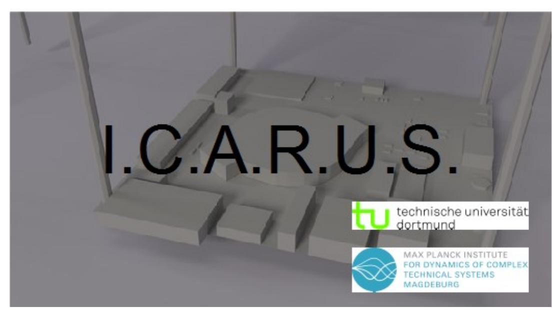
The ICARUS white paper: A scalable, energy-efficient, solar-powered HPC center based on low power GPUs

Markus Geveler, Dirk Ribbrock, Daniel Donner, Hannes Ruelmann, Christoph Höppke, David Schneider, Daniel Tomaschewski, Stefan Turek



Unconventional HPC, EuroPar 2016, Grenoble, 2016 / 8 / 23

markus.geveler@math.tu-dortmund.de

Outline

Introduction

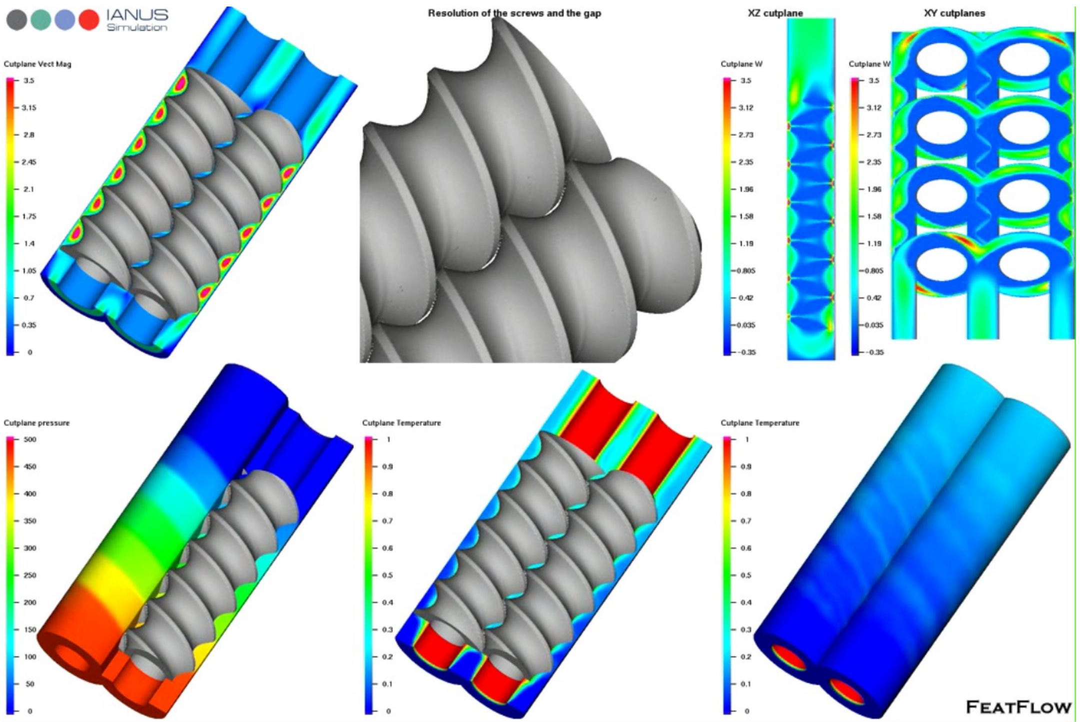
→ Simulation, Hardware-oriented numerics and energy-efficiency
→ We are energy consumers and why we can not continue like this
→ (Digital-)Computer hardware (in a general sense) of the future

