Confessions of an Accidental Greenie:  
*From Green Destiny to the Green500*

Wu FENG, wfeng@vt.edu
Dept. of Computer Science
Dept. of Electrical & Computer Engineering
Dept. of Biomedical Engineering and Mechanics
Health Sciences
Biocomplexity Institute

5th Workshop on Energy-Efficient Supercomputing
Decades of focus on performance, performance, performance (and price/performance).
Importance of High-Performance Computing (HPC)

Competitive Risk From Not Having Access to HPC

- Could exist and compete: 3%
- Could not exist as a business: 16%
- Could not compete on quality & testing issues: 34%
- Could not compete on time to market & cost: 47%

Data from Council of Competitiveness. Sponsored Survey Conducted by IDC

Only 3% of companies could exist and compete without HPC.

- 200+ participating companies, including many Fortune 500 (Proctor & Gamble and biological and chemical companies)
The Focus on Performance:
The TOP500 List
A ranking of the fastest 500 supercomputers in the world

• Benchmark
  – LINPACK: Solves a (random) dense system of linear equations in double-precision (64 bits) arithmetic.

• Evaluation Metric
  – Performance (i.e., Speed)
    • Floating-Operations Per Second (FLOPS)

• Web Site
  – http://www.top500.org/
The Focus on Performance: Gordon Bell Awards

• Metrics for Evaluating Supercomputers
  – Performance (i.e., Speed)
    • Metric: Floating-Operations Per Second (FLOPS)
  – Price/Performance ➔ Cost Efficiency
    • Metric: Acquisition Cost / FLOPS

• Performance & price/performance are important metrics, but …
The “Cost” of Achieving Performance

- Electrical power costs $$$$.

In 2001, the combined annual cost of “Infrastructure & Energy” for a 1U server surpassed its purchase cost.

In 2004, the annual “Infrastructure” cost for a 1U server in a data center surpassed its purchase cost.

In 2008, the annual “Energy” cost for a 1U server in a data center surpassed its purchase cost.
The “Cost” of Achieving Performance

- Too much power can affect efficiency, reliability, and availability.
    - **Winter**: “Machine Room” Temperature of 70-75° F
      - Failure approximately once per week.
    - **Summer**: “Machine Room” Temperature of 85-90° F
      - Failure approximately twice per week.
  - Arrhenius’ Equation (applied to microelectronics)
    (circa 1890s in chemistry → circa 1980s in computer & defense industries)
    - For every 10° C (18° F) increase in temperature, … the failure rate of a system doubles. *


“Necessity is the mother of invention.”

synergy.cs.vt.edu
Yet in 2002 …

• “In HPC, no one cares about power & cooling, and no one ever will …”

• “Moore’s Law for Power will stimulate the economy by creating a new market in cooling technologies.”
Moore’s Law for Power

Chip Maximum
Power in watts/cm²

1000
100
10
1

1985 1995 2001 Year

1.5µ 1µ 0.7µ 0.5µ 0.35µ 0.25µ 0.18µ 0.13µ 0.1µ 0.07µ

I386 – 1 watt
I486 – 2 watts
Pentium – 14 watts
Pentium Pro – 30 watts
Pentium II – 35 watts
Pentium III – 35 watts
Pentium 4 – 75 watts
Itanium – 130 watts

Not too long to reach Nuclear Reactor

Surpassed
Heating Plate

Source: Fred Pollack, Intel. New Microprocessor Challenges in the Coming Generations of CMOS Technologies, MICRO32 and Transmeta
Moore’s Law for Power

Can we build a low-power supercomputer that is efficient, reliable, and highly available but is still considered high performance?

Source: Fred Pollack, Intel. New Microprocessor Challenges in the Coming Generations of CMOS Technologies, MICRO32 and Transmeta
Demo at ACM/IEEE SC 2001

MetaBlade: 24 ServerBlade 633s

MetaBlade2: 24 ServerBlade 800s
(On-loan from RLX for SC 2001)

- MetaBlade Node
  - 633-MHz Transmeta TM5600
  - 512-KB cache, 256-MB RAM
  - 100-MHz front-side bus
  - 3 x 100-Mb/s Ethernet

- MetaBlade2 Node
  - 800-MHz Transmeta TM5800
  - 512-KB cache, 384-MB RAM
    (128-MB on-board DDR + 256-MB SDR DIMM)
  - 133-MHz front-side bus
  - 3 x 100-Mb/s Ethernet

Performance of an N-body Simulation of Galaxy Formation
- MetaBlade: 2.1 Gflops; MetaBlade2: 3.3 Gflops

No unscheduled failures in its lifetime despite no cooling facilities.
RLX System™ 324: A Web-Hosting Server!

- 3U vertical space
  - 5.25” x 17.25” x 25.2”
- Two hot-pluggable 450W power supplies
  - Load balancing
  - Auto-sensing fault tolerance
- System midplane
  - Integration of system power, management, and network signals
  - Elimination of internal system cables
  - Enabling hot-pluggable blades
- Network cards
  - Hub-based management
  - Two 24-port interfaces

RLX System™ 300ex
- Interchangeable blades
  - Intel, Transmeta, or both.
- Switched-based management
**RLX ServerBlade™ 633**

