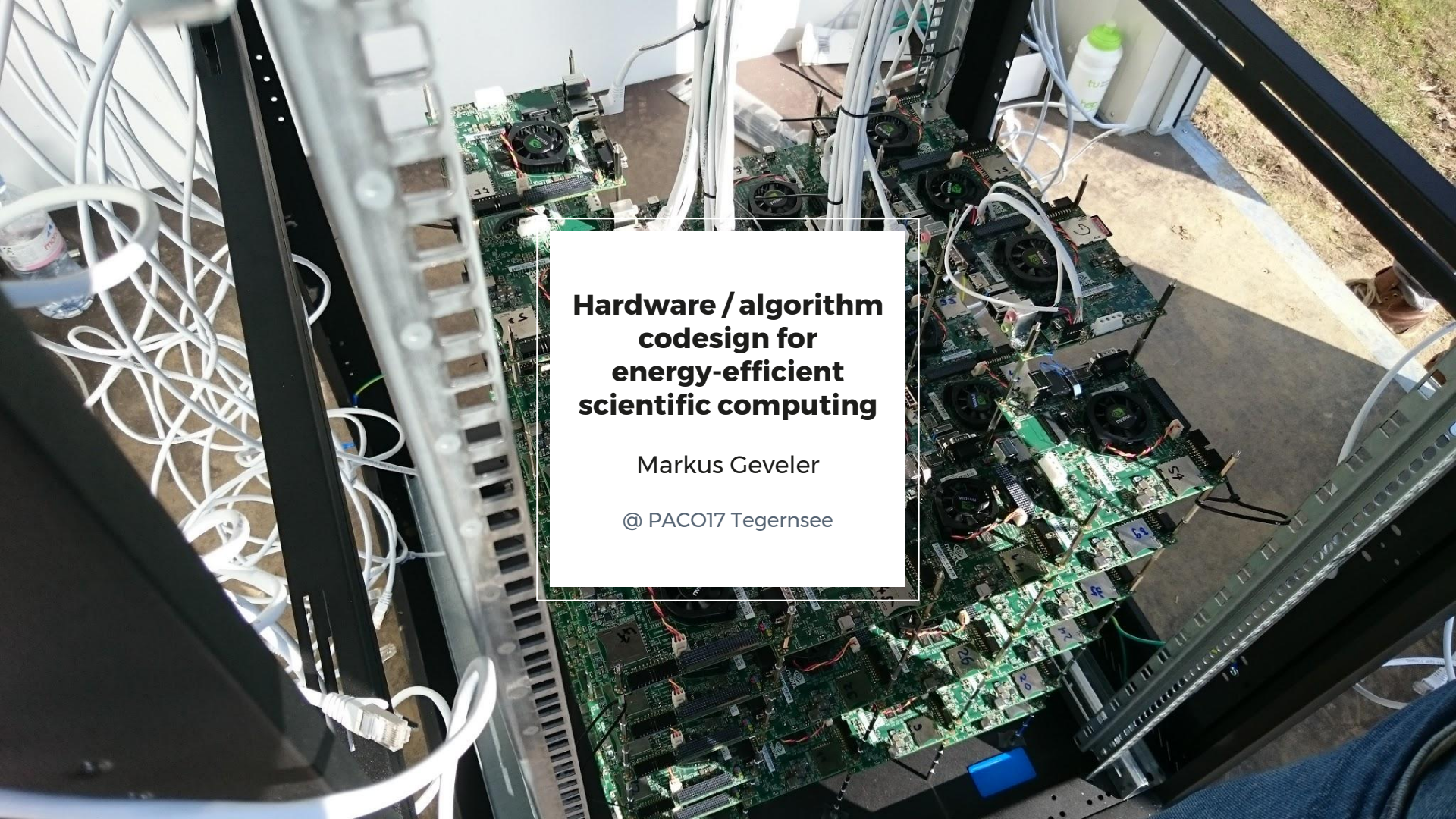
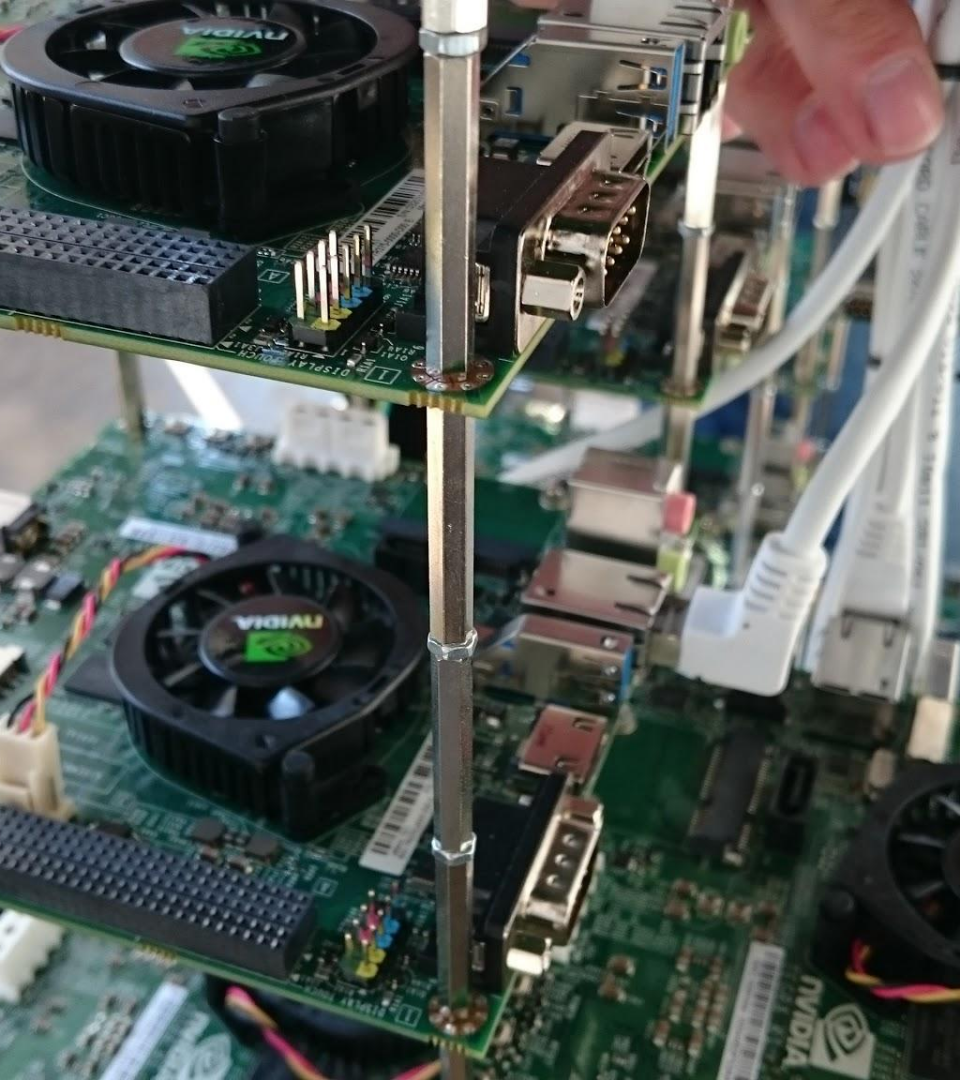


Hardware / algorithm codesign for energy-efficient scientific computing

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

technische universität dortmund

fakultät für mathematik

FEAT

MERCUR Mercator Research Center Ruhr

I.C.A.R.U.S.

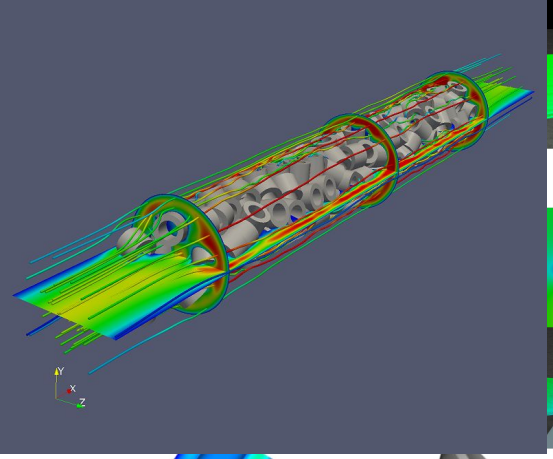
Ministerium für Innovation, Wissenschaft und Forschung des Landes Nordrhein-Westfalen

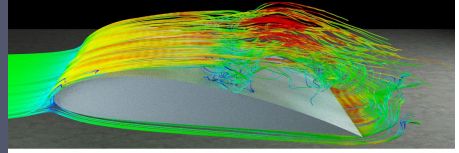
www.icarus-green-hpc.org

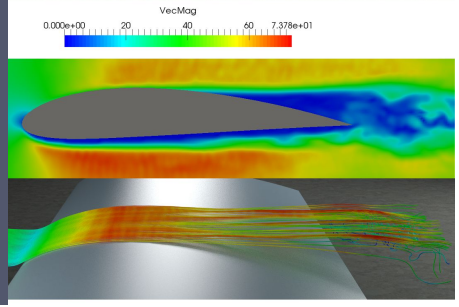
Insular
Compute center
Applied Sciences
Renewables
Unconventional
System Integration