ICARUS

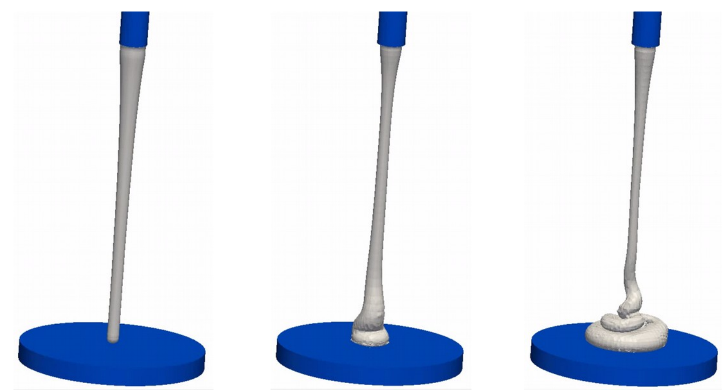
→ Architecture and white sheet
→ system integration of high-end photovoltaic-, battery- and embedded tech
→ benchmarking of basic kernels and applications
  → node level
  → cluster level

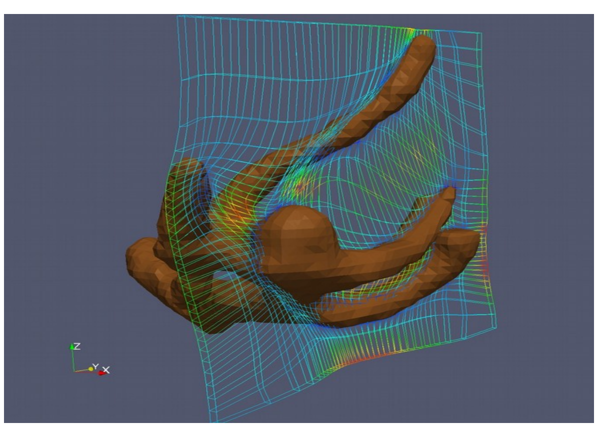
Preface: what we usually do

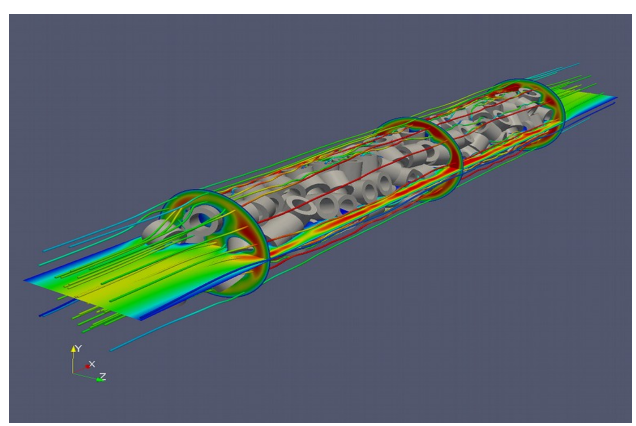
Simulation of technical flows with FEATFLOW

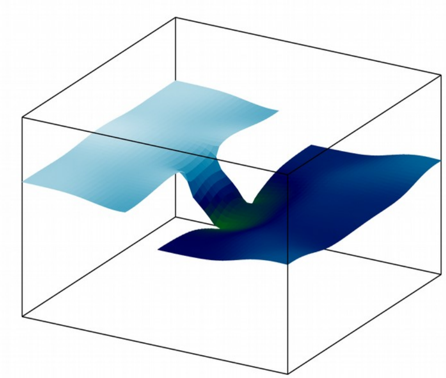


Performance engineering for
→ hardware efficiency
→ numerical efficiency
→ energy efficiency