**Transmeta™ TM5600 633 MHz**

- **Public NIC**
  - 33 MHz PCI

- **Private NIC**
  - 33 MHz PCI

- **Management NIC**
  - 33 MHz PCI

- **Code Morphing Software (CMS), 1 MB**

- **512KB Flash ROM**

- **128MB, 256MB, 512MB DIMM SDRAM PC-133**

- **ATA 66**
  - 0 or 1 or 2 - 2.5” HDD
  - 10 or 30 GB each

- **Status LEDs**

- **Serial RJ-45 debug port**

- **Reset Switch**

- **128KB L1 cache, 512KB L2 cache**

- **RLX ServerBlade™**
  - **667 @ $960**
  - **933 @ TBD**
  - **1066 @ alpha**

- **SySeRGI™**

- **Supercomputing in Small Spaces**

- **Synergy.cs.vt.edu**

( Circa 2001 )
Transmeta TM5600 CPU: VLIW + CMS

- **VLIW Engine**
  - Up to four-way issue
    - In-order execution only.
  - Two integer units
  - Floating-point unit
  - Memory unit
  - Branch unit

- **VLIW Transistor Count (“Anti-Moore’s Law”)**
  - 25% of Intel PIII → 7x less power consumption
  - Less power → lower “on-die” temp. → better reliability & availability
Moore’s Law for Power

Chip Maximum Power in watts/cm²

Not too long to reach Nuclear Reactor

Surpassed Heating Plate

Pentium – 14 watts

I386 – 1 watt

I486 – 2 watts

Pentium II – 35 watts

Pentium Pro – 30 watts

Pentium III – 35 watts

Pentium 4 – 75 watts

Itanium – 130 watts

Year

1985 1995 2001

1.5μ 1μ 0.7μ 0.5μ 0.35μ 0.25μ 0.18μ 0.13μ 0.1μ 0.07μ

Source: Fred Pollack, Intel. New Microprocessor Challenges in the Coming Generations of CMOS Technologies, MICRO32 and Transmeta
Outline

• Motivation & Background

• Supercomputing in Small Spaces (http://sss.cs.vt.edu/)
  – Origin: Green Destiny
  – Evolution
    • Architectural: Orion Multisystems → Sun Microsystems
    • Software: ENERGYFIT™ Power-Aware Run-Time OS Daemon

• The Future?

• The Green500 List: Past, Present, and Future

• Conclusion
Supercomputing in Small Spaces
Efficiency, Reliability, and Availability via Green HPC
(Started in 2001 at Los Alamos Nat’l Lab. Now at Virginia Tech.)

• Goal: Improve efficiency, reliability, and availability (ERA) in large-scale supercomputing systems.

• Analogy: Today’s Supercomputer vs. Supercomputing in Small Spaces
  – Formula One Race Car: Wins raw performance but reliability is so poor that it requires frequent maintenance. Throughput low.
  – Nissan 370Z: Loses raw performance but high reliability results in high throughput (i.e., miles driven → answers/month).

“Necessity is the mother of invention.”
Green Destiny Supercomputer
(circa December 2001 – February 2002)

• A 240-Node Cluster in Five Sq. Ft.

• Each Node
  – 1-GHz Transmeta TM5800 CPU w/ High-Performance Code-Morphing Software running Linux 2.4.x
    • CPU Power Consumption? **Only 6 watts!**
  – 640-MB RAM, 20-GB hard disk, 100-Mb/s Ethernet

• Total
  – 240 Gflops peak (**Linpack: 101 Gflops in March 2002.**)
  – **Power Consumption: Only 3.2 kW (diskless)**

• Reliability & Availability
  – **No unscheduled downtime in its 24-month lifetime**
    • Environment: A dusty 85°-90° F warehouse

Empirical Data on Temperature

• From off to system boot-up, after 25 seconds:

<table>
<thead>
<tr>
<th>Processor</th>
<th>Clock Freq.</th>
<th>Voltage</th>
<th>Peak Temp.*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel Pentium III-M</td>
<td>500 MHz</td>
<td>1.6 V</td>
<td>252° F (122° C)</td>
</tr>
<tr>
<td>Transmeta Crusoe TM5600</td>
<td>600 MHz</td>
<td>1.6 V</td>
<td>147° F (64° C)</td>
</tr>
</tbody>
</table>

*Peak temperature measured with no cooling.

• Arrehenius’ Equation
  – Every 10° C increase, doubles the failure rate.

Implication: Without cooling facilities, PIII-M is 32 times more likely to fail!
Yet in 2002 …

• “Green Destiny is so low power that it runs just as fast when it is unplugged.”

• “The slew of expletives and exclamations that followed Feng’s description of the system …”
Outline

• Motivation & Background

• Supercomputing in Small Spaces (http://sss.cs.vt.edu/)
  – Origin: Green Destiny
  – Evolution
    • Architectural: Orion Multisystems → Sun Microsystems
    • Software: ™ Power-Aware Run-Time OS Daemon

• The Future?

• The Green500 List: Past, Present, and Future

• Conclusion
The Start of a Controversial Movement (2002)

- **Traditional View**
  - Extend battery life in laptops, sensors, and embedded systems (such as PDAs, handhelds, and mobile phones)

- **Controversial View** (2002)
  - Potentially sacrifice a bit of performance to enhance efficiency, reliability, and availability in HPC systems
  - Gripe: HPC unwilling to “sacrifice” performance

- **The Start of a New Movement** (2004)
  - Feng et al. (Green Destiny II, 9/2003), Nakashima et al. (MegaScale: Green Destiny II, 9/2003); Cameron (U. South Carolina, 2004); Freeh (NCSU, 2004); and Lowenthal (U. Georgia, 2005).
But in the form factor of a workstation … a cluster workstation
ORION DT-12 DESKTOP CLUSTER WORKSTATION

Imagine a 36 Gflop cluster on your desk!

