning

• Presented by Shekhar Borkar in Berkeley in November 2000
Power will be a problem

Power delivery and dissipation will be prohibitive
Power density will increase

Power density too high to keep junctions at low temp
Traditional Sources of Performance Improvement are Flat-Lining (2004)

• New Constraints
  – 15 years of exponential clock rate growth has ended

• Moore’s Law reinterpreted:
  – How do we use all of those transistors to keep performance increasing at historical rates?
  – Industry Response: #cores per chip doubles every 18 months instead of clock frequency!
  – multicore

Figure courtesy of Kunle Olukotun, Lance Hammond, Herb Sutter, and Burton Smith
Estimated Exascale Power Requirements

- LBNL IJHPCA Study for ~1/5 Exaflop for Climate Science
  - Extrapolation of Blue Gene and AMD design trends
  - Estimate: 20 MW for BG and 179 MW for AMD

- DOE E3 Report
  - Extrapolation of existing design trends to exascale in 2016
  - Estimate: 130 MW

- DARPA Study
  - More detailed assessment of component technologies for exascale system
  - Estimate: more than 120 MW

- The current approach is not sustainable!
DARPA Exascale Study

- Commissioned by DARPA to explore the challenges for Exaflop computing
- Two model for future performance growth
  - Simplistic: ITRS roadmap; power for memory grows linear with #of chips; power for interconnect stays constant
  - Fully scaled: same as simplistic, but memory and router power grow with peak flops per chip
We won’t reach Exaflops with this approach
... and the power costs will still be staggering

From Peter Kogge, DARPA Exascale Study
## Extrapolating to Exaflop/s in 2018

<table>
<thead>
<tr>
<th></th>
<th>BlueGene/L (2005)</th>
<th>Exaflop Directly scaled</th>
<th>Exaflop compromise using expected technology</th>
<th>Assumption for “compromise guess”</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Node Peak Perf</strong></td>
<td>5.6GF</td>
<td>20TF</td>
<td>20TF</td>
<td>Same node count (64k)</td>
</tr>
<tr>
<td><strong>hardware concurrency/node</strong></td>
<td>2</td>
<td>8000</td>
<td>1600</td>
<td>Assume 3.5GHz</td>
</tr>
<tr>
<td><strong>System Power in Compute Chip</strong></td>
<td>1 MW</td>
<td>3.5 GW</td>
<td>35 MW</td>
<td>100x improvement (very optimistic)</td>
</tr>
<tr>
<td><strong>Link Bandwidth (Each unidirectional 3-D link)</strong></td>
<td>1.4Gbps</td>
<td>5 Tbps</td>
<td>1 Tbps</td>
<td>Not possible to maintain bandwidth ratio.</td>
</tr>
<tr>
<td><strong>Wires per unidirectional 3-D link</strong></td>
<td>2</td>
<td>400 wires</td>
<td>80 wires</td>
<td>Large wire count will eliminate high density and drive links onto cables where they are 100x more expensive. Assume 20 Gbps signaling</td>
</tr>
<tr>
<td><strong>Pins in network on node</strong></td>
<td>24 pins</td>
<td>5,000 pins</td>
<td>1,000 pins</td>
<td>20 Gbps differential assumed. 20 Gbps over copper will be limited to 12 inches. Will need optics for in rack interconnects. 10Gbps now possible in both copper and optics.</td>
</tr>
<tr>
<td><strong>Power in network</strong></td>
<td>100 KW</td>
<td>20 MW</td>
<td>4 MW</td>
<td>10 mW/Gbps assumed. Now: 25 mW/Gbps for long distance (greater than 2 feet on copper) for both ends one direction. 45mW/Gbps optics both ends one direction. + 15mW/Gbps of electrical Electrical power in future: separately optimized links for power.</td>
</tr>
<tr>
<td><strong>Memory Bandwidth/node</strong></td>
<td>5.6GB/s</td>
<td>20TB/s</td>
<td>1 TB/s</td>
<td>Not possible to maintain external bandwidth/Flop</td>
</tr>
<tr>
<td><strong>L2 cache/node</strong></td>
<td>4 MB</td>
<td>16 GB</td>
<td>500 MB</td>
<td>About 6-7 technology generations</td>
</tr>
<tr>
<td><strong>Data pins associated with memory/node</strong></td>
<td>128 data pins</td>
<td>40,000 pins</td>
<td>2000 pins</td>
<td>3.2 Gbps per pin</td>
</tr>
<tr>
<td><strong>Power in memory I/O (not DRAM)</strong></td>
<td>12.8 KW</td>
<td>80 MW</td>
<td>4 MW</td>
<td>10 mW/Gbps assumed. Most current power in address bus. Future probably about 15mW/Gbps maybe get to 10mW/Gbps (2.5mW/Gbps is c^v^2+f for random data on data pins) Address power is higher.</td>
</tr>
<tr>
<td><strong>QCD CG single iteration time</strong></td>
<td>2.3 msec</td>
<td>11 usec</td>
<td>15 usec</td>
<td>Requires: 1) fast global sum (2 per iteration) 2) hardware offload for messaging (Driverless messaging)</td>
</tr>
</tbody>
</table>