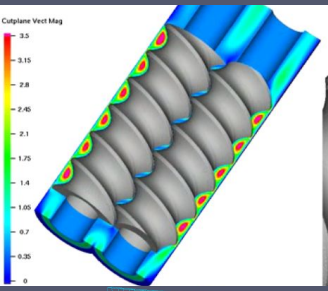
@ TU Dortmund:
Markus Geveler, Stefan Turek, Dirk Ribbrock, Hannes Ruelmann, Daniel Donner

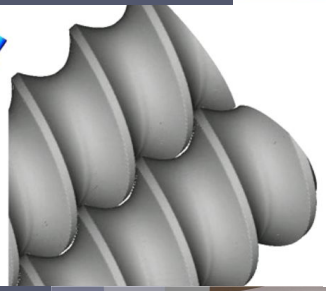
@ MPI Magdeburg
Peter Benner, Jens Saak, Martin Köhler, Gerry Truschkewitz

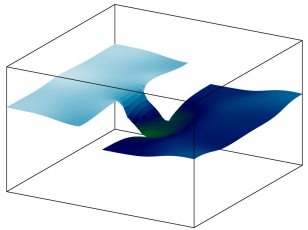


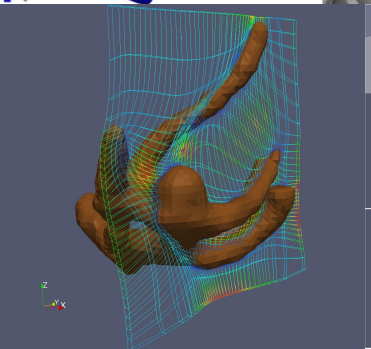


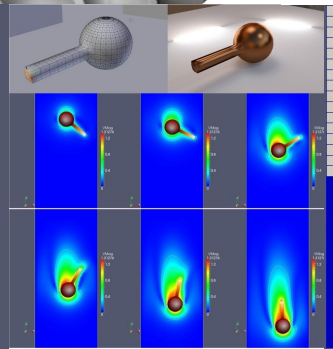


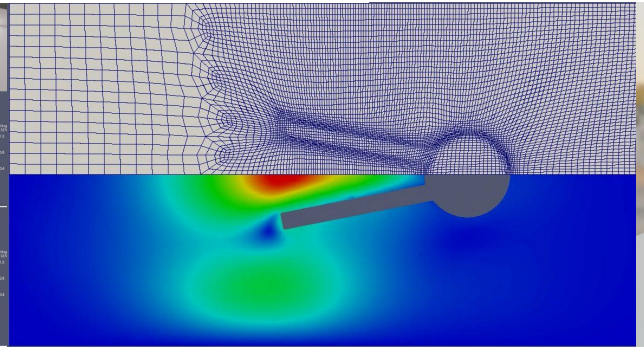










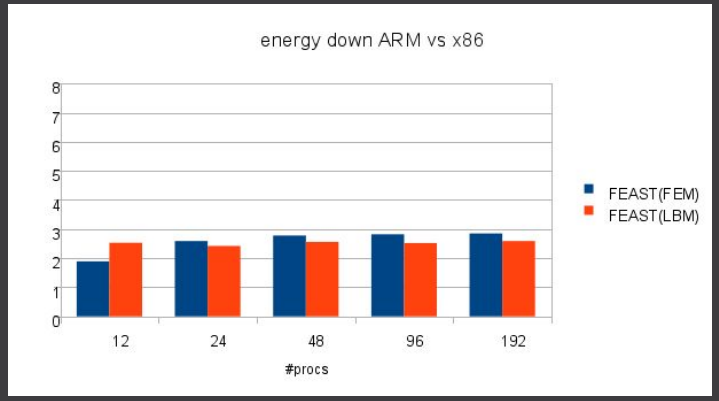


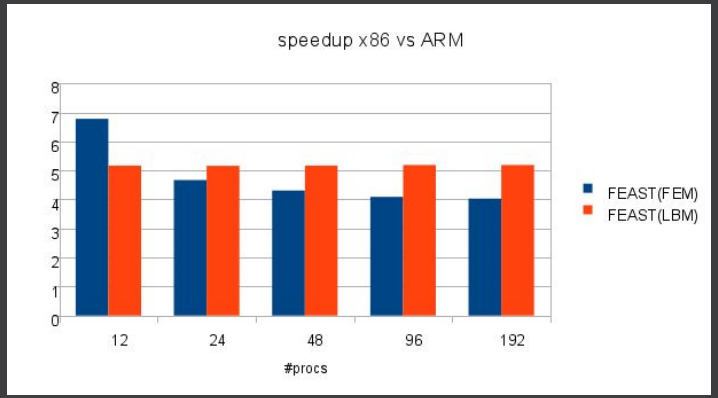
Where everything is leading: simulation of technical flow

- **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

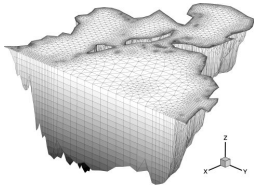




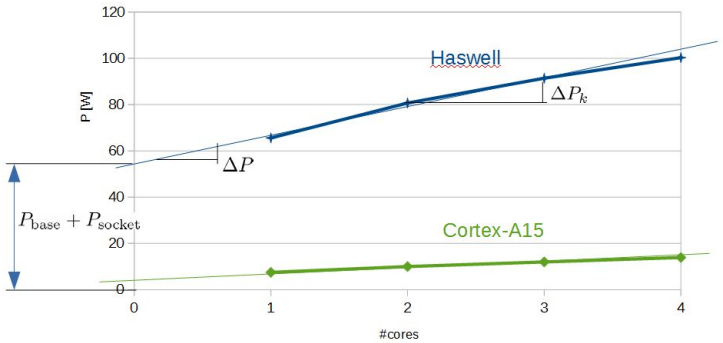
History: FEAT and low power SoCs: embedded v commodity

- **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

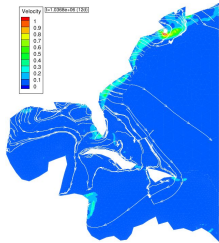
Applications on ARM Cortex A15

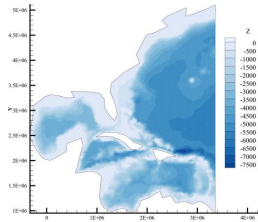


Global ocean circulation simulation, saline transport, 3D, DG



k	T [s]	P [W]	E [J]	ΔT [s]	ΔP [W]
			Haswell		
1	0.0768	65.6	5.038	–	–
2	0.0434	80.7	3.502	0.0334	15.1000
3	0.0352	91.4	3.217	0.0082	10.7000
4	0.0316	100.3	3.169	0.0036	8.9000
			Cortex-A15		
1	0.477	7.5	3.577	–	–
2	0.249	10	2.49	0.228	2.5000
3	0.173	12	2.076	0.076	2.0000
4	0.170	13.9	2.363	0.003	1.9000



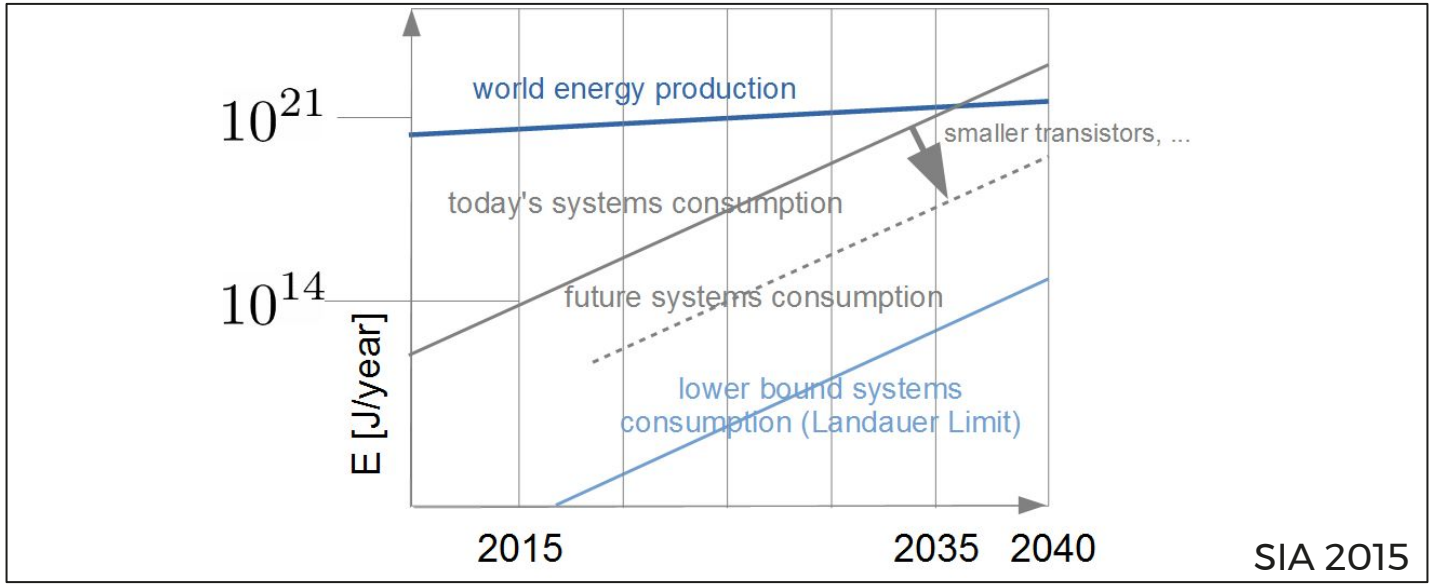


University of Stuttgart Germany
SimTech Cluster of Excellence

technische universität dortmund
fakultät für mathematik

FAU Friedrich-Alexander Universität Erlangen-Nürnberg Naturwissenschaftliche
Department Mathematik