12 Nodes
in a single computer

36 Gflops
peak processing power

24 GBytes
memory capacity

1 TByte
internal storage

DESIGNED FOR THE INDIVIDUAL
The Orion DT-12 cluster workstation is a fully integrated, completely self-contained, personal workstation based on the best of today's cluster technologies. Designed to be an affordable individual resource it is capable of 36 Gflops peak performance (18 Gflops sustained) with models starting at under $10k.

The Orion DT-12 cluster workstation provides supercomputer performance for the engineering, scientific, financial and creative professionals who need to solve computationally complex problems without waiting in the queue of the back-room cluster.

FASTER SOFTWARE DEVELOPMENT
The Orion DT-12 cluster workstation is the perfect platform for developers writing and deploying cluster software packages. It comes with cluster software development tools pre-installed, including libraries and a parallel compiler that allows you to spread one multiple-file compile to all the nodes in the system. Also included is a suite of system monitoring and management software.

NO ASSEMBLY REQUIRED
Orion workstations are designed from the ground up as a single computer. The entire system boots with the push of a button and has the ergonomics and ease of use of a personal computer. The modular design allows for flexible configurations and scalability by stacking up to 4 systems as one 48 node cluster.

PRESERVE SOFTWARE INVESTMENTS
Orion workstations are built around industry standards for clustering: x86 processors, Ethernet, the Linux operating system and standard parallel programming libraries, including MPICH, PVM and SGE. Existing Linux cluster applications run without modification.

PERFORMANCE AND FEATURES
The Orion DT-12 is a cluster of 12 x86-compatible nodes linked by a switched Gigabit Ethernet fabric. The cluster operates as a single computer with a single on-off switch and a single system image rapid boot sequence, which allows the entire system to boot in less than 90 seconds.

The Orion DT-12 cluster workstation is highly efficient, consuming a maximum of 220 Watts of power under peak load—about the same as an average desktop PC. It operates quietly, plugs into a standard 110V 15A wall socket and fits unobtrusively on a desk or lab bench.

(Circa 2003-2004)

Orion DT-12

- Footprint
  - 3 sq. ft. (24” x 18”)
  - 1 cu. ft. (24” x 4” x 18”)

- Power Consumption
  - 170 watts at load

Power Efficient
Performance/Core
Price
Proprietary Hardware
(Limited Trickle-Down)
IBM Blue Gene/L

System
(64 cabinets, 64x32x32)

Cabinet
(32 Node boards, 8x8x16)

Node Card
(32 chips, 4x4x2)
16 Compute Cards

Compute Card
(2 chips, 2x1x1)

Chip
(2 processors)

2.8/5.6 GF/s
4 MB

5.6/11.2 GF/s
0.5 GB DDR

90/180 GF/s
8 GB DDR

2.9/5.7 TF/s
256 GB DDR

October 2003
BG/L half rack prototype
500 Mhz
512 nodes/1024 proc.
2 TFlop/s peak
1.4 Tflop/s sustained

© 2004 IBM Corporation

ss.s.cs.vt.edu

synergy.cs.vt.edu
SiCortex SC 648 and SC 5832

(Circa 2006)

Sources: SiCortex, Google, and BigNComputing

CPU Power: 0.6 W
SiCortex SC 648 (648 Gflops peak)
- 2 kW for 648-CPU system
SiCortex SC 5832 (5.8 Tflops peak)
- 18 kW for 5832-CPU system

Green Computing Performance Index (GCPI)
Sample Comparisons of Leading HPC Systems
GCPI = (System Performance) / kWatt

- Power Efficient
- Performance/Core
- Price
- Proprietary Hardware (Limited Trickle-Down)

\(^1\)GCPI = n(HPCC results)/kWatt, where n = results normalized to Cray XT3 reference system. HPCC is an industry-standard benchmark suite comprising 7 tests and a total of 28 benchmarks.

\(^2\)Intel 'Nehalem' GCPI results estimated from posted HPCC benchmark results (30 March 09) for the Intel Endeavor Xeon 5560 and derived energy consumption.
Outline

• Motivation & Background

• Supercomputing in Small Spaces (http://sss.cs.vt.edu/)
  – Origin: Green Destiny
  – Evolution
    • Architectural: Orion Multisystems → Sun Microsystems
    • Software: ENERGYFIT™ Power-Aware Run-Time OS Daemon

• The Future?

• The Green500 List: Past, Present, and Future

• Conclusion
Software-Based Evolution of **Green Destiny**: Dynamic Voltage & Frequency Scaling (DVFS)

- **DVFS Mechanism**
  - Trades CPU performance for power reduction by allowing the CPU supply voltage and/or frequency to be adjusted at run-time.

- **Why was/is DVFS important?**
  - Recall: Moore’s Law for Power.
  - CPU power consumption is directly proportional to the square of the supply voltage and to frequency.

- **DVFS Algorithm**
  - Determines when to adjust the current frequency-voltage setting and what the new frequency-voltage setting should be.
Automated Power-Aware Run-Time OS Daemon

- Self-Adapting Software for Energy Efficiency
  Tag Line: Conserve power & energy **WHILE** maintaining performance.


- Observations
  - Support for *dynamic voltage & frequency scaling (DVFS)*, which allows changes to processor voltage and frequency at run time.
  
  - Trade off of processor performance for power reduction.
    - *Power* $= V^2 \cdot f$ where $V$ is the supply voltage of the processor and $f$ is its frequency.
    - *Processor performance* $= a \cdot f$
Software-Based Energy Efficiency via DVFS:

Key Observation

• Execution time of many programs is insensitive to CPU speed change.