*Source: David Turek, IBM*
Power Efficiency related to Processors
Green Flash: 
Ultra-Efficient Climate Modeling

• Project by Shalf, Oliker, Wehner and others at LBNL

• An alternative route to exascale computing
  – Target specific machine designs to answer a scientific question
  – Use of new technologies driven by the consumer market.
Ultra-Efficient “Green Flash” Computing at NERSC: 100x over Business as Usual

Radically change HPC system development via application-driven hardware/software co-design

- Achieve 100x power efficiency and 100x capability of mainstream HPC approach for targeted high-impact applications
- Accelerate development cycle for exascale HPC systems
- Approach is applicable to numerous scientific applications
- Proposed pilot application: Ultra-high resolution climate change simulation
Path to Power Efficiency
Reducing Waste in Computing

• Examine methodology of low-power embedded computing market
  – optimized for low power, low cost and high computational efficiency

  “Years of research in low-power embedded computing have shown only one design technique to reduce power: reduce waste.”
  — Mark Horowitz, Stanford University & Rambus Inc.

• Sources of waste
  – Wasted transistors (surface area)
  – Wasted computation (useless work/speculation/stalls)
  – Wasted bandwidth (data movement)
  – Designing for serial performance
Design for Low Power: More Concurrency

- Cubic power improvement with lower clock rate due to V^2F
- Slower clock rates enable use of simpler cores
- Simpler cores use less area (lower leakage) and reduce cost
- Tailor design to application to reduce waste

This is how iPhones and MP3 players are designed to maximize battery life and minimize cost.
Low Power Design Principles

- IBM Power5 (server)
  - 120W@1900MHz
  - Baseline
- Intel Core2 sc (laptop):  
  - 15W@1000MHz
  - 4x more FLOPs/watt than baseline
- IBM PPC 450 (BG/P - low power)  
  - 0.625W@800MHz
  - 90x more
- Tensilica XTensa (Moto Razor):  
  - 0.09W@600MHz
  - 400x more

Even if each core operates at 1/3 to 1/10th efficiency of largest chip, you can pack 100s more cores onto a chip and consume 1/20 the power.
Customization Continuum: Green Flash

- **Application-driven does NOT necessitate a special purpose machine**
- **MD-Grape: Full custom ASIC design**
  - 1 Petaflop performance for one application using 260 kW for $9M
- **D.E. Shaw Anton System: Full and Semi-custom design**
  - Simulate 100x–1000x timescales vs any existing HPC system (~200kW)
- **Application-Driven Architecture (Green Flash): Semicustom design**
  - Highly programmable core architecture using C/C++/Fortran
  - Goal of 100x power efficiency improvement vs general HPC approach
  - Better understand how to build/buy application-driven systems
  - **Potential:** 1km-scale model (~200 Petaflops peak) running in O(5 years)
Green Flash Strawman System Design