Computers and energy consumption: a (pessimistic) forecast



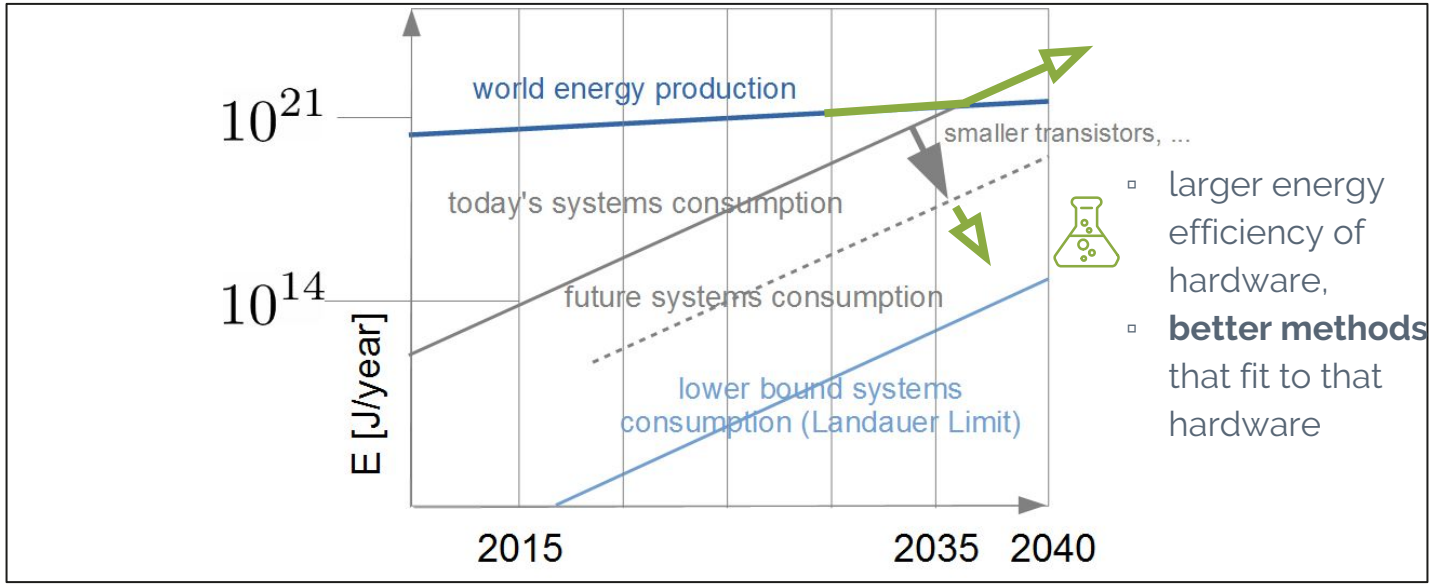
“

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

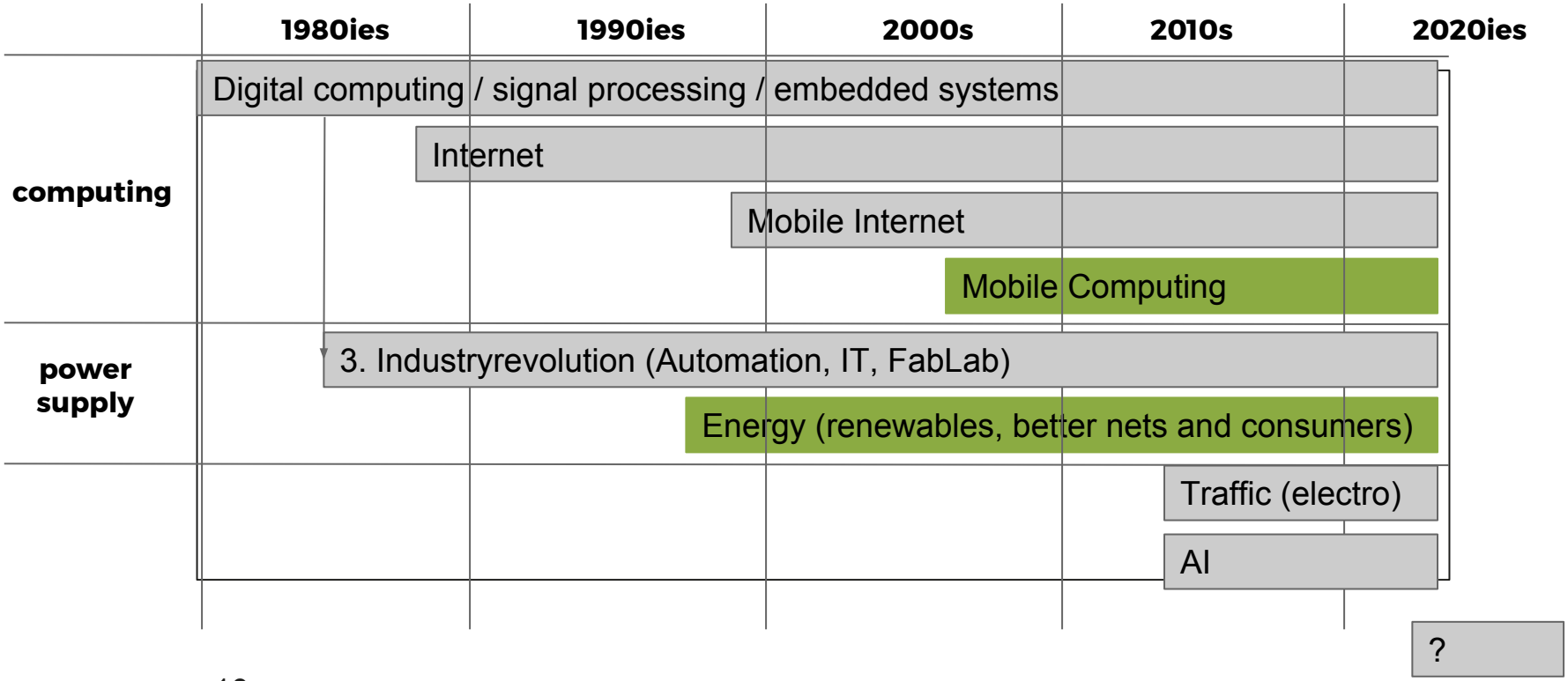
The device level is not the end of the story! What can Scientific Computing do?

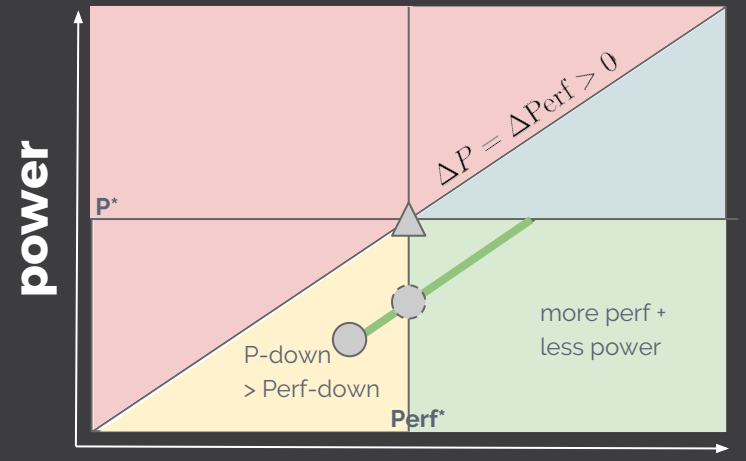
- integrated renewable power source

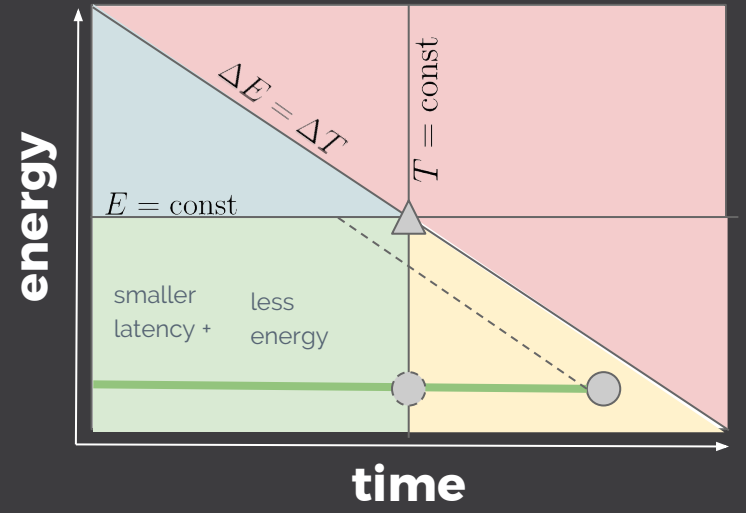


- larger energy efficiency of hardware,
- **better methods** that fit to that hardware