• Why?
  – The Memory & I/O Wall → The Brick Wall
Software-Based Energy Efficiency via DVFS:

**Key Idea**

- Applying DVFS to these programs will result in significant power & energy savings at a minimal performance impact.
Why is Power Awareness via DVFS Hard?

• What is cycle time of a processor?
  – Frequency $\approx 2 \text{ GHz } \rightarrow \text{ Cycle Time } \approx \frac{1}{(2 \times 10^9)} = 0.5 \text{ ns}$

• How long does the system take to scale voltage and frequency?

  $O(10,000,000 \text{ cycles})$
Problem Formulation:

Linear Programming for an Energy-Optimal DVFS Schedule

• Definitions
  – A DVFS system exports \( n \) \( \{ (f_i, P_i) \} \) settings.
  – \( T_i \): total execution time of a program running at setting \( i \)

• Given a program with deadline \( D \), find a DVFS schedule \((t_1^*, \ldots, t_n^*)\) such that
  – If the program is executed for \( t_i \) seconds at setting \( i \), the total energy usage \( E \) is minimized, the deadline \( D \) is met, and the required work is completed.

\[
\min E = \sum_i P_i \cdot t_i
\]

subject to

\[
\sum_i t_i \leq D
\]
\[
\sum_i t_i/T_i = 1
\]
\[
t_i \geq 0
\]

C. Hsu and W. Feng.
“A Power-Aware Run-Time System for High-Performance Computing,”

Embrace the power wall
… select the right setting
… at the right time
for the workload at hand

Hardware-Software Co-Design
β-Adaptation DVFS Scheduling Algorithm

- **Input:** Relative slowdown $\delta$ and performance model $T(f)$.
- **Output:** Constraint-based DVS schedule.
- **For every $I$ seconds do**
  1. Compute coefficient $\beta$
  2. Compute the desired frequency $f^*$
     - If $f^*$ is not a supported frequency, then
       1. Identify $f_j$ and $f_{j+1}$.
       2. Compute the ratio $r$.
       3. Run $r \cdot I$ seconds at frequency $f_j$.
       4. Run $(1 - r) \cdot I$ seconds at frequency $f_{j+1}$.
       5. Update $\text{mips}(f_j)$ and $\text{mips}(f_{j+1})$.
  - Else run at $f^*$.

$$r = \frac{(1 + \delta/\beta)/f_{\text{max}} - 1/f_{j+1}}{1/f_{j} - 1/f_{j+1}}$$

$$f^* = \begin{cases} f_{\text{min}} & \text{if } \beta \leq \delta \\ f_{\text{max}}/(1 + \delta/\beta) & \text{otherwise} \end{cases}$$
Experimental Set-Up

• Tested Computer Platforms with PowerNow! Enabled
  – Mobile AMD Athlon XP (with five frequency-voltage settings) – same processor used in the Sun BladeSystem.
  – 64-bit AMD Athlon 64
  – 64-bit AMD Opteron → CAFfeine Power-Aware Cluster

• Digital Power Meter
  – Yokogawa WT210: Continuously samples every 20 µs.

• Benchmarks Used
  – Uniprocessor: SPEC
  – Multiprocessor: Linpack and NAS Parallel Benchmarks
### β - Adaptation with Sequential Codes (SPEC CPU)

<table>
<thead>
<tr>
<th>program</th>
<th>β</th>
<th>2step</th>
<th>nqPID</th>
<th>freq</th>
<th>mips</th>
<th>beta</th>
</tr>
</thead>
<tbody>
<tr>
<td>swim</td>
<td>0.02</td>
<td>1.00/1.00</td>
<td>1.04/0.70</td>
<td>1.00/0.96</td>
<td>1.00/1.00</td>
<td>1.04/0.61</td>
</tr>
<tr>
<td>tomcatv</td>
<td>0.24</td>
<td>1.00/1.00</td>
<td>1.03/0.69</td>
<td>1.00/0.97</td>
<td>1.03/0.83</td>
<td>1.00/0.85</td>
</tr>
<tr>
<td>su2cor</td>
<td>0.27</td>
<td>0.99/0.99</td>
<td>1.05/0.70</td>
<td>1.00/0.95</td>
<td>1.01/0.96</td>
<td>1.03/0.85</td>
</tr>
<tr>
<td>compress</td>
<td>0.37</td>
<td>1.02/1.02</td>
<td>1.13/0.75</td>
<td>1.02/0.97</td>
<td>1.05/0.92</td>
<td>1.01/0.95</td>
</tr>
<tr>
<td>mgrid</td>
<td>0.51</td>
<td>1.00/1.00</td>
<td>1.18/0.77</td>
<td>1.01/0.97</td>
<td>1.00/1.00</td>
<td>1.03/0.89</td>
</tr>
<tr>
<td>vortex</td>
<td>0.65</td>
<td>1.01/1.00</td>
<td>1.25/0.81</td>
<td>1.01/0.97</td>
<td>1.07/0.94</td>
<td>1.05/0.90</td>
</tr>
<tr>
<td>turb3d</td>
<td>0.79</td>
<td>1.00/1.00</td>
<td>1.29/0.83</td>
<td>1.03/0.97</td>
<td>1.01/1.00</td>
<td>1.05/0.94</td>
</tr>
<tr>
<td>go</td>
<td>1.00</td>
<td>1.00/1.00</td>
<td>1.37/0.88</td>
<td>1.02/0.99</td>
<td>0.99/0.99</td>
<td>1.06/0.96</td>
</tr>
</tbody>
</table>

**relative time / relative energy**
with respect to total execution time and system energy usage

SMALLER numbers are BETTER.

