Connecting the Dots between Parallel Programming and Energy

Connecting the dots between parallel programming and energy

Professor Dimitrios S. Nikolopoulos
d.nikolopoulos@qub.ac.uk
Points to get across

• PDP and HPC should be leading (not lagging) research on energy-efficient computing
• Parallel programming, runtimes systems and energy-efficiency are connected
• What lies ahead for the parallel programmer
Acknowledgments
Our resources

EPSRC
Pioneering research and skills

DEPARTMENT OF ENERGY
UNITED STATES OF AMERICA

NSF

7
PEOPLE

IBM

7
COOPERATION

Lawrence Livermore National Laboratory
Energy in HPC

Papers tagged "Power/Energy/HPC" ACM Digital Library

2012: 270
2011: 260
2010: 110
2009: 60
2008: 40
2007: 60
2006: 30
2005: 20
2004: 10
2003: 5
2002: 5
2001: 1
Leader or laggard?

- The history of BlueGene
  - Based on processors for the embedded systems market (PowerPC)
  - Pioneered “scale-out” idea, now common in datacentres
    - Many nodes with simple cores, custom interconnect
  - Dominated Top-500 list for performance
BlueGene on the Green500

BlueGene and the Green-500 List

MFLOPS per Watt

- BlueGene/L-P-Q
- Peak Green-500

Nov-07 Jan-08 Mar-08 May-08 Jul-08 Sep-08 Nov-08 Jan-09 Mar-09 May-09 Jul-09 Sep-09 Nov-09 Jan-10 Mar-10 May-10 Jul-10 Sep-10 Nov-10 Jan-11 Mar-11 May-11 Jul-11 Sep-11 Nov-11 Jan-12 Mar-12 May-12 Jul-12 Sep-12 Nov-12

BlueGene/L-P-Q
Peak Green-500

Friday, March 1, 13
EuroMicro PDP'13 Keynote
# Leader or laggard?

<table>
<thead>
<tr>
<th>Processor</th>
<th>Type</th>
<th>GFLOPS (32bit)</th>
<th>GFLOPS (64bit)</th>
<th>Watt (TDP)</th>
<th>GFLOPS/Watt (32bit)</th>
<th>FLOPS/Watt (64bit)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adapteva Epiphany-IV</td>
<td>Epiphany</td>
<td>100</td>
<td>N/A</td>
<td>2</td>
<td>50</td>
<td>N/A</td>
</tr>
<tr>
<td>Movidius Myriad</td>
<td>ARM SoC: LEON3+SHAPE</td>
<td>15.28</td>
<td>N/A</td>
<td>0.32</td>
<td>48</td>
<td>N/A</td>
</tr>
<tr>
<td>ZiiLabs</td>
<td>ARM SoC</td>
<td>58</td>
<td>N/A</td>
<td>?</td>
<td>20?</td>
<td>N/A</td>
</tr>
<tr>
<td>Nvidia Tesla K10</td>
<td>X86 GPU</td>
<td>4577</td>
<td>190</td>
<td>225</td>
<td>20.34</td>
<td>?</td>
</tr>
<tr>
<td>ARM + MALI T804</td>
<td>ARM SoC</td>
<td>8468</td>
<td>N/A</td>
<td>4?</td>
<td>19?</td>
<td>N/A</td>
</tr>
<tr>
<td>Nvidia GTX 690</td>
<td>X86 GPU x 2</td>
<td>5621</td>
<td>234?</td>
<td>300</td>
<td>18.74</td>
<td>0.78</td>
</tr>
<tr>
<td>GeForce GTX 680</td>
<td>X86 GPU</td>
<td>3090</td>
<td>118</td>
<td>195</td>
<td>15.85</td>
<td>0.65</td>
</tr>
<tr>
<td>AMD Radeon HD 7970 GHz</td>
<td>X86 GPU</td>
<td>4300</td>
<td>1075</td>
<td>300+</td>
<td>14.3</td>
<td>3.58</td>
</tr>
<tr>
<td>Intel Knight’s Corner (Xeon Phi)</td>
<td>X87+</td>
<td>2000?</td>
<td>1000</td>
<td>200?</td>
<td>10?</td>
<td>5?</td>
</tr>
<tr>
<td>AMD A10-5800K + HD 7660D</td>
<td>X86 SoC</td>
<td>121 + 614</td>
<td>?</td>
<td>100</td>
<td>7.35</td>
<td>?</td>
</tr>
<tr>
<td>Intel Core i7-3770 + HD4000</td>
<td>X86 SoC</td>
<td>225 + 294.4</td>
<td>112 + 73.6</td>
<td>77</td>
<td>6.74</td>
<td>2.41</td>
</tr>
<tr>
<td>NVIDIA CARMA (complete board)</td>
<td>ARM + GPU</td>
<td>? + 200</td>
<td>?</td>
<td>40</td>
<td>5.00</td>
<td>?</td>
</tr>
<tr>
<td>IBM Power A2</td>
<td>Power CPU</td>
<td>204?</td>
<td>204</td>
<td>55</td>
<td>3.72?</td>
<td>3.72</td>
</tr>
<tr>
<td>Intel Core i7-3770</td>
<td>X86 CPU</td>
<td>225</td>
<td>112</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>AMD A10-5800K</td>
<td>X86 CPU</td>
<td>121</td>
<td>60?</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>
Leader or laggard?