Technology revolutions we can use in scientific computing



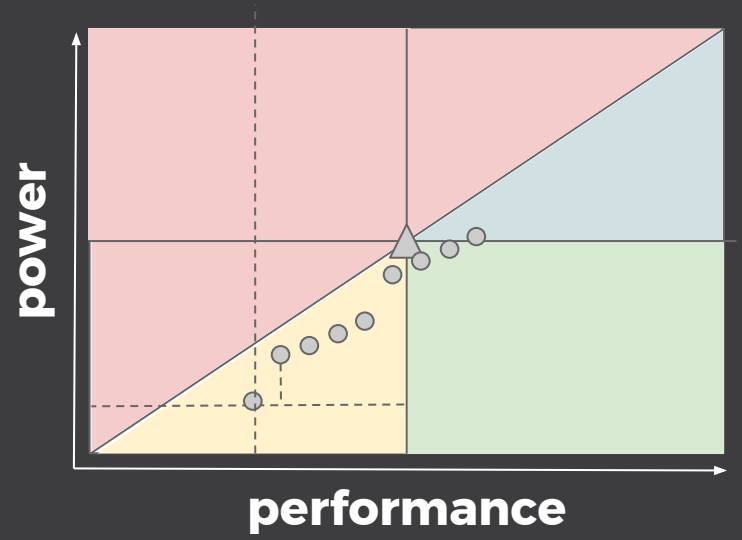


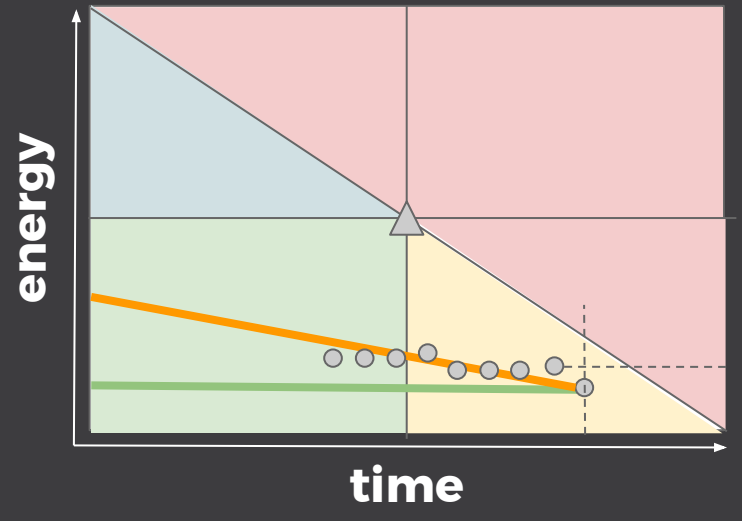


Single devices, clusters and energy efficiency

- **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)

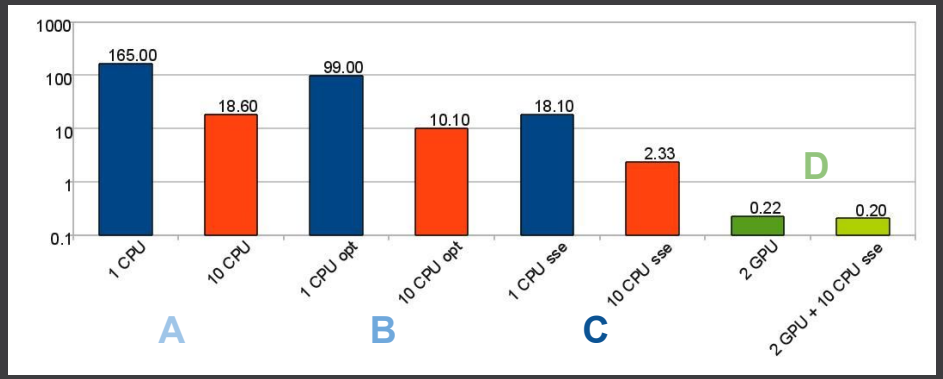
- **expected scaling penalty** to power (switches)
- **expected parallelisation penalty** to performance (communication, numerics)

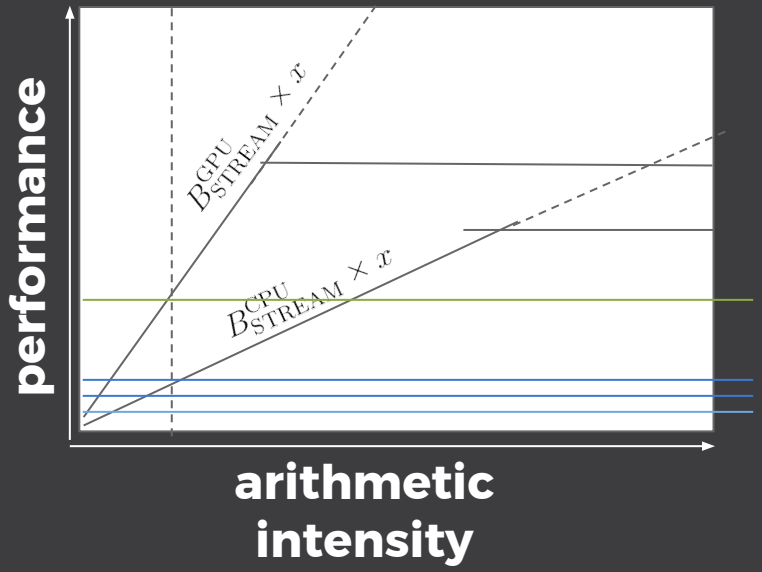
$$E = \left(n_{\text{nodes}} P_{\text{node}} + \left\lceil \frac{n_{\text{nodes}} - 1}{n_{\text{ports}}} \right\rceil \right) \left(\frac{t_{n_{\text{nodes}}=1}}{n_{\text{nodes}}} + t_{\text{overhead}}(n_{\text{nodes}}) \right)$$

$$E_{A,B} = \frac{\Delta P_{A,B}}{\Delta t_{A,B}}$$

2
Hardware-oriented numerics and energy efficiency

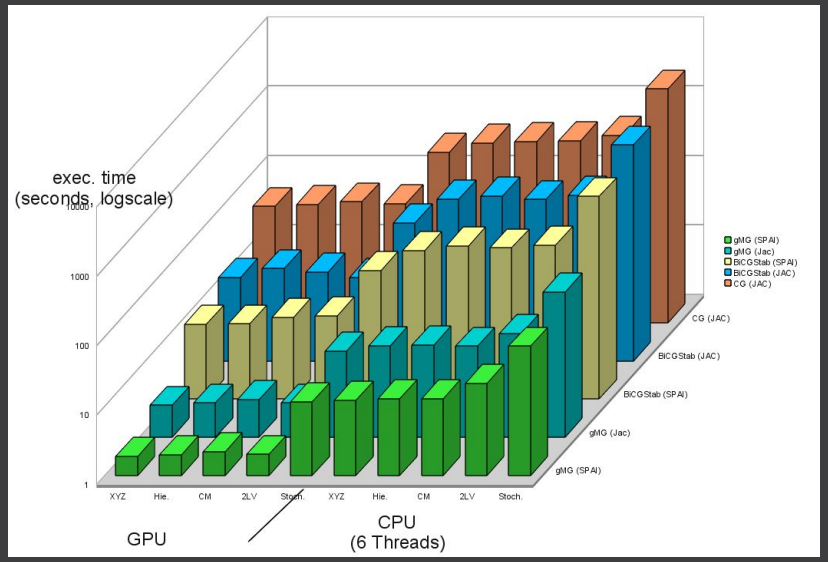
How hardware and numerics determine energy to solution in (FEM-based) technical flow simulation