We examined three different approaches (in 2008 technology)

Computation $0.015\times0.02\times100$: 10 PFlops sustained, ~200 PFlops peak

- **AMD Opteron**: Commodity approach, lower efficiency for scientific applications offset by cost efficiencies of mass market
- **BlueGene**: Generic embedded processor core and customize system-on-chip (SoC) to improve power efficiency for scientific applications
- **Tensilica XTensa**: Customized embedded CPU w/SoC provides further power efficiency benefits but maintains programmability

<table>
<thead>
<tr>
<th>Processor</th>
<th>Clock</th>
<th>Peak/ Core (Gflops)</th>
<th>Cores/ Socket</th>
<th>Sockets</th>
<th>Cores</th>
<th>Power</th>
<th>Cost 2008</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMD Opteron</td>
<td>2.8GHz</td>
<td>5.6</td>
<td>2</td>
<td>890K</td>
<td>1.7M</td>
<td>179 MW</td>
<td>$1B+</td>
</tr>
<tr>
<td>IBM BG/P</td>
<td>850MHz</td>
<td>3.4</td>
<td>4</td>
<td>740K</td>
<td>3.0M</td>
<td>20 MW</td>
<td>$1B+</td>
</tr>
<tr>
<td>Green Flash / Tensilica XTensa</td>
<td>650MHz</td>
<td>2.7</td>
<td>32</td>
<td>120K</td>
<td>4.0M</td>
<td>3 MW</td>
<td>$75M</td>
</tr>
</tbody>
</table>
Climate System Design Concept
Strawman Design Study

10PF sustained
~120 m²
<3MWatts
< $75M

32 boards
per rack

100 racks @
~25KW

32 chip + memory
clusters per board
(2.7 TFLOPS @ 700W)

VLIW CPU:
- 128b load-store + 2 DP MUL/ADD + integer op/ DMA per cycle
- Synthesizable at 650MHz in commodity 65nm
- 1mm² core, 1.8-2.8mm² with inst cache, data cache data RAM, DMA interface, 0.25mW/MHz
- Double precision SIMD FP: 4 ops/cycle (2.7GFLOPs)
- Vectorizing compiler, cycle-accurate simulator, debugger GUI (Existing part of Tensilica Tool Set)
- 8 channel DMA for streaming from on/off chip DRAM
- Nearest neighbor 2D communications grid

83 GFLOPS @ 7W

32 processors per 65nm chip

8 chan DMA

32K

64-128K D
2x128b

VLIW CPU:
- 128b load-store + 2 DP MUL/ADD + integer op/ DMA per cycle:
- Synthesizable at 650MHz in commodity 65nm
- 1mm² core, 1.8-2.8mm² with inst cache, data cache data RAM, DMA interface, 0.25mW/MHz
- Double precision SIMD FP: 4 ops/cycle (2.7GFLOPs)
- Vectorizing compiler, cycle-accurate simulator, debugger GUI (Existing part of Tensilica Tool Set)
- 8 channel DMA for streaming from on/off chip DRAM
- Nearest neighbor 2D communications grid
Green Flash Hardware Demo at SC08

- Demonstrated during SC ‘08
- Proof of concept
  - CSU atmospheric model ported to Tensilica Architecture
  - Single Tensilica processor running atmospheric model at 50MHz
- Emulation performance advantage
  - Processor running at 50MHz vs. Functional model at 100 kHz
  - 500x Speedup
- Actual code running - not representative benchmark
Silicon photonics enables optics to be integrated with conventional CMOS

- Enables up to 27x improvement in communication energy efficiency!
Summary

• Power consumption is a huge problem in HPC
  – “Bits”: we may not be able to scale to Exaflops without new technologies
  – “Buildings”: we may have to spend more $$ on infrastructure and less on computing
Extra Slides
Outline