C. Hsu and W. Feng.
“A Power-Aware Run-Time System for High-Performance Computing,”
**β - Adaptation with Sequential Codes (SPECjbb)**

<table>
<thead>
<tr>
<th>Power Management</th>
<th>Watts</th>
<th>% Power Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>264</td>
<td>0%</td>
</tr>
<tr>
<td>Cpuspeed</td>
<td>257</td>
<td>3%</td>
</tr>
<tr>
<td>Ondemand</td>
<td>253</td>
<td>4%</td>
</tr>
<tr>
<td>β</td>
<td>196</td>
<td>25%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Power Management</th>
<th>bops/watt</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>100.00%</td>
</tr>
<tr>
<td>Cpuspeed</td>
<td>102.56%</td>
</tr>
<tr>
<td>Ondemand</td>
<td>104.37%</td>
</tr>
<tr>
<td>β</td>
<td>123.70%</td>
</tr>
</tbody>
</table>
$\beta$ – Adaptation with Linpack

Blue Line is $\beta$ – adaptation

15% Power Saving

1% (10 ms) Performance Loss
\( \beta \) – Adaptation on NAS Parallel Benchmarks

C. Hsu and W. Feng.
“A Power-Aware Run-Time System for High-Performance Computing,”
Finally … in 2006 … Why Be Green?

- **K Computer**
  - Power & Cooling: 12.66 MW → $12M/year

- **Google in The New York Times, June 14, 2006**

  Hiding in Plain Sight, Google Seeks More Power

Google is building two computing centers, top and left, each the size of a football field, in The Dalles, Ore.
Google Details and Defends Its Use of Electricity

“Google disclosed Thursday that it continuously uses enough electricity to power 200,000 homes.”

The New York Times, September 18, 2011
Outline

• Motivation & Background
• Supercomputing in Small Spaces (http://sss.cs.vt.edu/)
  – Origin: Green Destiny
  – Evolution
    • Architectural: Orion Multisystems → Sun Microsystems
    • Software: ENERGYFIT™ Power-Aware Run-Time OS Daemon
  • The Future? Energy Proportionality
• The Green500 List: Past, Present, and Future
• Conclusion
Severe Underutilization in Data Centers

Average CPU utilization of 5000 servers over a six-month period

Severe Underutilization in Data Centers

Average CPU utilization of 5000 servers over a six-month period

Servers are heavily underutilized

Severe Underutilization in Data Centers

Average CPU utilization of 5000 servers over a six-month period

Servers typically operate between 20%-50% utilization

Region of data center operation

Consume power proportional to utilization (or load-level)
Consume power proportional to utilization (or load-level)
The Case for Energy-Proportional Computing

Luiz André Barroso and Urs Hölzle

Consume power proportional to utilization (or load-level)
Consume power proportional to utilization (or load-level)
The Case for Energy-Proportional Computing

Luiz André Barroso and Urs Hölzle

- Consume power proportional to utilization
- Advocate improvements in non-peak power efficiency
• Properties of an energy-proportional system

Low idle power
(0 watts in ideal case)
Properties of an energy-proportional system
The Case for Energy-Proportional Computing

Luiz André Barroso and Urs Hölzle

- Consume power proportional to utilization
- Advocate improvements in non-peak power efficiency
- Enhance data center energy efficiency via non-peak power efficiency improvements
The Case for Energy-Proportional Computing

Luiz André Barroso and Urs Hölzle

- Consume power proportional to utilization
- Advocate improvements in non-peak power efficiency
- Enhance data center energy efficiency via non-peak power efficiency improvements

Power-Capping Mechanism (e.g., Intel’s Running Average Power Limit (RAPL))
- A way to cap power consumption when a system is underutilized
Power Efficiency at Different Levels of Utilization

Towards Energy-Proportional Computing for Enterprise-Class Server Workloads.
Balaji Subramaniam, Wu-chun Feng.
In Proceedings of the 3rd ACM/SPEC International Conference on Performance Engineering (ICPE), Prague, Czech Republic, April 2013. Best Paper Award
Power Efficiency at Different Levels of Utilization

- Near energy-proportional power consumption at high levels of utilization
Power Efficiency at Different Levels of Utilization

• Near energy-proportional power consumption at high levels of utilization
• Significant gap between ideal and energy-proportional power consumption in the region of typical data center operation

Opportunity for significant improvement in power efficiency
Power Savings – Package+Memory Subsystems
Power Savings – Package+Memory Subsystems

- Energy-proportional operation for 80% load-level
Power Savings – Package+Memory Subsystems

- Energy-proportional operation for 80% load-level
- Near energy-proportional operation for 60% load-level
Power Savings – Core Subsystem

The graph illustrates the power savings in the core subsystem as a function of target load. The x-axis represents the target load, while the y-axis shows the percentage of peak power. Several lines represent different scenarios:

- Vanilla
- CPUOnly
- MemOnly
- CPU + Mem
- Ideal

The lines show how each configuration compares to the ideal scenario, highlighting the effectiveness of power management techniques.
• Better than energy-proportional operation in nearly all cases
  – All of the power saving comes from the Core subsystem
CPU+Mem achieves best overall power savings (19% power saved)
Energy-proportional operation for 80% load-level
Power Savings – Full System

- CPU+Mem achieves best overall power savings (19% power saved)
- Energy-proportional operation for 80% load-level
Outline

