Hardware/algorithm codesign for energy-efficient scientific computing

Markus Geveler

@ PACO17 Tegernsee
Markus @ TUDo

www.mathematik.tu-dortmund.de/lsiii/_Markus_Geveler
markus.geveler@math.tu-dortmund.de

Energy-efficient simulations,
Hardware-oriented numerics,
Multilevel domain decomposition solvers

@ TU Dortmund:
Markus Geveler, Stefan Turek,
Dirk Ribbrock, Hannes Ruelmann,
Daniel Donner

@ MPI Magdeburg
Peter Benner, Jens Saak,
Martin Köhler, Gerry Truschkewitz

Insular
Compute center
Applied Sciences
Renewables
Unconventional
System Integration
Where everything is leading: simulation of technical flow

- **Multiphysics**: increasing complexity of problems and thus methods
- **Models are versatile**: applications in many fields
- **Discrete continuum mechanics**: determine physical quantities for a huge number of points in space and repeat very often
- **Ressource-hungry**: memory, time, energy
- **Methods**: DD Newton-Krylov-Multigrid schemes
1 Unconventional hardware for scientific computing

ARM-based SoCs with mobile GPUs plus Photovoltaic in a (big) box
**History: FEAT and low power SoCs: embedded v commodity**

- **2012**: FEAT family one of the first frameworks to run applications on the then new Tibidabo cluster (Montblanc project) @ BSC
  - exploration of energy savings / performance tradeoffs using ARM processors on Tegra 2 SoCs (Cortex A9 only)
  - results were very promising

- **2013**: ICARUS accepted, Tegra K1 announced (GPU!)

- **2013 - 2014**: preliminary experiments with Tegra 3, Tegra 4, Tegra K1 for ICARUS

- **2015**: small cluster with 4 nodes @ PACO15

- **2015 - march 2016**: ICARUS construction
Applications on ARM Cortex A15

Global ocean circulation simulation, saline transport, 3D, DG

<table>
<thead>
<tr>
<th>$k$</th>
<th>$T$ [s]</th>
<th>$P$ [W]</th>
<th>$E$ [J]</th>
<th>$\Delta T$ [s]</th>
<th>$\Delta P$ [W]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Haswell</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.0768</td>
<td>65.6</td>
<td>5.038</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>2</td>
<td>0.0434</td>
<td>80.7</td>
<td>3.502</td>
<td>0.0334</td>
<td>15.1000</td>
</tr>
<tr>
<td>3</td>
<td>0.0352</td>
<td>91.4</td>
<td>3.217</td>
<td>0.0082</td>
<td>10.7000</td>
</tr>
<tr>
<td>4</td>
<td>0.0316</td>
<td>100.3</td>
<td>3.169</td>
<td>0.0036</td>
<td>8.9000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Cortex-A15</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.477</td>
<td>7.5</td>
<td>3.577</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>2</td>
<td>0.249</td>
<td>10</td>
<td>2.49</td>
<td>0.228</td>
<td>2.5000</td>
</tr>
<tr>
<td>3</td>
<td>0.173</td>
<td>12</td>
<td>2.076</td>
<td>0.076</td>
<td>2.0000</td>
</tr>
<tr>
<td>4</td>
<td>0.170</td>
<td>13.9</td>
<td>2.363</td>
<td>0.003</td>
<td>1.9000</td>
</tr>
</tbody>
</table>
Computers and energy consumption: a (pessimistic) forecast

![Graph showing energy consumption forecast](image)

- World energy production
- Today's systems consumption
- Future systems consumption
- Lower bound systems consumption (Landauer Limit)

- Smaller transistors, ...

- E [J/year]
- 2015, 2035, 2040

SIA 2015
Conventional approaches are running into physical limits. Reducing the 'energy cost' of managing data on-chip requires coordinated research in new materials, devices, and architectures.

- the Semiconductor Industry Association (SIA), 2015
The device level is not the end of the story! What can Scientific Computing do?

- integrated renewable power source
- larger energy efficiency of hardware,
- better methods that fit to that hardware
### Technology revolutions we can use in scientific computing

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Digital computing / signal processing / embedded systems</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Internet</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mobile Internet</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mobile Computing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>3. Industry revolution (Automation, IT, FabLab)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Energy (renewables, better nets and consumers)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Traffic (electro)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| ? | | | | |

Single devices, clusters and energy efficiency

- **mobile processors**: less power, less performance than commodity (x86, ..., desktop GPUs)
- **same task, more devices of same type**: more power, more performance
- Can we scale to same performance and spend less?
- **scaling and energy**: less execution time, same energy (because $E$ is $P$ integrated in time)

For two computer architectures A and B must hold: If A is more energy-efficient than B in executing a task then the powerdown from A to B executing this task must be larger than the respective speeddown.
A performance model for a low energy cluster (dropping infrastructure)

- expected scaling penalty to power (switches)
- expected parallelisation penalty to performance (communication, numerics)

\[
E = \left( n_{\text{nodes}} P_{\text{node}} + \left[ \frac{n_{\text{nodes}} - 1}{n_{\text{ports}}} \right] \frac{t_{n_{\text{nodes}}=1}}{n_{\text{nodes}}} + t_{\text{overhead}}(n_{\text{nodes}}) \right)
\]

\[
E_{A,B} = \frac{\Delta P_{A,B}}{\Delta t_{A,B}}
\]
Hardware-oriented numerics and energy efficiency