1. Power consumption has become an industry-wide issue for computing

2. Building and computer room energy efficiency

3. Computer architecture for energy efficiency - the Green Flash project

4. Future
Processor Technology Trend

- 1990s - R&D computing hardware dominated by desktop/COTS
  - Had to learn how to use COTS technology for HPC
- 2010 - R&D investments moving rapidly to consumer electronics/ embedded processing
  - Must learn how to leverage embedded processor technology for future HPC systems
Consumer Electronics Convergence

Market in Japan (B$)
Consumer Electronics Convergence

Market in Japan (B$)

- PC
- TV + DVD + DSC
- Analog
- Digital
- DSC
- DVD

2001 2002 2003 2004 2005
Consumer Electronics has Replaced PCs as the Dominant Market Force in CPU Design!!
Consumer Electronics has Replaced PCs as the Dominant Market Force in CPU Design!!

Brief History of PC
1975: Altair/ MITS
1978: Apple II
1981: IBM PC (MSDOS+i8088)
1985: Windows 1.0

Source: IDC

IBM Sold PC Business to Lenovo

Revenue

Shipment (Units)

Revenue

Units

1000M

100M

10M

1B

10B

100B

1000B


Source: IDC
Consumer Electronics has Replaced PCs as the Dominant Market Force in CPU Design!!

- Apple Introduces iPod
- iPod+iTunes exceeds 50% of Apple’s Net Profit
- Apple Introduces Cell Phone (iPhone)
Consumer Electronics has Replaced PCs as the Dominant Market Force in CPU Design!!

Apple

Introduces

IPod

IPod+iTunes exceeds 50% of Apple’s Net Profit

Apple Introduces Cell Phone (iPhone)
Power fundamentals 2018--2020

Processor budget: **15 MW** for a sustained HPL Exaflops (10pJ/op)  \{250\}

Memory budget: **25– 50 MW** (25 pJ/op)  \{300\}  
[1/2 Byte/sec/Flops]

Interconnect budget: **50 MW** (5 pJ/op)  \{0.1 B/F\}  
\{30\}

I/O Budget: **5 MW** (5 pJ/byte) 1 petabyte/sec

Power and Cooling Budget @30%: **30 MW**

**Total Power required** **125 MW!**
Power Ranking and How Not to do it!

- To rank objects by “size” one needs extensive properties:
  - Weight or Volume
  - Rmax (TOP500)
    - A ‘larger’ system should have a larger Rmax.
- The ratio of 2 extensive properties is an intensive one:
  - \(\frac{\text{weight}}{\text{volume}} = \text{density}\)
  - \(\frac{\text{Performance}}{\text{Power Consumption}} = \text{Power efficiency}\)
- One can-not ‘rank’ objects with densities BY SIZE:
  - Density does not tell anything about size of an object
  - A piece of lead is not heavier or larger than one piece of wood.
- Linpack (sub-linear) / Power (linear) will always sort smaller systems before larger ones!

![Graph showing Rmax/Power relationship for TOP500 systems.](image)
The Transition to Low-Power Technology is Inevitable

Does it make sense to build systems that require the electric power equivalent of an aluminum smelter?

• Information “factories” are only affordable for a few government labs and large commercial companies (Google, MSN, Yahoo …)
  – Midrange installations will soon hit the 1 - 2 MW wall, requiring costly new installations
  – Economics will change if operating expenses of a server exceed acquisition cost
• The industry will switch to low-power technology within 2 - 3 years
• Embedded processors or game processors will be the next step (BG, Cell, Nvidia, SiCortex, Tensilica)
  – Example RR, first Petaflops system
Power Efficiency related to Processors

Power Efficiencies of Systems with different Processors

- PowerXCell 8i
- PowerPC 450
- Intel EM64T Xeon L54xx (Harpertown)
- Intel EM64T Xeon E54xx (Harpertown)
- PowerPC 440
- Intel EM64T Xeon S3xx (Clovertown)
- AMD x86_64 Opteron Quad Core
- Intel EM64T Xeon S1xx (Woodcrest)
- AMD x86_64 Opteron Dual Core

Legend:
- (8+1) core
- QC embedded
- Quadcore
- DC embedded
- Dualcore
Frequencies and Power Efficiency

Power rating is 80 Watts each!