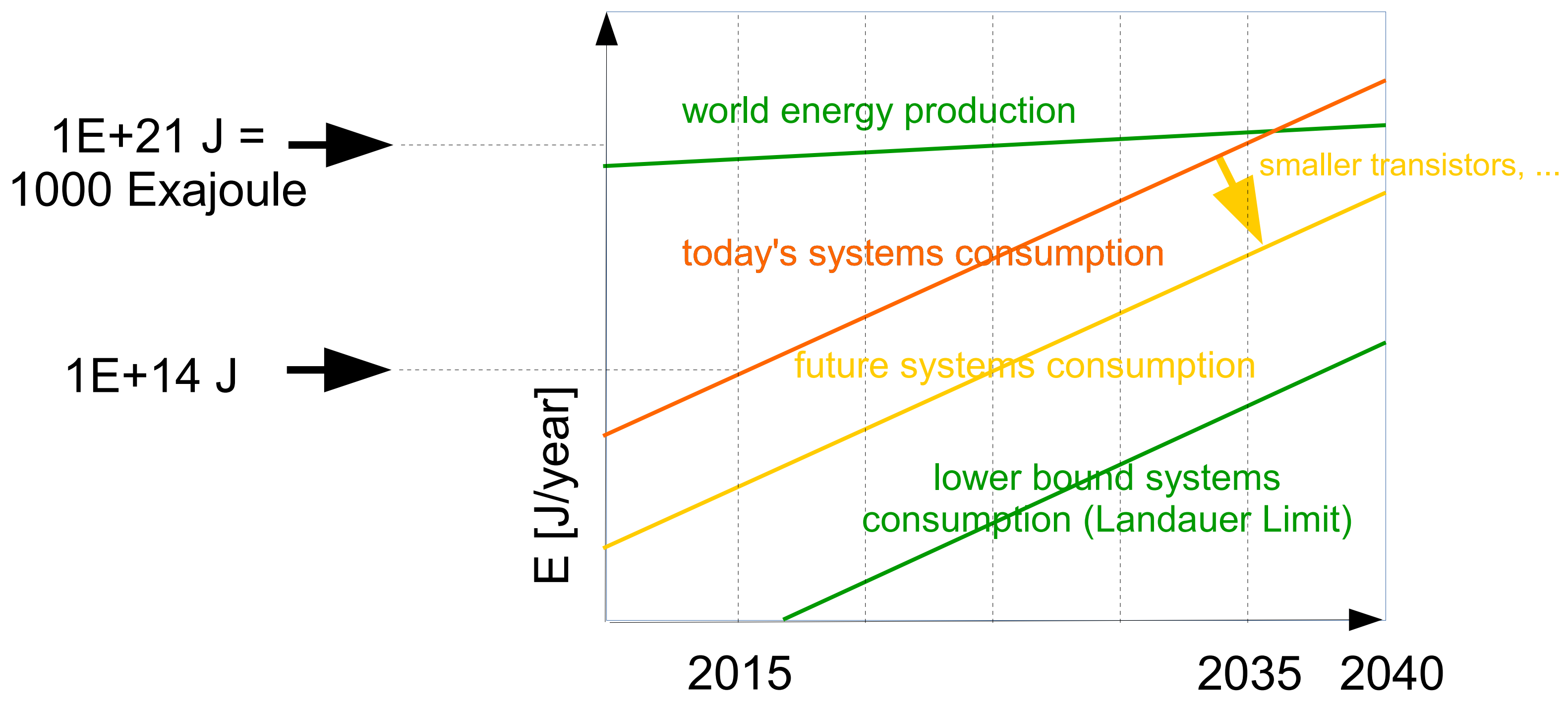


Motivation

Computers will require more energy than the world generates by 2040

"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"