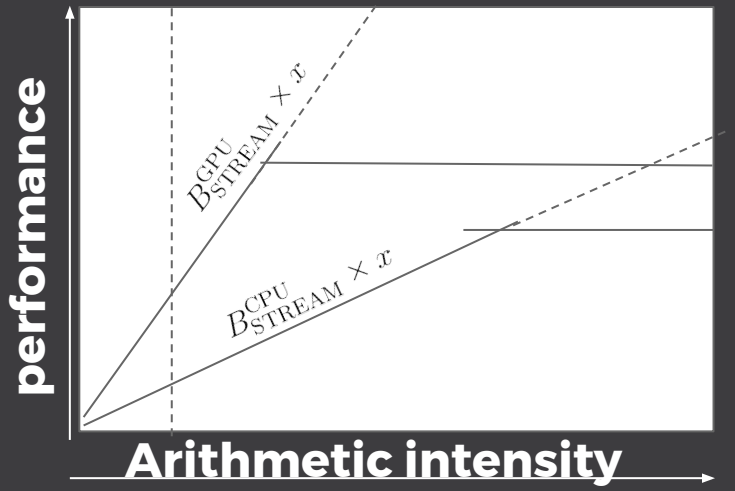




Hardware efficiency and performance engineering for technical flow simulation

- 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**





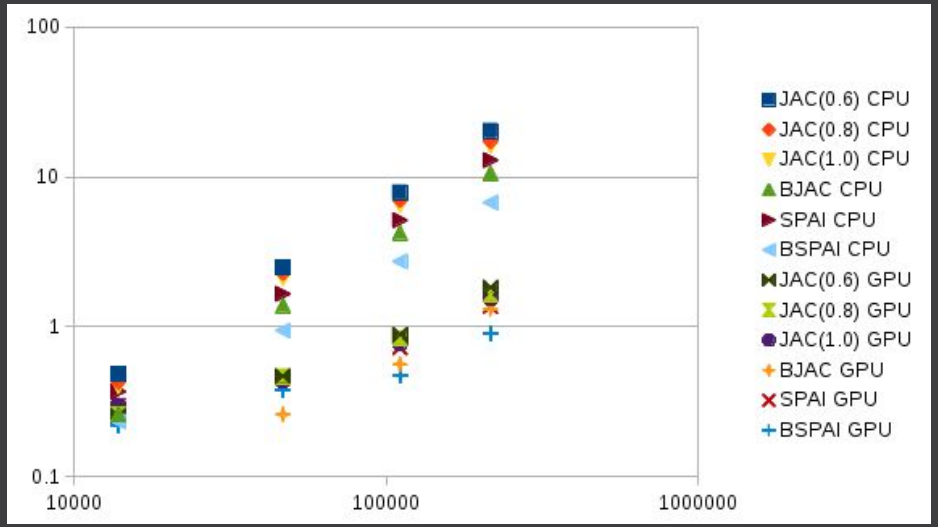
Numerical efficiency and performance engineering for technical flow simulation

- 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**

$$\mathbf{x}^{k+1} \leftarrow \mathbf{x}^k + \omega M(\mathbf{b} - A\mathbf{x}^k)$$

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

$$min_{m_k} \| A^{\mathrm{T}} m_k - e_k \|_2, \quad k = 1, \ldots n.$$



Many Householder transforms -based QR decompositions, Batched on GPU

cuBLAS, cuSPARSE and MKL

~4x GPU v CPU

Partial randomization via Markov Chains

$$m_{ij} = \lim_{N\to\infty} \frac{1}{N} \sum_{s=1}^{N} \left(\sum_{q=0}^{\infty} \frac{W_q}{A_{ii}} \delta_{s_q j} \right)$$

CUDA and MKL

~5x GPU v CPU

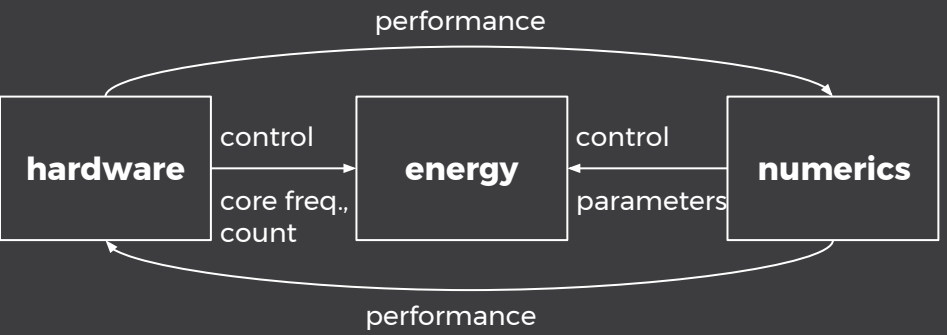
Treat matrices as discrete function in a function regression, train a Deep Learning Net with pairs (A, A^{-1}) make inference for M

first results with fixed sparsity, implementation with TensorFlow CPU and GPU

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/smoother
- 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)

SPPEXA

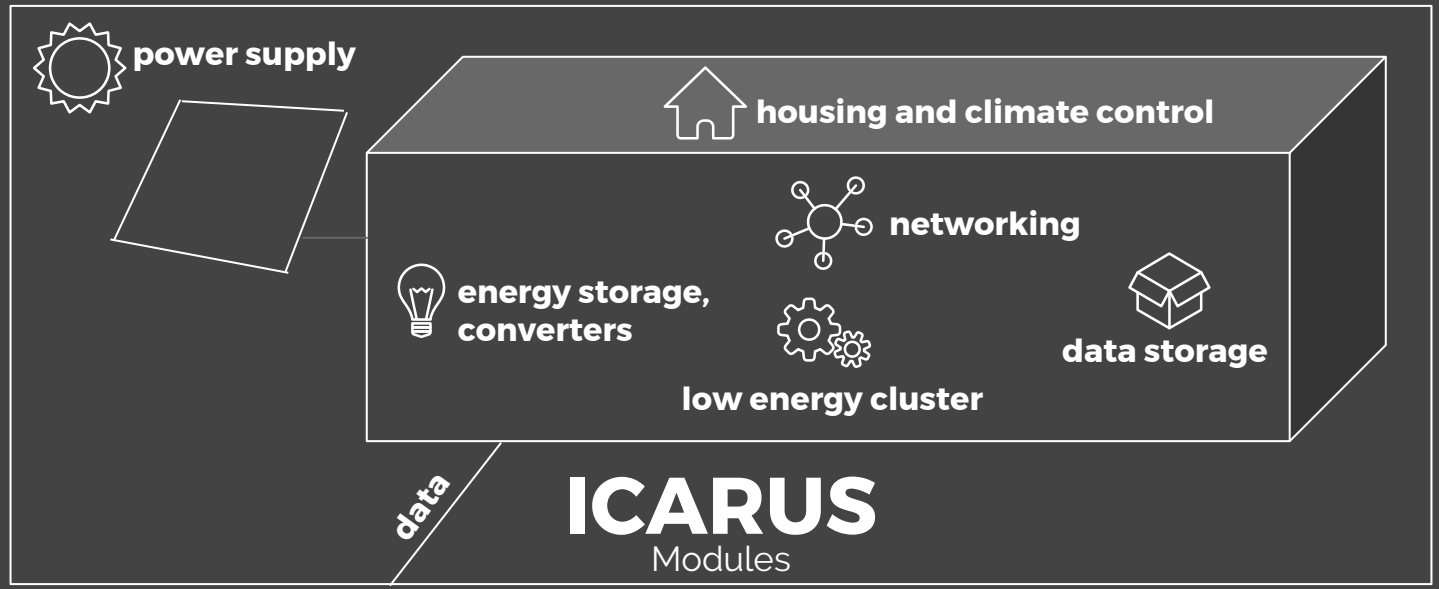


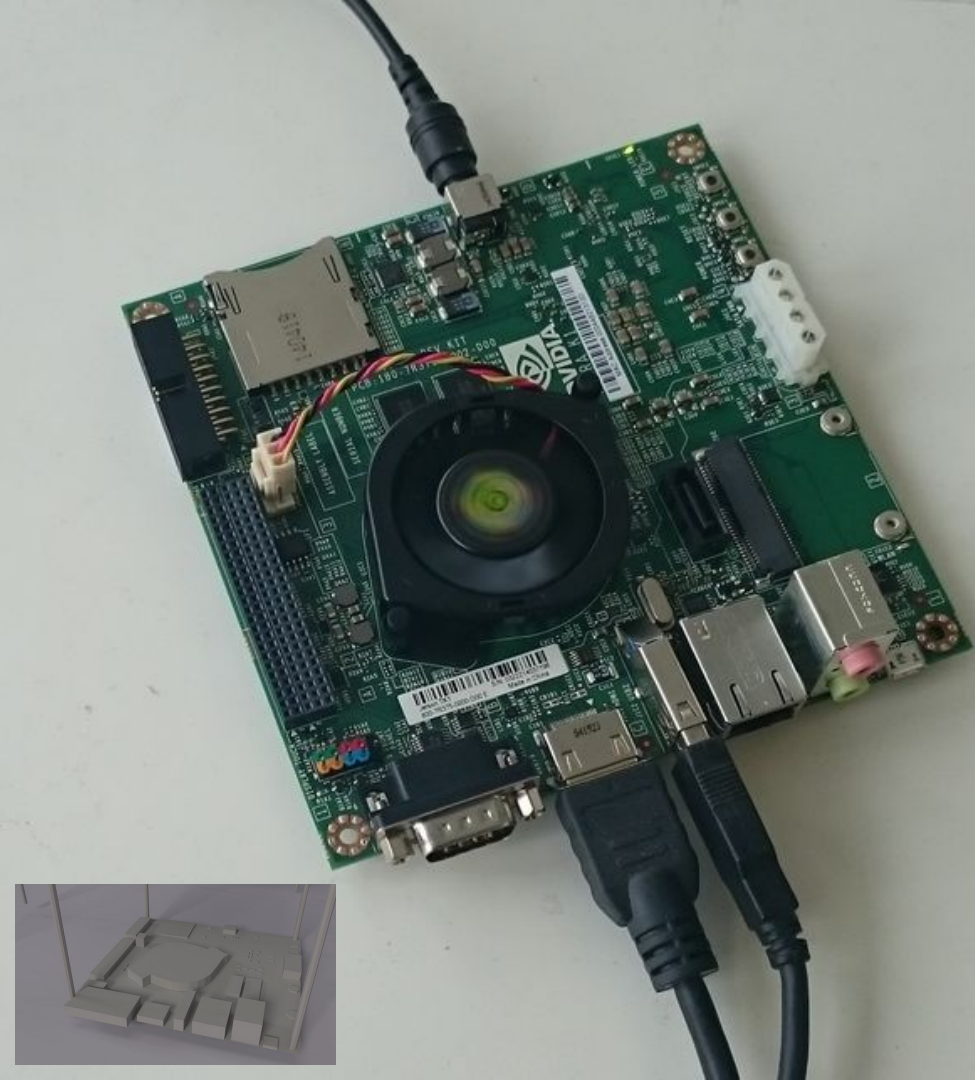
Energy efficiency and performance engineering for technical flow simulation