- Is HPC just reusing solutions?
  - Processors originally designed for the mobile phones market
  - Clock gating, DVFS, device sleep states well known for 20 years
Still looking for a proper metric...

- FLOPS/Watt
- FLOPS/Watt/square inch
- Energy \times Delay
- Energy \times Delay \times Delay
- Energy \times Time to solution
- £, $, €…

- Or should we just measure execution time and forget about the rest?
Is there a role for HPC & PDP in green computing innovation?
Exascale projections

• Assume that currently most energy-efficient supercomputer sustains improvement towards an Exaflop
  – Factor of $2384 \times$ in performance
  – 202.7 MegaWatt of power

• Assume target power cap of 20 MegaWatt
  – Need two orders of magnitude improvement in FLOPS/Watt
  – What hardware devices can achieve this improvement without compromising performance?
Parallel software is (or should be) energy-efficient

- Parallel software can scale up, down and out, if written with appropriate programming models

- Scale-free software can adapt to power constraints and control energy budgets
Concurrency throttling

- Dynamic power increases linearly with number of active cores
- Scalability curves have knees
  - Locate the knee
  - Energy-efficient parallel programs execute at the starting point of the knee
Exploiting the concurrency trade-off

- Programs execute distinct phases
  - Compute-, memory- or communication-bound
- Phase characterization indicates concurrency sweet spot
  - DCT can both improve performance and save power if the sweet spot is found
- DCT is one of several optimizations
  - Can easily complement DVFS to maximize energy-efficiency
Unified program adaptation framework

- Multi-dimensional online performance predictor [Curtis-Maury et al., TPDS08, PACT08, ICS06]
  - Statistically analyzes samples of hardware event rates collected at runtime
  - Derives predictions based on data from short execution samples
- Runtime adaptation per program phase
  - Thread-based programming model (OpenMP, Cilk)
  - DCT and DVFS based on phase scaling predictions

Sample

Targets

<table>
<thead>
<tr>
<th>Active core at full frequency</th>
<th>Idle core</th>
</tr>
</thead>
</table>

Friday, March 1, 13

EuroMicro PDP'13 Keynote
Training an adaptation model

- Map event rates collected during sample configurations to predict performance on alternate, untested configurations

\[ u_{IPC}(t) = u_{IPC}(s) \cdot a(e_1(s),\ldots,e_n(s)) + \beta \]

\[ a = \sum_{i=1}^{n} (e_i(s) \cdot x_i + y_i) + z \]

\[ u_{IPC}(t) = u_{IPC}(s) \cdot \sum_{i=1}^{n} (e_i(s) \cdot x_i) + u_{IPC}(s) \cdot \lambda + \beta \]

- Regression, ANN, classifiers (e.g. SVM),...
- Simple models capture scalability curves well, albeit with lower absolute accuracy
Taming locality issues

- Original DCT failed to consider implications of cache refills, thread and data mapping

Example: 45% execution time variation across 85 mappings

4, 4-core nodes: 43,680 mappings.
16, 4-core nodes: 63 million mappings.
1000, 4-core nodes: $10^{43}$ mappings.
DyNUMA training using ANN

Optimize for concurrency, vertical and horizontal locality

[Su et al., IISWC12, SIGMETRICS PER12]
Modeling accuracy

Predicted Wall-clock time

Measured Wall-clock time

Normalized Prediction

Time (second)