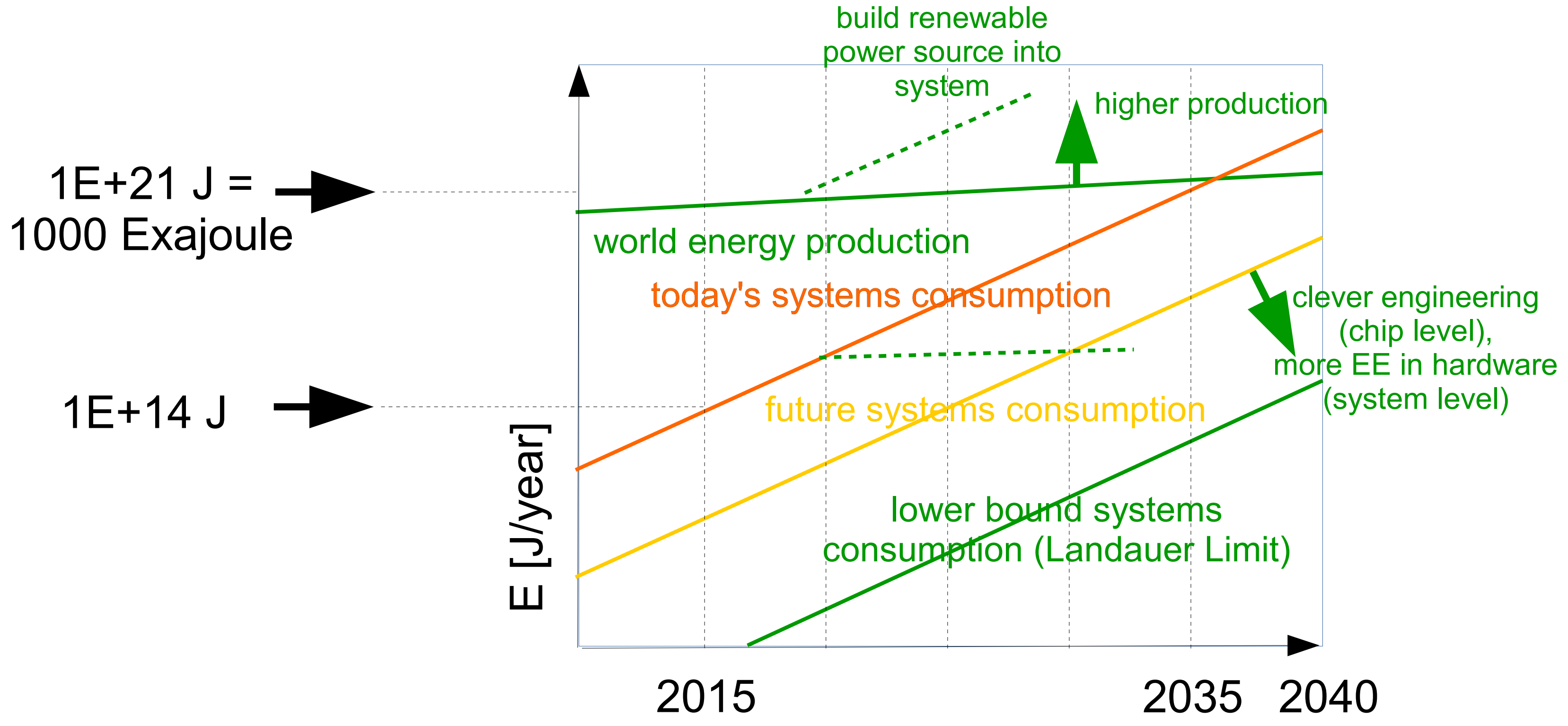
- the Semiconductor Industry Association (SIA), 2015



Motivation

What we can do

→ enhance supplies
→ enhance energy efficiency



HPC Hardware

Today's HPC facilities (?)