- 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

3
Building an insular compute center with a low energy cluster

And actually running it


power supply
housing and climate control
networking
energy storage,
converters
data storage
low energy cluster
data
ICARUS
Modules



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, ...

$$\frac{\mathrm{Perf}_{\mathrm{peak}}^{\mathrm{TegraK1}}}{P_{\mathrm{peak}}^{\mathrm{TegraK1}}} = \frac{35\mathrm{GFlop}/s}{W}$$

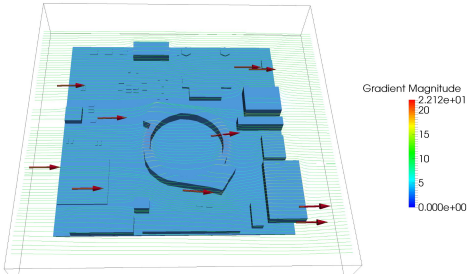
$$\frac{\mathrm{Perf}_{\mathrm{peak}}^{\mathrm{Ref}}}{P_{\mathrm{peak}}^{\mathrm{Ref}}} = \frac{11\mathrm{GFlop}/s}{W}$$

Ref.: Xeon + K40

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

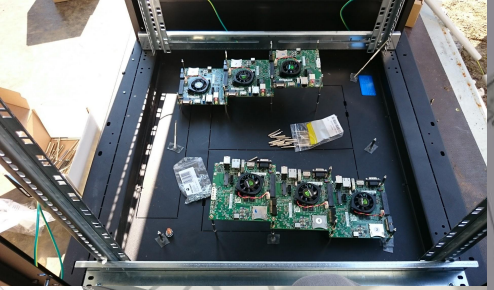
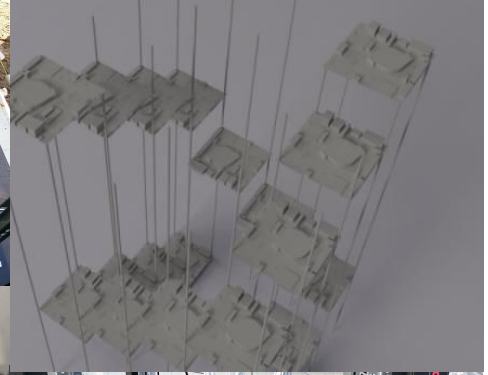
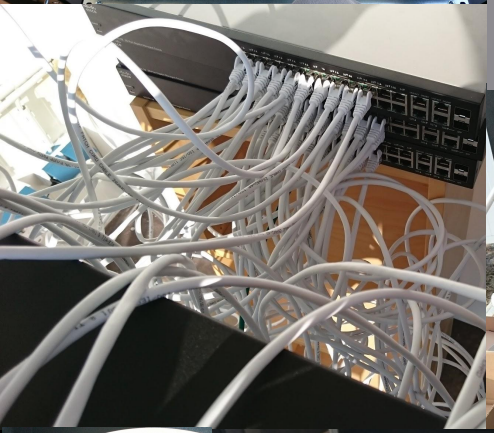
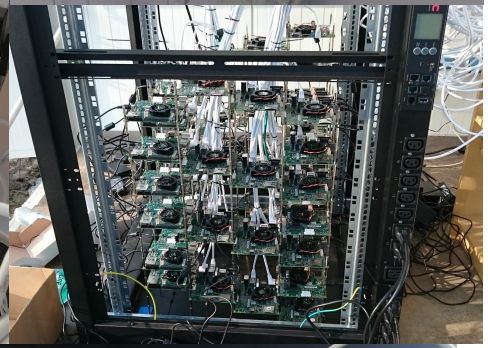
Challenges: area, weather, operation at night

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)
power dissipation (peak): ~ 1kW
theoretical peak performance: ~ 20 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

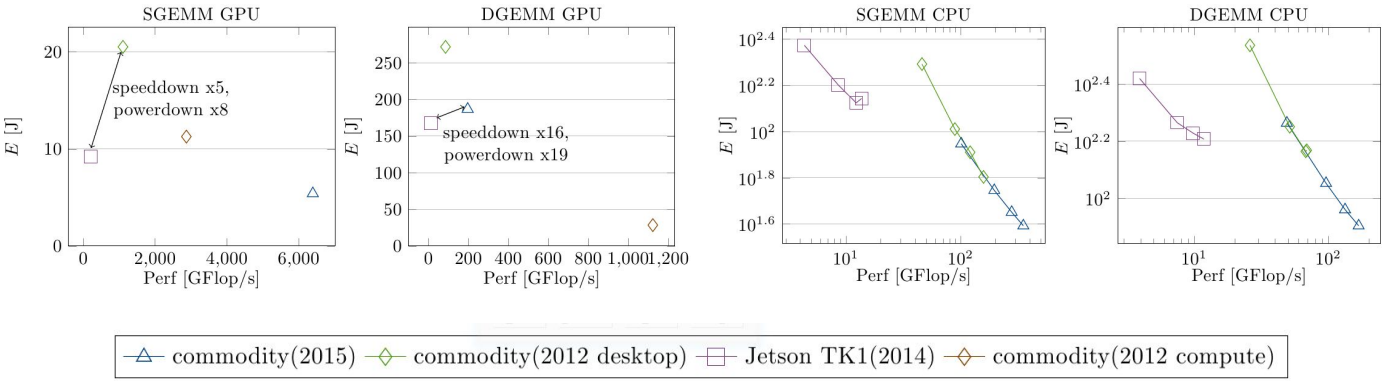
MAX PLANCK INSTITUTE
FOR DYNAMICS OF COMPLEX
TECHNICAL SYSTEMS
MAGDEBURG

Peak sustainable speed:
90 MByte/s (for comparison: 140 with commodity) -> **Speeddown = 1,6**