Maximum Power Efficiency of Harpertown E54xx

- 2500: 138
- 2666: 147
- 2800: 213
- 2830: 213
- 3000: 240

31st List / June 2008
Most Power Efficient Systems

Power Efficiencies of different Systems

- BladeCenter OS22 Cluster
- BlueGene/L
- SGI Altix ICE 8200
- Cluster Platform 3000 BL 2x220
- BlueGene/L
- PowerEdge 1950 QC
- Cray XT4 QuadCore
- BladeCenter HS21 Cluster, Xeon quad core
- Bladecenter HS21 Cluster, Xeon quad core
- PowerEdge 1950 QC
- Cray XT4 QuadCore
- BladeCenter HS21 Cluster, Xeon dual core
- pSeries 575
- Cluster Platform 3000 BL 460c
- Cray XT4
- BLadecenter LS21 Cluster
- PowerEdge 1950 DC
- xSeries x3550 Cluster Xeon dual core
- xSeries x3650 Cluster Xeon dual core

Legend:
- (8+1) core
- Embedded
- Quadcore
- Dualcore
Convergence of Platforms

- Multiple parallel general-purpose processors (GPPs)
- Multiple application-specific processors (ASPs)

"The Processor is the new Transistor" [Rowen]

Intel 4004 (1971):
- 4-bit processor,
- 2312 transistors,
- ~100 KIPS,
- 10 micron PMOS,
- 11 mm² chip

Sun Niagara
- 8 GPP cores (32 threads)

Intel Network Processor
- 1 GPP Core
- 16 ASPs (128 threads)

IBM Cell
- 1 GPP (2 threads)
- 8 ASPs

Picochip DSP
- 1 GPP core
- 248 ASPs

Cisco CRS-1
- 188 Tensilica GPPs

"1000s of processor cores per die"
BG/L—the Rise of the Embedded Processor

TOP 500 Performance by Architecture

Aggregate $R_{\text{max}}$ (Tflop/s)

- MPP
- SMP
- Cluster
- Constellations
- Single Processor
- SIMD
- Others
- MPP embedded

06/1993 to 06/2005
Summary (1)

- LBNL has taken a comprehensive approach to the power in computing problem
  - Component level (investigate use of low-power components and build new system)
  - System level (measuring and understanding energy consumption of system)
  - Computer Room level (understand airflow and cooling technology)
  - Building Level (enforce rigorous energy standards in new computer building and use of innovative energy savings technology)
Summary (2)

• Economic factors are driving us already to more energy efficient solutions in computing

• Incremental improvements are well on track, but we may ultimately need revolutionary new technology to reach the Exaflop/s level and beyond
Outline

1. Power consumption has become an industry-wide issue for computing

2. Building and computer room energy efficiency

3. Computer architecture for energy efficiency - the Green Flash project

4. Towards a better understanding of “green computing”
Focus on PUE

- PUE = “power usage effectiveness” metric promoted by “Green Grid”
- PUE = total facility power/ computer equipment power
- Reduce PUE by consistent application of facilities improvements

<table>
<thead>
<tr>
<th></th>
<th>PUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current Trends</td>
<td>1.9</td>
</tr>
<tr>
<td>Improved Operations</td>
<td>1.7</td>
</tr>
<tr>
<td>Best Practices</td>
<td>1.3</td>
</tr>
<tr>
<td>State-of-the-Art</td>
<td>1.2</td>
</tr>
</tbody>
</table>