www.green500.org Green500 list Nov 2015

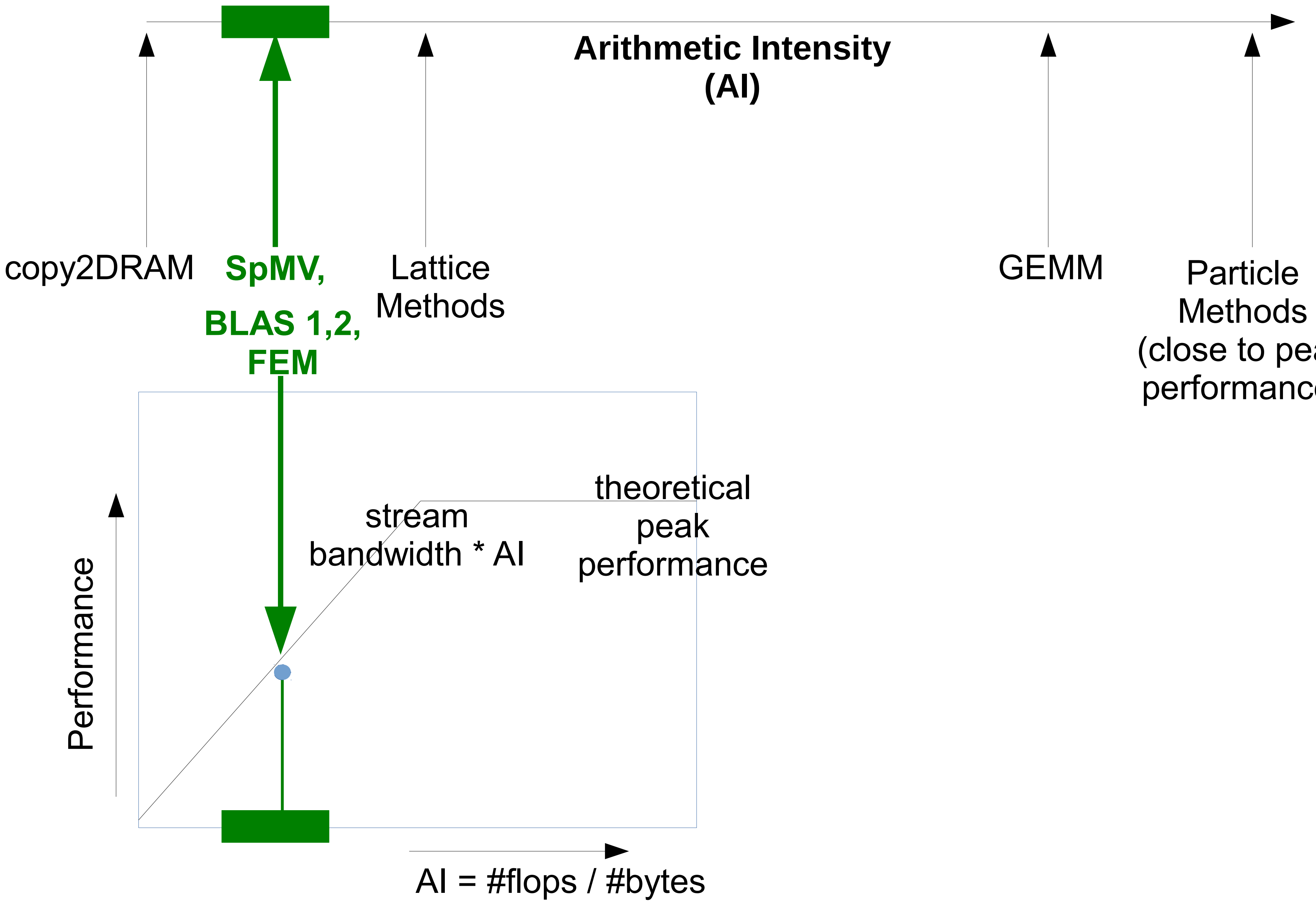
Green 500 rank	Top 500 rank	Total power [kW]	MFlops per watt	Year	Hardware architecture
1	133	50	7031	2015	ExaScaler-1.4 80Brick, Xeon E5-2618Lv3 8C 2.3GHz, Infiniband FDR, **PEZY-SC**
2	392	51	5331	2013	LX 1U-4GPU/104Re-1G Cluster, Intel Xeon E5-2620v2 6C 2.1GHz, Infiniband FDR, **NVIDIA Tesla K80**
3	314	57	5271	2014	ASUS ESC4000 FDR/G2S, Intel Xeon E5-2690v2 10C 3GHz, Infiniband FDR, **AMD FirePro S9150**
4	318	65	4778	2015	Sugon Cluster W780I, Xeon E5-2640v3 8C 2.6GHz, Infiniband QDR, **NVIDIA Tesla K80**
5	102	190	4112	2015	Cray CS-Storm, Intel Xeon E5-2680v2 10C 2.8GHz, Infiniband FDR, **Nvidia K80**
6	457	58	3857	2015	Inspur TS10000 HPC Server, Xeon E5-2620v3 6C 2.4GHz, 10G Ethernet, **NVIDIA Tesla K40**
7	225	110	3775	2015	Inspur TS10000 HPC Server, Intel Xeon E5-2620v2 6C 2.1GHz, 10G Ethernet, **NVIDIA Tesla K40**

No Top500 top-scorer

Top500 #1: 17,000 kW, 1,900 MFlop/s/W

unconventional hardware

What we can expect from hardware



Total efficiency of simulation software

Aspects

→ **Numerical efficiency** dominates asymptotic behaviour and wall clock time

→ **Hardware-efficiency**
→ exploit all levels of parallelism provided by hardware (SIMD, multithreading on a chip/device/socket, multiprocessing in a cluster, hybrids)
→ then try to reach good scalability (communication optimisations, block comm/comp)

→ **Energy-efficiency**
→ by hardware:
→ what is the most energy-efficient computer hardware? What is the best core frequency? What is the optimal number of cores used?
→ by software as a direct result of performance
→ but: its not all about performance