Average power: 30W (for comparison: 500W with commodity) ->
Powerdown = 16,8

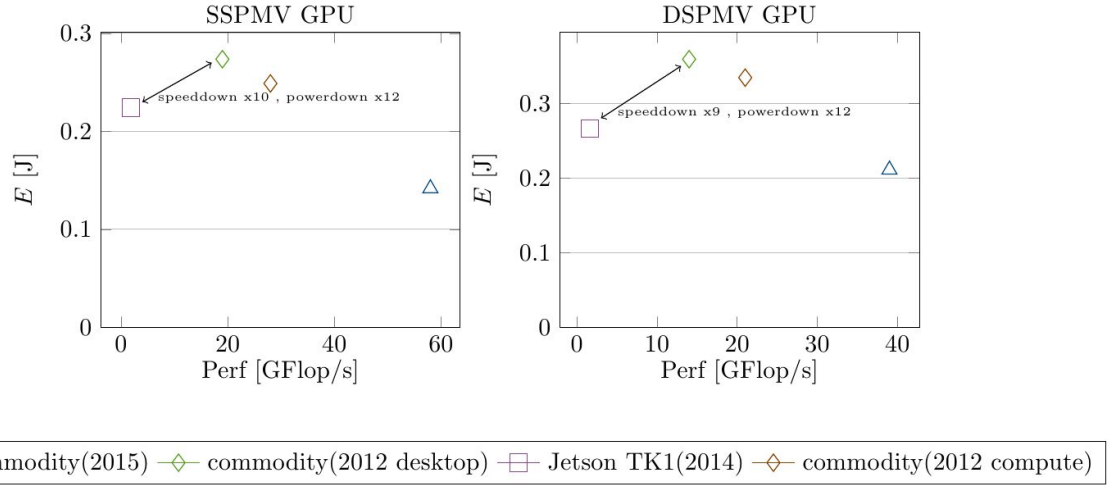
Benchmarks:
basic kernels, single node

Compute-bound



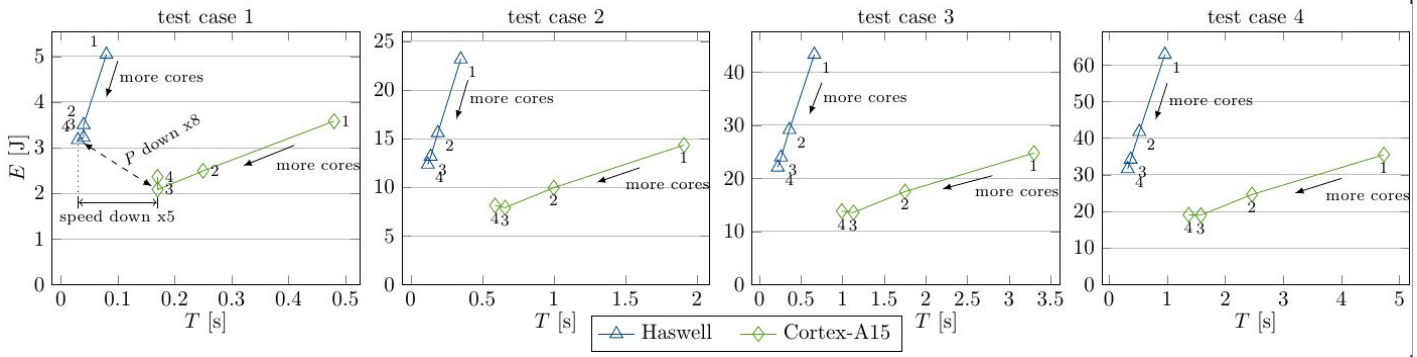
Benchmarks:
basic kernels, single node

Memory bandwidth- bound



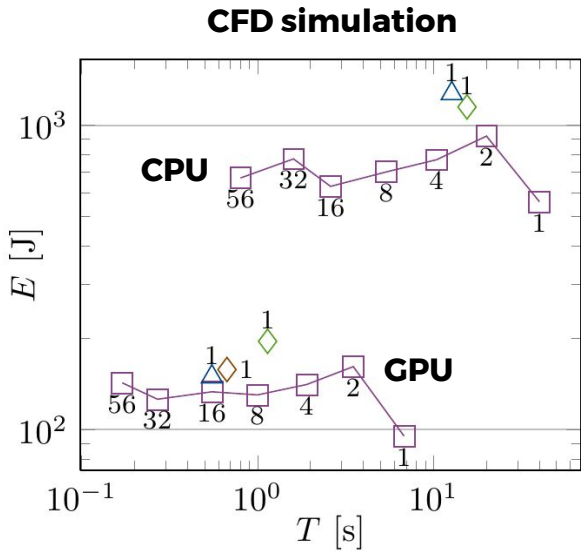
Applications on ARM Cortex A15

Global ocean circulation simulation on ARM Cortex A15



University of Stuttgart Germany
SimTech Cluster of Excellence
technische universität dortmund
fakultät für mathematik
FAU Friedrich-Alexander Universität Erlangen-Nürnberg Naturwissenschaftliche
Department Mathematik

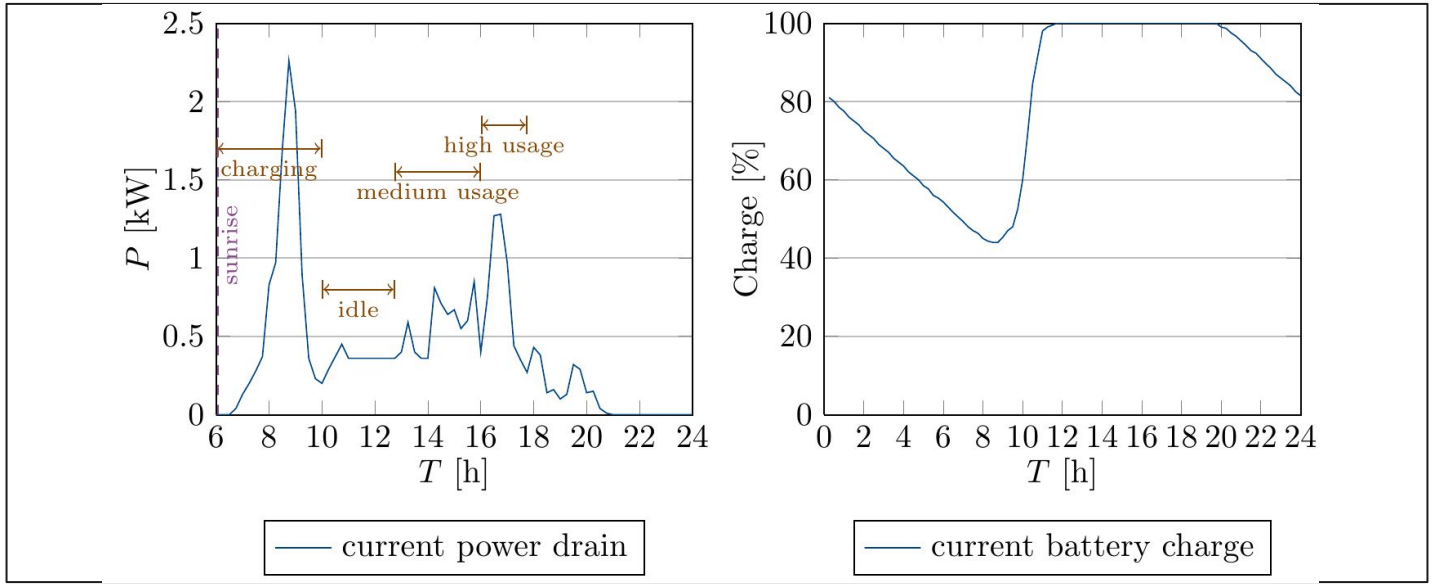
Benchmarks:
applications, full cluster, GPU, CPU

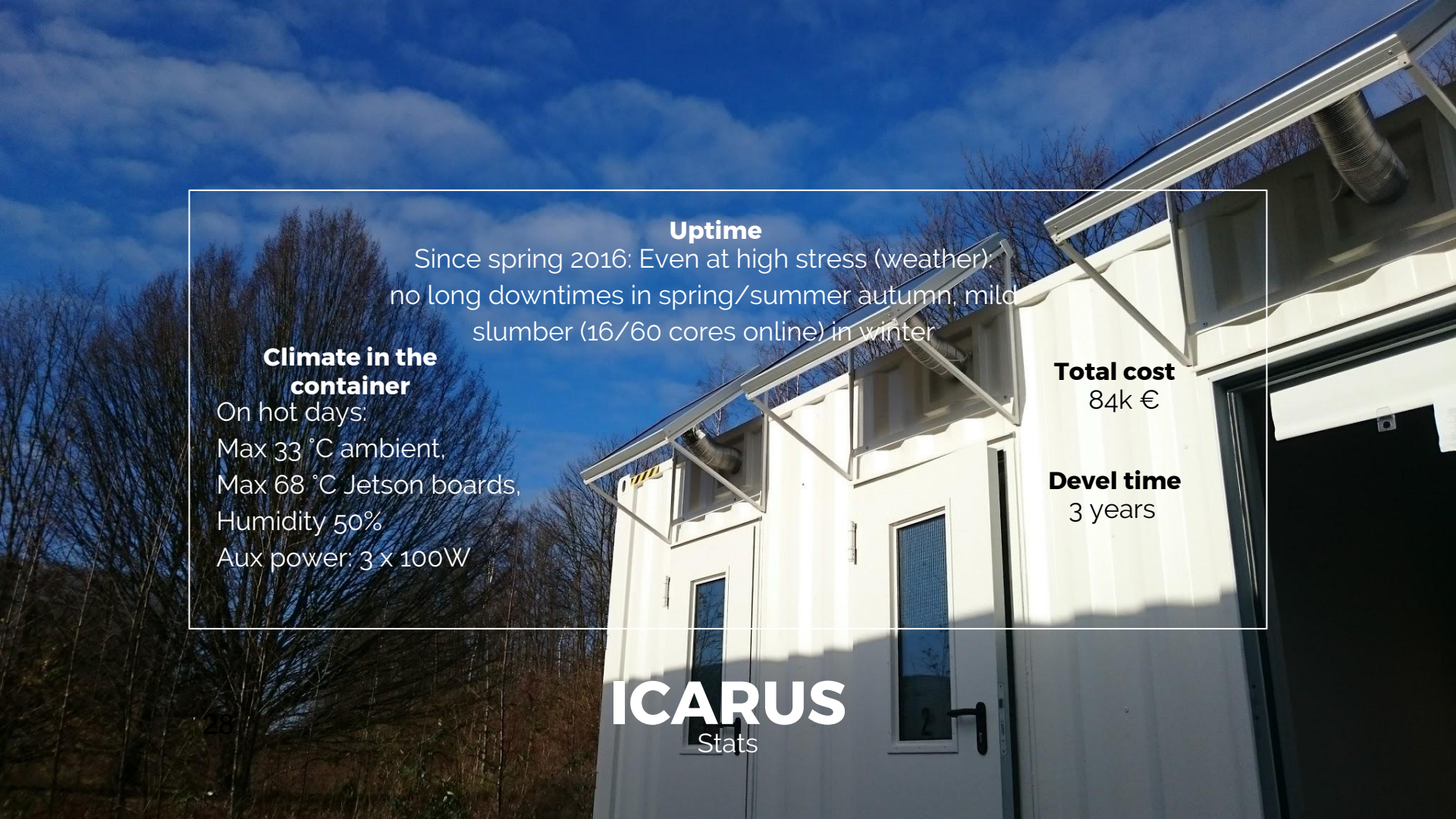


- **scaling** induces power penalties
- **parallelisation** induces performance penalties
- **all compared architectures beaten nonetheless**

commodity(2015) — commodity(2012 desktop) — Jetson TK1(2014) — commodity(2012 compute)