How hardware and numerics determine energy to solution in (FEM-based) technical flow simulation
Hardware efficiency and performance engineering for technical flow simulation

- most of our codes are memory bandwidth-bound
- proper exploitation of SIMD is key to single core performance. Often: optimised SpMV.
- the memory interface is saturated with a (small) amount of cores
- GPGPU usually gives us a speedup of 5 - 10 through larger on-chip memory bandwidth. GPUs can also saturate that bandwidth.
- mixed precision provides another x1.5 max sustainable (double-single). More possible (double-half)
- low precision: with some methods
- baseline power of all devices has to be amortized via hybrid computation with careful load balancing
Numerical efficiency and performance engineering for technical flow simulation

- clever smoother construction example: SPAI-types
- in theory: SPAI with same structure as A gives convergence rates like GS (SPAI-1)
- works very well as MG smoother
- construction phase: different ways, we make progress (next)
- application phase: SpMV

\[
x^{k+1} \leftarrow x^k + \omega M(b - Ax^k)
\]

\[
\| I - MA \|_F^2 = \sum_{k=1}^{n} \| e_k^T - m_k^T A \|_2^2 = \sum_{k=1}^{n} \| A^T m_k - e_k \|_2^2
\]

\[\min_{m_k} \| A^T m_k - e_k \|_2, \ k = 1, \ldots, n.\]
Numerical efficiency: recent SPAI preconditioner results from EXA-DUNE project

- SPAI is exceptionally adaptable
- allows for good balancing of **effort/energy to effectivity** of preconditioner/smoothers
- high **reuse potential** of once created approximate inverse
- many screws to adapt to hardware (assembly stage)
  - predefined sparsity pattern (SPAI-1)
  - refinement of sparsity pattern
  - refinement of coefficients
  - rough inverses often good enough
  - (half precision, use Machine Learning / Interpolation in knowledgebase)
Energy efficiency and performance engineering for technical flow simulation

- Influences on incore performance of technical flow simulation (hardware efficiency):
  - FEM space(s)
  - mesh adjacencies (fully unstructured)
  - DOF numbering
  - matrix storage (SELL)
  - accuracy (mixed, low)
  - assembly of matrices (SPAI methods)

- Influences on incore performance of technical flow simulation (numerical efficiency):
  - assembly of matrices (SPAI-order)
  - solver scheme
  - preconditioners / smoothers
Building an insular compute center with a low energy cluster

And actually running it
ICARUS
Modules

- power supply
- housing and climate control
- energy storage, converters
- networking
- low energy cluster
- data storage
- data
Tegra K1 SoC

History of embedded/mobile processors is different than commodity archs’

32 Bit architecture,
4 x Cortex-A15 CPU,
programmable Kepler GPU,
LPDDR3,
high SP Performance,
most energy-efficient SoC of its time
nowadays: Tegra X1, Tegra X2, ...

Jetson TK1 carrier board

Everything we need in a compute node on a single carrier board

Tegra K1 Carrier Board.
P = 10 - 15W incl. fan (overkill),
GiBit Ethernet,
much I/O: serial, USB, SATA
Linux, CUDA,
nowadays: Jetson TX1, Jetson TX2
Energy supply & storage

Challenges: area, weather, operation at night

7.5 kWp Photovoltaik, freeland
2.5m x 16m,
8kWh Li-ion battery,
2+1 inverter, charge
management
Aim: Full operation at daytime,
mild operation at night (3 seasons), just stay alive (winter)

Housing, cooling, heating

Challenges: isolation, area, ventilation, (cooling, heating)

Custom design modified high cube cargo container
90mm isolation, fireproof,
steel safety doors,
heatable ventilation with separate power supply.
Cluster, networking, data storage

Challenges: space usage, avoid heat nests

60 x NVIDIA Jetson TK1, 240 ARM Cortex-A15 cores, 60 Kepler GPUs, 120 GB RAM (LPDDR3), 3 PDUs, sensors, 3 + 1 low power switches (Netgear)

correct power dissipation (peak): \(~ 1kW\)

theoretical peak performance: \(~ 20 \text{TFlop/s} mixed precision,\)

1 rack skeleton, many standoffs, cables, 3D-printed custom parts

Energy savings in the data storage system

Fully portable, BananaPi-based

10 x 1 TByte SSDs, redundant, e.g.: 5 TByte usable

Peak sustainable speed:

90 MByte/s (for comparison: 140 with commodity) \(\rightarrow \text{Speeddown} = 1.6\)

Average power:

30W (for comparison: 500W with commodity) \(\rightarrow \text{Powerdown} = 16.8\)
Benchmarks:
basic kernels, single node

Compute-bound

- ▲ commodity(2015)
- ◇ commodity(2012 desktop)
- □ Jetson TK1(2014)
- ◇ commodity(2012 compute)
Benchmarks:
basic kernels, single node

Memory bandwidth-bound

Applications on ARM Cortex A15

Global ocean circulation simulation on ARM Cortex A15
Benchmarks: applications, full cluster, GPU, CPU

- scaling induces power penalties
- parallelisation induces performance penalties
- all compared architectures beaten nonetheless

![CFD simulation diagram]

---

- ▲ commodity(2015)
- ◇ commodity(2012 desktop)
- □ Jetson TK1(2014)
- ▽ commodity(2012 compute)
Photovoltaic subsystem results

![Graph showing power consumption and battery charge over time](image)
Uptime
Since spring 2016: Even at high stress (weather): no long downtimes in spring/summer autumn, mild slumber (16/60 cores online) in winter.