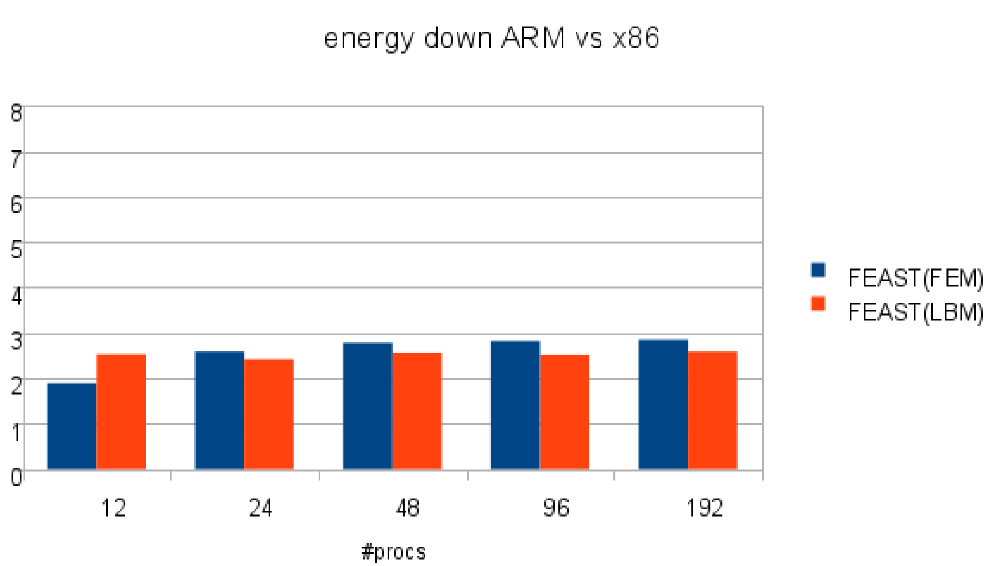
Hardware-oriented Numerics: Enhance hardware- and numerical efficiency simultaneously, use (most) energy-efficient Hardware(-settings) where available! Attention: codependencies!

Hardware-oriented Numerics

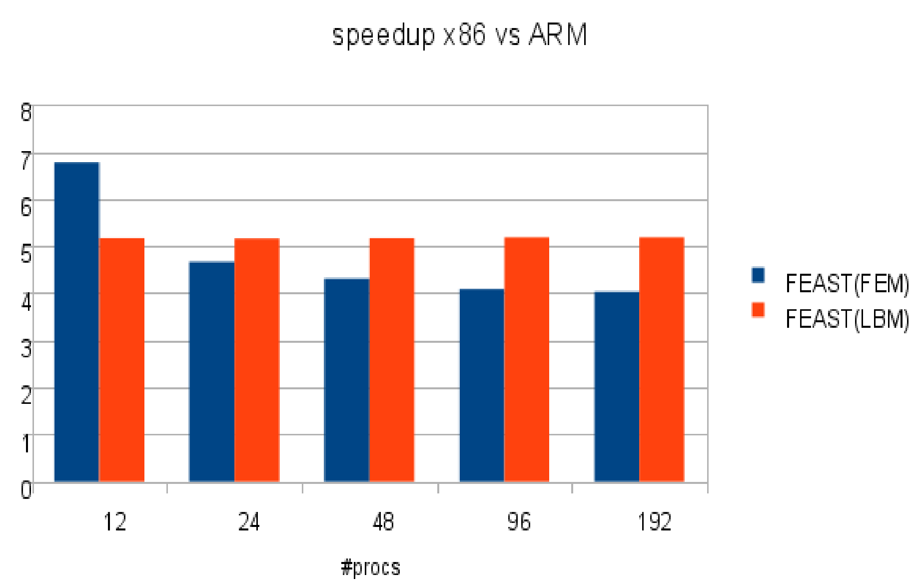
Energy Efficiency

- energy consumption/efficiency is one of the major challenges for future supercomputers → 'exascale'-challenge

- in 2012 we proved: we can solve PDEs for less energy 'than normal'
- simply by switching computational hardware from commodity to embedded
- Tegra 2 (2x ARM Cortex A9) in the Tibidabo system of the MontBlanc project
- tradeoff between energy and wall clock time