Photovoltaic subsystem results





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

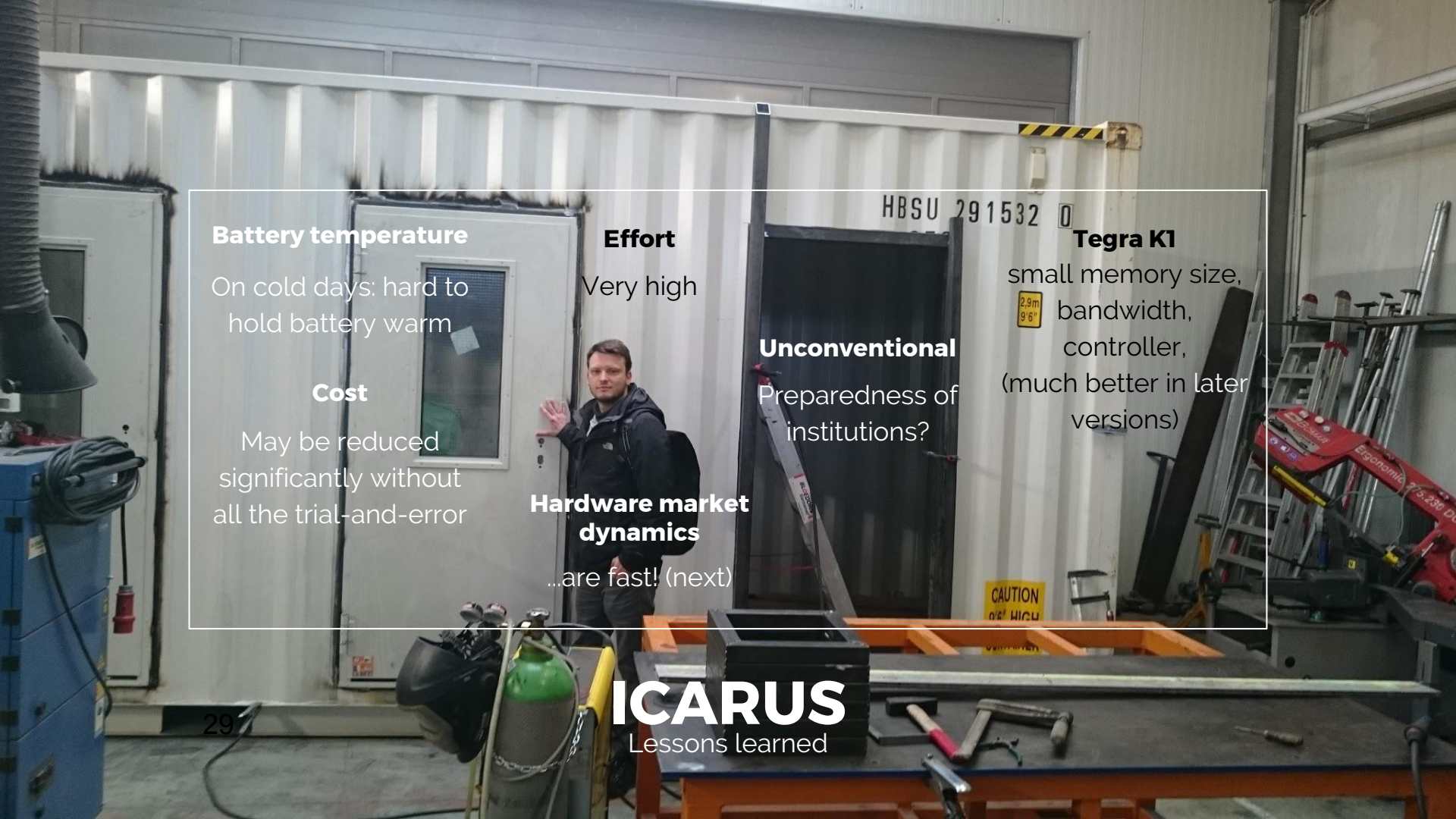
On hot days:
Max 33 ˚C ambient,
Max 68 ˚C Jetson boards,
Humidity 50%
Aux power: 3 x 100W

Total cost
84k €

Devel time
3 years

ICARUS

Stats



Battery temperature

On cold days: hard to hold battery warm

Cost

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

Effort

Very high

Hardware market dynamics

...are fast! (next)

Unconventional

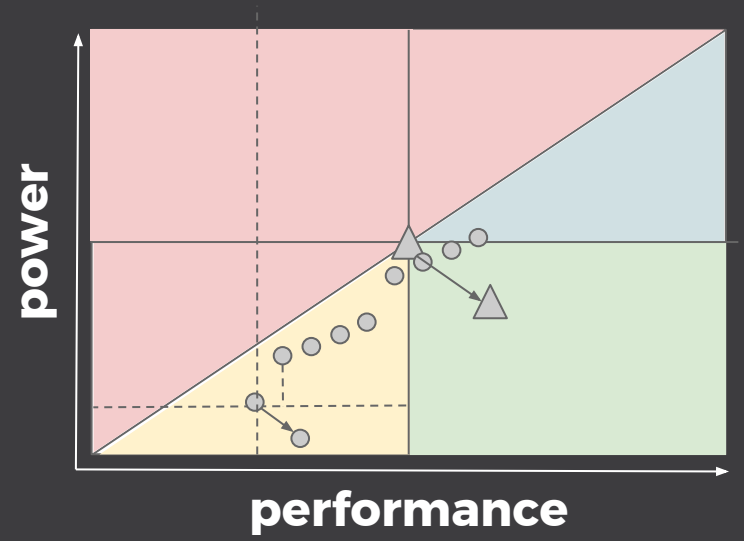
Preparedness of institutions?

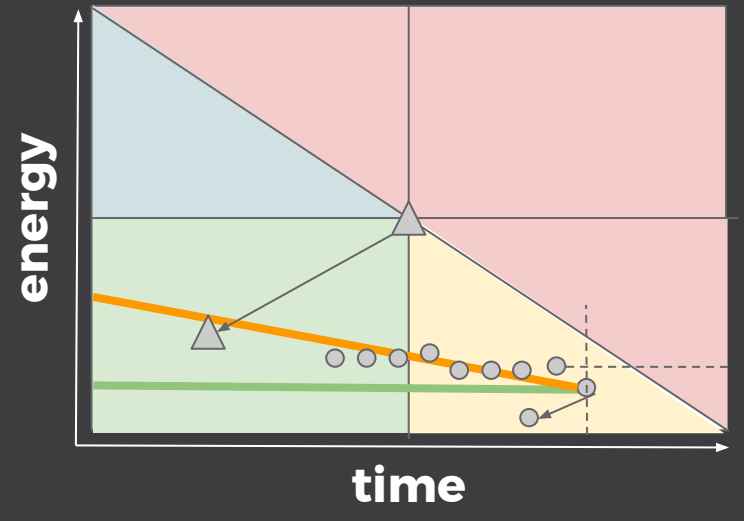
Tegra K1

small memory size, bandwidth, controller, (much better in later versions)

ICARUS

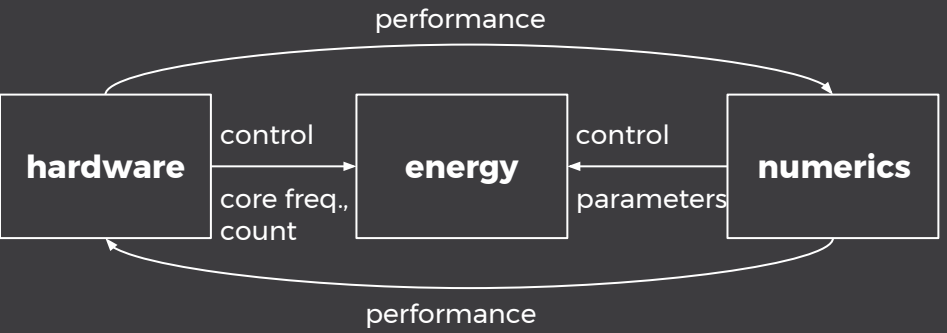
Lessons learned





Next?

- ‘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 of?



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'.

SPPEXA

ICARUS hardware is financed by MIWF NRW under the lead of MERCUR.

MERCUR Mercator Research Center Ruhr

Ministerium für Innovation, Wissenschaft und Forschung des Landes Nordrhein-Westfalen

(Super-)Computers and power dissipation

(when we thought ICARUS in 2014)

Green500 rank	architecture	Top500 rank	Performance / P [GFlop/s / W]	P [MW]
1	**Xeon + FirePro**	**168**	**5.2**	**0.057**
2	**Xeon + PEZY**	**369**	**4.9**	**0.037**
3	**Xeon + K20x**	**392**	**4.4**	**0.035**
4	**Xeon + K40m**	**361**	**4.0**	**0.044**

- **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)

Green500 rank	architecture	Top500 rank	Performance / P [GFlop/s / W]	P [MW]
1	**Xeon + P100**	**28**	**9.5**	**0.350**
2	**Xeon + P100**	**8**	**7.5**	**1.310**
3	**Xeon + PEZY**	**116**	**6.7**	**0.150**
4	**Sunway**	**1**	**6.0**	**15.370**

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

www.top500.org/green500/list/**2016**/11

(Super-)Computers and power dissipation

(2017)

Green500 rank	architecture	Top500 rank	Performance / P [GFlop/s / W]	P [MW]
1	**Xeon + P100**	**61**	**14**	**0.14**
2	**Xeon + P100**	**465**	**14**	**0.03**
3	**Xeon + P100**	**148**	**12**	**0.08**
4	**Xeon + P100**	**305**	**10**	**0.06**

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

www.top500.org/green500/list/**2017**/07