Climate in the container
On hot days:
Max 33 °C ambient,
Max 68 °C Jetson boards,
Humidity 50%
Aux power: 3 x 100W

Total cost
84k €

Devel time
3 years
Lessons learned

Battery temperature
On cold days: hard to hold battery warm

Cost
May be reduced significantly without all the trial-and-error

Hardware market dynamics
...are fast! (next)

Effort
Very high

Unconventional
Preparedness of institutions?

Tegra K1
small memory size, bandwidth, controller,
(much better in later versions)
'Mobile/embedded' is becoming more 'multi-purpose'.

All compute hardware is becoming more energy-efficient.

When will the employment of a newer architecture pay off?

How much energy can we save during that time?

Who will win the race? Will there be convergence?

For which kind of codes does this pay off?
Next?

- autotuning for parameter settings (Exa-DUNE)
- better SPAI-eps and SAINV (Exa-DUNE)
- exploit Machine Learning more
- spread the lore:
  Taking control of energy-consumption can make a huge difference
Thanks!

This work has been supported in part by the German Research Foundation (DFG) through the Priority Program 1648 'Software for Exascale Computing'.

ICARUS hardware is financed by MIWF NRW under the lead of MERCUR.
(Super-)Computers and power dissipation

(when we thought ICARUS in 2014)

<table>
<thead>
<tr>
<th>Green500 rank</th>
<th>architecture</th>
<th>Top500 rank</th>
<th>Performance / P [GFlop/s / W]</th>
<th>P [MW]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Xeon + FirePro</td>
<td>168</td>
<td>5.2</td>
<td>0.057</td>
</tr>
<tr>
<td>2</td>
<td>Xeon + PEZY</td>
<td>369</td>
<td>4.9</td>
<td>0.037</td>
</tr>
<tr>
<td>3</td>
<td>Xeon + K20x</td>
<td>392</td>
<td>4.4</td>
<td>0.035</td>
</tr>
<tr>
<td>4</td>
<td>Xeon + K40m</td>
<td>361</td>
<td>4.0</td>
<td>0.044</td>
</tr>
</tbody>
</table>

- **no Top500 topscorer** (Top500#1: P=18)
- **Xeon CPU:** commodity
- **GPUs (Fire,Kxx):** also commodity
- **PEZY CPU:** unconventional

www.top500.org/green500/list/2014/11
## (Super-)Computers and power dissipation (2016)

<table>
<thead>
<tr>
<th>Green500 rank</th>
<th>architecture</th>
<th>Top500 rank</th>
<th>Performance / P [GFlop/s / W]</th>
<th>P [MW]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Xeon + P100</td>
<td>28</td>
<td>9.5</td>
<td>0.350</td>
</tr>
<tr>
<td>2</td>
<td>Xeon + P100</td>
<td>8</td>
<td>7.5</td>
<td>1.310</td>
</tr>
<tr>
<td>3</td>
<td>Xeon + PEZY</td>
<td>116</td>
<td>6.7</td>
<td>0.150</td>
</tr>
<tr>
<td>4</td>
<td>Sunway</td>
<td>1</td>
<td>6.0</td>
<td>15.370</td>
</tr>
</tbody>
</table>

- **Perf./P: x1.5!**
- **Sunway**: also unconventional and Top500#1!
- **Powerdown Top500#1: x0.8**

[www.top500.org/green500/list/2016/11](http://www.top500.org/green500/list/2016/11)
### (Super-)Computers and power dissipation

#### (2017)

<table>
<thead>
<tr>
<th>Green500 rank</th>
<th>architecture</th>
<th>Top500 rank</th>
<th>Performance / P [GFlop/s / W]</th>
<th>P [MW]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Xeon + P100</td>
<td>61</td>
<td>14</td>
<td>0.14</td>
</tr>
<tr>
<td>2</td>
<td>Xeon + P100</td>
<td>465</td>
<td>14</td>
<td>0.03</td>
</tr>
<tr>
<td>3</td>
<td>Xeon + P100</td>
<td>148</td>
<td>12</td>
<td>0.08</td>
</tr>
<tr>
<td>4</td>
<td>Xeon + P100</td>
<td>305</td>
<td>10</td>
<td>0.06</td>
</tr>
</tbody>
</table>

- Perf./P.: x1.5
- PEZY, Sunway systems: now ranking 7, 15, 17

[www.top500.org/green500/list/2017/07](http://www.top500.org/green500/list/2017/07)