~3x less energy



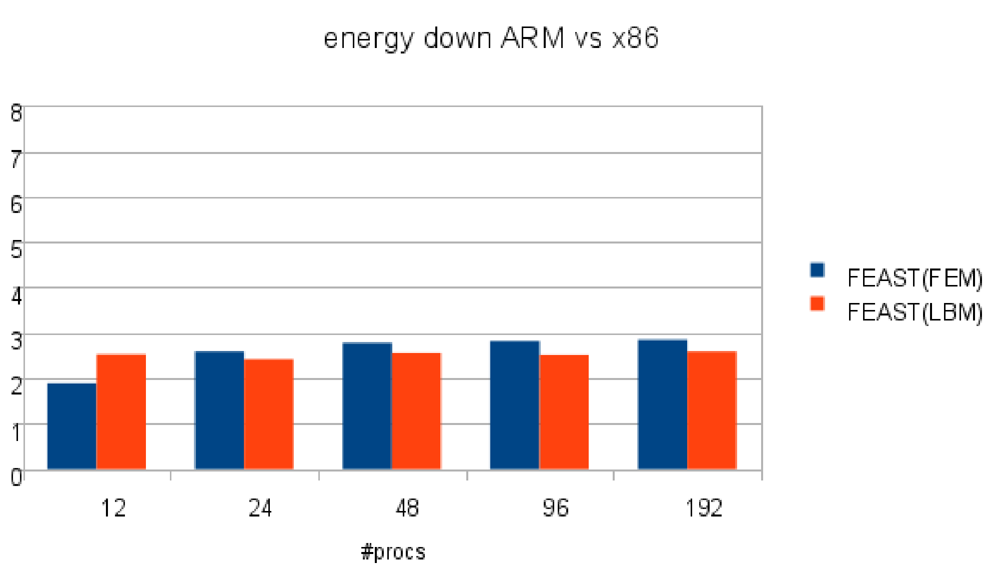
~5x more time!

Hardware-oriented Numerics

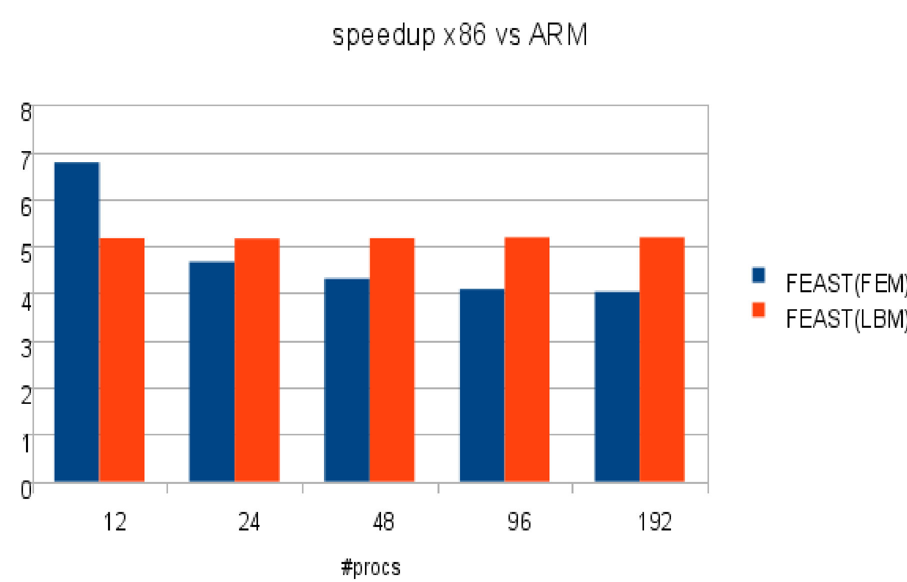
Energy Efficiency

To be more energy-efficient with different computational hardware, this hardware would have to dissipate *less power* at the *same performance* as the other!

→ More performance per Watt! → powerdown > speeddown



~3x less energy



~5x more time!

Hardware-oriented Numerics

Energy Efficiency: technology of ARM-based SoCs since 2012

Something has been happening in the mobile computing hardware evolution:

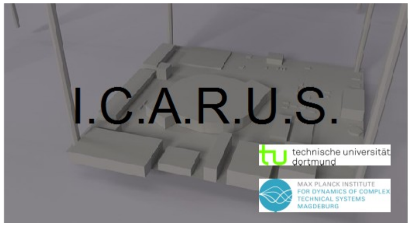
→ Tegra 3 (late 2012) was also based on A9 but had 4 cores
→ Tegra 4 (2013) is build upon the A15 core (higher frequency) and had more RAM and LPDDR3 instead of LPDDR2
→ Tegra K1 (32 Bit, late 2014) CPU pretty much like Tegra 4 but higher freq., more memory