• Motivation & Background
• Supercomputing in Small Spaces (http://sss.cs.vt.edu/)
  – Origin: Green Destiny
  – Evolution
    • Architectural: Orion Multisystems → Sun Microsystems
    • Software: EnergFr™ Power-Aware Run-Time OS Daemon
• The Future? Data Movement
• The Green500 List: Past, Present, and Future
• Conclusion
Data Movement: A Major Source of Power Consumption

- Data movement through the memory hierarchy thought to be a major source of power consumption

Based on Shalf et al., “Exascale Computing Technology Challenges,” VECPAR 2010

- Measurements on real systems lacking
  - Current estimates based on simulation studies
  - Real-world measurements are coarse-grained and do not give breakdown for data movement
State-of-the-Art Measurement Approach

\[ \text{Energy consumed by L2 microbenchmark} = E_{L2} \]

\[ \text{Energy consumed by L1 microbenchmark} = E_{L1} \]

\[ \text{Energy cost of moving data from L2 to L1} = E_{L2} - E_{L1} \]

**Issue:** Over-estimation of data-movement energy

G. Kestor et al., "Quantifying the energy cost of data movement in scientific applications," IISWC 2013
State-of-the-Art Measurement Approach: Limitation

**Issue:** Data-movement power also includes L2-access power
A Bit of Background …

Representative block diagram of AMD FirePro™ W9100 GPU
(Previously Code-named “Hawaii”)

Distance between shader engines and L2 banks differ
A Bit of Background …

Physical distance traversed by data thought to affect data-movement power

Representative block diagram of AMD FirePro™ W9100 GPU
(Previously Code-named “Hawaii”)
Our Proposed Approach

Short-path microbenchmark

Long-path microbenchmark

Design microbenchmarks based on data-movement distance to properly isolate the interconnect.
Our Proposed Approach

Short-path microbenchmark

Long-path microbenchmark

Design microbenchmarks based on data-movement distance to properly isolate the interconnect
CHALLENGES
Challenges

OpenCL™ lacks native support to pin threads to programmer-specified compute units

Temperature of device during tests will affect the power consumption

Power difference between the two sets of microbenchmarks can be hard to observe
Challenges

Only a small difference between used-L2 and L1 cache size which can make it challenging to write L2-only microbenchmarks.

Ensuring that the same amount of work is done by the two microbenchmarks can be challenging due to NUCA effects.
Challenges

OpenCL™ lacks native support to pin threads to programmer-specified compute units

Temperature of device during tests will affect the power consumption

Power difference between the two sets of microbenchmarks can be hard to observe
Addressing the challenges

**Pinning Threads to Cores with OpenCL™**

Can be accomplished with some binary hacking
Challenges

OpenCL™ lacks native support to pin threads to programmer-specified compute units

Temperature of device during tests will affect the power consumption

Power difference between the two sets of microbenchmarks can be hard to observe
Addressing the challenges

Eliminating Temperature Effects

Solution 1: Run GPU fans at very high speed to limit temperature difference between runs
Addressing the challenges

Eliminating Temperature Effects

**Solution 2:** Model idle power separately and subtract from measured power
Challenges

OpenCL™ lacks native support to pin threads to programmer-specified compute units.

Temperature of device during tests will affect the power consumption.

Power difference between the two sets of microbenchmarks can be hard to observe.
Addressing the challenges

**Saturating Interconnect Bandwidth**

- Power difference between microbenchmarks can be too low to be reliably observed
  - Unless the amount of data going through the interconnect increases
- Difficult to saturate interconnect bandwidth without increasing the number of wavefronts
  - Fewer wavefronts can result in stalling exposing latency difference issues
- More wavefronts can lead to one of the following issues
  - More L1 hits when accesses per thread is low
  - Register pressure and memory spills if access per thread is high

**Solution:** Modify firmware to artificially shrink L1 cache size and then increase number of wavefronts
Parameters Affecting Interconnect Power

- Average interconnect distance
- Data toggle rate
- Bandwidth observed on the interconnect
- Interconnect’s voltage and frequency
Our experiments confirm that the distance traversed by data affects power consumption.
Interconnect Power vs Distance

Within 15% of industrial estimates for energy/bit/mm
Interconnect Power vs Distance

Linear relationship observed between data-movement distance and interconnect power
Interconnect Power vs Voltage and Frequency

Impact of voltage and frequency, as expected
Energy Cost of Data Movement on Real Applications

28 nm AMD FirePro™ W9100 GPU architecture

Up to 14% dynamic power spent on interconnects in today’s chips
Energy Cost of Data Movement on Real Applications

Even non-memory bound applications can show high interconnect power (but at different hierarchy)
Energy Cost of Data Movement on Real Applications

We use Borkar’s scaling factors for wires and transistors to scale the total dynamic power and interconnect power from 28nm to 7nm.

Data-movement power problem problem exacerbated in future technology nodes
Interconnect Power Optimizations

- **Layout-Based Optimizations**
  - Explore how much impact layout-based optimization can have on real world applications
  - Examine two layouts – one reduces distance between L1 and L2 and the other reduces distance between L2 and memory controller

- **Cache Resizing**
  - Increasing cache size decreases average data movement distance (most data is fetched from nearer memories)
  - Quantify the magnitude of difference when we increase the cache size four (4) times
Interconnect Power Optimizations – Layout Optimization

L1-$ \quad L2-$ \quad Mem Controller

L1 to L2 = 17.0 units  
L2 to MC = 7.6 units  

Optimized to reduce L2-MC distance.
Not too different from current GPUs.
Interconnect Power Optimizations – Layout Optimization