Normalized Value (%)
DyNUMA performance

Normalized Time, EDP (%)

wall-clock time | EDP | wall-clock time | EDP | wall-clock time | EDP

Barcelona | Magny-Cours | Tilera64

NPB.FT | NPB.CG | NPB.SP | NPB.BT | NPB.MG | NPB.UA

AMG.Relax | AMG.Matvec

wall-clock;me; EDP

0% 20% 40% 60% 80% 100% 120%
Performance on TilePro64

Tile64Pro OS default Linux mapping is inefficient
More concurrency does not necessary improve performance
Power-aware hybrid parallel programming
Energy-Aware Hybrid Programming

DVFS can save power and thereby energy by reducing time by reducing imbalance.
Critical path based modeling

• Predicting time vs. predicting scaling function

\[ t_i = \sum_{j=1}^{M} \min_{1 \leq \text{thr} \leq X \cdot Y} t_{i,j,\text{thr}} \]

\[ t_c = \max_{1 \leq i \leq N} \sum_{j=1}^{M} \min_{1 \leq \text{thr} \leq X \cdot Y} t_{i,j,\text{thr}} \]
Time modeling enables slack dispersion

- Slack dispersing DCT&DVFS (Li et. al, TPDS13)

Use critical path time to determine slack (essentially imbalance)

\[
\Delta t_i^{slack} = t_c - t_i - t_i^{comm} - t_{dvfs}
\]

Time constraint:

\[
\sum_{1 \leq j \leq M} \Delta t_{ijk} \leq \Delta t_i^{slack}
\]

Energy constraint:

\[
\sum_{1 \leq j \leq M} t_{ijk} f_k \leq t_i f_0
\]
Consistent (or improving) energy savings with weak scaling
Where is the connection?
Energy and the programmer

- Understand performance and power implications:
  - Thread mapping
  - Scheduling
  - Data placement
- Use adaptable, scale-free programming models
  - Phase-aware optimization infeasible otherwise
- Domain-specific knowledge can save energy
  - DAG, dataflow representations have enough information for a first-order approximation of the optimization problem
  - Proactive application-specific program control likely to outperform any system-level approach
Connecting the dots

- Parallel programs are unfortunately not prepared for a power crisis
- Programming models and runtime systems still are not (yet) up for the task
  - Adapt concurrency at runtime on non-coherent many-core architectures?
  - Control power states within the runtime with input from the application?
  - Scale-free, scale-aware?
  - How much waste is there in the runtime and the OS?
We educate two kinds of (parallel) programmers

Conclusive proof it pays more to do nothing than it does to get a Ph.D.:

<table>
<thead>
<tr>
<th>Average Maximum Annual Unemployment Benefit</th>
<th>Average Graduate Student Stipend</th>
</tr>
</thead>
<tbody>
<tr>
<td>$21,060</td>
<td>$18,779</td>
</tr>
</tbody>
</table>

Sources: U.S. Department of Labor (via SF Chronicle), The Chronicle of Higher Education 2008-2009 survey of pay and benefits for teaching and research assistants. Unemployment benefits computed from average maximum state weekly benefits (typically 50% of base wages, capped by state) multiplied by 52 (in some cases, benefits can be extended up to 79 weeks). Academic year stipends extrapolated to 12 months.

www.phdcomics.com
We educate two kinds of (parallel) programmers

Joe the Plumber
Programmer

Joe's story is an American tale. He's a plumber who became the CEO of a company that cleans up the mess of an American cookie. As Joe said, "As a plumber, I could be a great American."

Fighting for the American Dream
Samuel J. Wurzelbacher with Thomas N. Tabback

Friday, March 1, 13

EuroMicro PDP'13 Keynote
Revisiting parallel programming

- Environmentally conscious programmers know the hard limits of their algorithms given architectural or power constraints (e.g. DAG, roofline models)
- Environmentally conscious runtime systems eliminate waste
  - No more runtimes antagonizing operating systems
  - No more cross-runtime system noise
Revisiting parallel programming

• Your power budget may well be your most critical resource
  – Awareness of this issue must somehow become part of programmer education
  – Neither Joe nor graduate students are ready for this
  – Hardware and system software will not solve the problem by themselves, without a steep performance penalty

• Defensive parallel programming
  – Scale-free, react to power caps
  – Explicit resource proportioning
More information

http://www.qub.ac.uk/research-centres/HPDC/
A power-aware parallel program of the future:
- adapts concurrency
- eliminates waste (idle time, redundant communication)
- controls component power at a fine granularity (e.g. per task)

Manousakis et al, [SBAC-PAD12]