L1 to L2 = 17.0 units
L2 to MC = 7.6 units

L1 to L2 = 3.5 units
L2 to MC = 12.0 units

Optimized to reduce L1-L2 distance.
Interconnect Power Optimizations – Layout Optimization

Interconnect power reduces by 48% on average when we use an L1-L2 distance optimized layout.
The Future? Data Movement

• Distance-based microbenchmarking – a promising approach to measure data movement power
• Over 14% of dynamic power can go towards on-chip data movement in today’s chips
  – Lesser than past estimates as we separated out data access power from data movement power
  – Can increase to 22% by 7nm technology
• Optimizing on-chip interconnects to reduce the distance of frequently accessed portions can reduce on-chip data movement power by 48%

Measuring and Modeling On-chip Interconnect Power on Real Hardware.
V. Adhinarayanan, I. Paul, J. Greathouse, W. Huang, A. Pattnaik, W. Feng.
In IEEE International Symposium on Workload Characterization (IISWC), September 2016. Best Paper Award
Outline

• Motivation & Background
• Supercomputing in Small Spaces (http://sss.cs.vt.edu/)
  – Origin: Green Destiny
  – Evolution
    • Architectural: Orion Multisystems → Sun Microsystems
    • Software: EnergyFit™ Power-Aware Run-Time OS Daemon
• The Future?
• The Green500 List: Past, Present, and Future
• Conclusion
Brief History:

From Green Destiny to the Green500 List

  – “Honey, I Shrank the Beowulf!”
  31st Int’l Conf. on Parallel Processing, August 2002

• 04/2005: Keynote Talk by W. Feng at the IEEE Workshop
  on High-Performance, Power-Aware Computing
  – Generates initial discussion for Green500 List

• 04/2006 and 09/2006: Making a Case for a Green500 List
  (Sharma, Hsu, and Feng)
  – Workshop on High-Performance, Power-Aware Computing
  – Jack Dongarra’s CCGSC Workshop “The Final Push”

• 09/2006: Launch of Green500 Web Site & RFC (Feng & Hsu)
  – http://www.green500.org
Brief History:
From Green Destiny to the Green500 List

• 11/2007: First official Green500 list released (Feng & Cameron)
• 11/2010: First official Green500 run rules released
• 06/2011: Collaborations begin on standardizing metrics, methodologies, and workloads for energy-efficient parallel computing
  – Energy-Efficient High-Performance Computing Working Group (EE HPC WG) – Bates
  – The Green Grid
  – TOP500 – Strohmaier
  – Green500 – Feng and Scogland
Exascale Computing Study: Technology Challenges in Achieving Exascale Systems

• Goal
  – “Because of the difficulty of achieving such physical constraints, the study was permitted to assume some growth, perhaps a factor of 2X, to something with a maximum limit of 500 racks and 20 MW for the computational part of the 2015 system.”

• Realistic Projection?
  – “Assuming that Linpack performance will continue to be of at least passing significance to real Exascale applications, and that technology advances in fact proceed as they did in the last decade (both of which have been shown here to be of dubious validity), then […] an Exaflop per second system is possible at around 67 MW.”
Projections for Energy-Efficient Exascale Systems

• How to achieve 20 MW exascale system?
Projections for Energy-Efficient Exascale Systems

• How to achieve 20 MW exascale system?
  – Current fastest supercomputer, Sunway TaihuLight
    • 98.01 PFLOPS consuming 15.37 MW power.
Projections for Energy-Efficient Exascale Systems

• How to achieve 20 MW exascale system?
  – Current fastest supercomputer, Sunway TaihuLight
    • 93.01 PFLOPS consuming 15.37 MW power.
  – Required: 10.75-fold increase in performance with only 1.30-fold increase in power. (4.36-fold for more realistic 67 MW system.)
Projections for Energy-Efficient Exascale Systems

• How to achieve 20 MW exascale system?
  – Current fastest supercomputer, Sunway TaihuLight
    • 93.01 PFLOPS consuming 15.37 MW power.
    – Required: 10.75-fold increase in performance with only 1.30-fold increase in power. (4.36-fold for more realistic 67 MW system.)

• Is this target realistic?
Projections for Energy-Efficient Exascale Systems

• How to achieve 20 MW exascale system?
  – Current fastest supercomputer, Sunway TaihuLight
    • 93.01 PFLOPS consuming 15.37 MW power.
    – Required: 10.75-fold increase in performance with only 1.30-fold increase in power. (4.36-fold for more realistic 67 MW system.)

• Is this target realistic?
  – Track energy efficiency in the past
  – Project energy efficiency for the future

What does the DARPA Exascale Report predict?
What does the Green500 data project?
Relook at Exascale Strawman

In 2015, core energy per flop for Linpack is < 10pJ

<table>
<thead>
<tr>
<th>Operation</th>
<th>Energy (pJ/bit)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Register File Access</td>
<td>0.16</td>
</tr>
<tr>
<td>SRAM Access</td>
<td>0.23</td>
</tr>
<tr>
<td>DRAM Access</td>
<td>1</td>
</tr>
<tr>
<td>On-chip movement</td>
<td>0.0187</td>
</tr>
<tr>
<td>Thru Silicon Vias (TSV)</td>
<td>0.011</td>
</tr>
<tr>
<td>Chip-to-Board</td>
<td>2</td>
</tr>
<tr>
<td>Chip-to-optical</td>
<td>10</td>
</tr>
<tr>
<td>Router on-chip</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Step</th>
<th>Target</th>
<th>pJ</th>
<th>#Occurrences</th>
<th>Total pJ</th>
<th>% of Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read Alphas</td>
<td>Remote</td>
<td>13,819</td>
<td>4</td>
<td>55,276</td>
<td>16.5%</td>
</tr>
<tr>
<td>Read pivot row</td>
<td>Remote</td>
<td>13,819</td>
<td>4</td>
<td>55,276</td>
<td>16.5%</td>
</tr>
<tr>
<td>Read 1st Y[i]</td>
<td>Local</td>
<td>1,380</td>
<td>88</td>
<td>121,1</td>
<td>3.9%</td>
</tr>
<tr>
<td>Read Other Y[i]</td>
<td>L1</td>
<td>39</td>
<td>264</td>
<td>10,070</td>
<td>3.1%</td>
</tr>
<tr>
<td>Write Y's</td>
<td>L1</td>
<td>39</td>
<td>352</td>
<td>13,900</td>
<td>4.2%</td>
</tr>
<tr>
<td>Flush Y's</td>
<td>Local</td>
<td>891</td>
<td>88</td>
<td>78,480</td>
<td>23.4%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td>334,056</td>
<td></td>
</tr>
<tr>
<td>Ave per Flop</td>
<td></td>
<td></td>
<td></td>
<td>475</td>
<td></td>
</tr>
</tbody>
</table>

If this is true, 1 EF/s = 0.5 GW!  
Source: Peter Kogge

If this is true, 1 EF/s = 0.5 GW!
Projection to Exascale via Green500

- Efficiency of an Exaflop System
  - 20-MW power envelope: 50 GFLOPS/watt
  - 100-MW power envelope: 10 GFLOPS/watt
Projection to Exascale via Green500

• Efficiency of an Exaflop System
  – 20-MW power envelope: 50 GFLOPS/watt
  – 100-MW power envelope: 10 GFLOPS/watt

• State of the Art
  – Greenest system
  – Fastest system:
Projection to Exascale via Green500

- **Efficiency of an Exaflop System**
  - 20-MW power envelope: 50 GFLOPS/watt
  - 100-MW power envelope: 10 GFLOPS/watt

- **State of the Art (2014)**
  - Greenest system: 4.50 GFLOPS/watt
  - Fastest system: 2.14 GFLOPS/watt

- **Projections: Efficiency (GFLOPS/watt) in 2018 and 2020**

<table>
<thead>
<tr>
<th>Class</th>
<th>In 2018</th>
<th>In 2020</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green5</td>
<td>5.78</td>
<td>6.93</td>
</tr>
<tr>
<td>Green10</td>
<td>5.48</td>
<td>6.60</td>
</tr>
<tr>
<td>Green50</td>
<td>3.93</td>
<td>4.75</td>
</tr>
<tr>
<td>Green100</td>
<td>2.88</td>
<td>3.48</td>
</tr>
</tbody>
</table>

11-fold improvement needed for 20-MW envelope.
Projection to Exascale via Green500

- Efficiency of an Exaflop System
  - 20-MW power envelope: 50 GFLOPS/watt
  - 100-MW power envelope: 10 GFLOPS/watt

- State of the Art (2017)
  - Greenest system: 16.7 GFLOPS/watt
  - Fastest system: 6.05 GFLOPS/watt

- Projections: Efficiency (GFLOPS/watt) in 2018 and 2020

<table>
<thead>
<tr>
<th>Class</th>
<th>In 2018</th>
<th>In 2020</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green5</td>
<td>5.78</td>
<td>6.93</td>
</tr>
<tr>
<td>Green10</td>
<td>5.48</td>
<td>6.60</td>
</tr>
<tr>
<td>Green50</td>
<td>3.93</td>
<td>4.75</td>
</tr>
<tr>
<td>Green100</td>
<td>2.88</td>
<td>3.48</td>
</tr>
</tbody>
</table>
Projection to Exascale: Extrapolating to Exaflop

Power extrapolated to exaflop (Gigawatts)

List release year

08 09 10 11 12 13 14

Green500
Top500

222 MW
Projection to Exascale: Extrapolating to Exaflop

Top in

Green500
Top500

Power extrapolated to exaflop (MW)

1500
1000
500
0

List release year

165.3 MW
105.7 MW
Outline

• Motivation & Background
• Supercomputing in Small Spaces (http://sss.cs.vt.edu/)
  – Origin: Green Destiny
  – Evolution
    • Architectural: Orion Multisystems → Sun Microsystems
    • Software: EnergyFit™ Power-Aware Run-Time OS Daemon
• The Future?
• The Green500 List: Past, Present, and Future
• Conclusion
Conclusion

• Recall that …
  – We need **horsepower** to compute but **horsepower** also translates to electrical **power** consumption.

• Consequences
  – Electrical power costs $$$$.
  – “Too much” power affects efficiency, reliability, availability.

• Approaches
  – Autonomic energy and power savings while maintaining performance in the supercomputer and datacenter.
  – Low-Power Architectural and Power-Aware Commodity
  – Future? Energy Proportionality and Data Movement

• Venues of Evaluation
  – Green500
For More Information

W. Feng, wfeng@vt.edu, +1-540-231-1192

The Synergy Lab (http://synergy.cs.vt.edu)

Supercomputing in Small Spaces (http://sss.cs.vt.edu/)

NSF Center for High-Performance Reconfigurable Computing (http://chrec.cs.vt.edu/ → http://www.chrec.org/)

MyVICE: My Virtual Instances of Computing Environments for K-12 CS Education (http://myvice.cs.vt.edu/)

Green500 (http://www.green500.org/)

mpiBLAST (http://www.mpiblast.org/)

HokieSpeed (http://hokiespeed.cs.vt.edu/)