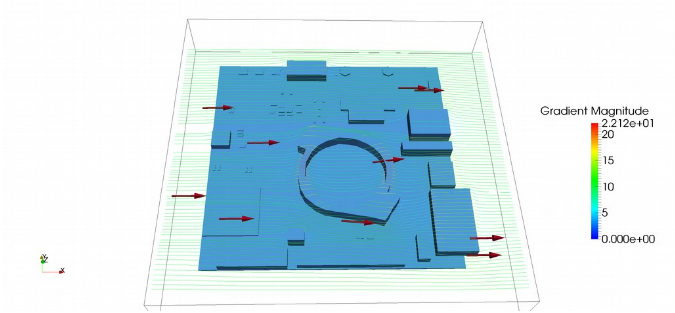
→ TK1 went GPGPU and comprises a programmable Kepler GPU on the same SoC!
→ the promise: 350+ Gflop/s for less than 11W
→ for comparison: Tesla K40 + x86 CPU: 4200 Gflop/s for 385W

→ 2.5x higher EE promised

→ interesting for Scientific Computing! Higher EE than commodity accelerator (of that time)!

An off-grid compute center of the future



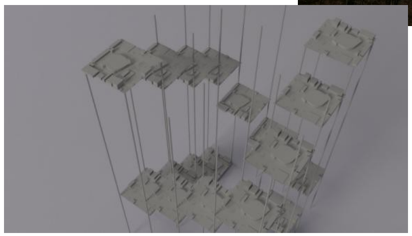


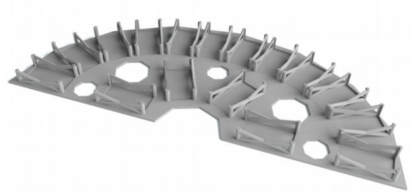
Vision

- **I**nsular
- **C**ompute center for
- **A**pplied Mathematics with
- **R**enewables-provided power supply based on
- **U**nconventional compute hardware empaired with
- **S**imulation Software for technical processes

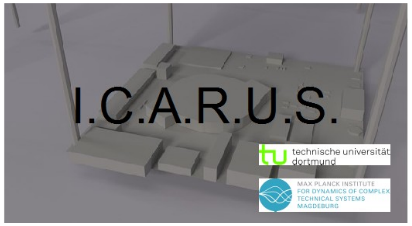
Motivation

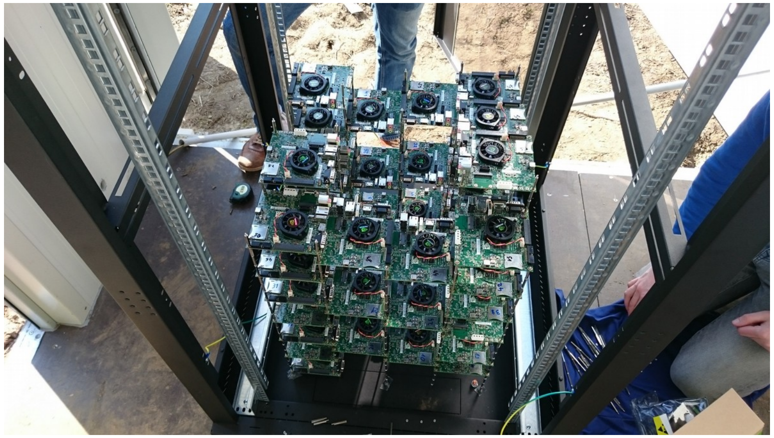
- **system integration** for Scientific HPC
  → high-end unconventional compute hardware
  → high-end renewable power source (photovoltaic)
  → specially tailored numerics, simulation software
- **no future spendings due to energy consumtion**
- **SME-class resource: <80K€**
- **Scalability, modular design**
- (simplicity)
- (maintainability)
- (safety)
- ...





Cluster





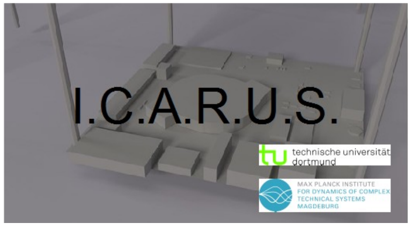
Whitesheet

→ **nodes:** 60 x NVIDIA Jetson TK 1
→ #cores (ARM Cortex-A15): 240
→ #GPUs (Kepler, 192 cores): 60
→ RAM/core: 2GB LPDDR3
→ switches (GiBit Ethernet): 3xL1, 1xL2
→ **cluster theoretical peak perf: ~20TFlop/s SP**
→ **cluster peak power: < 1kW, provided by PV**

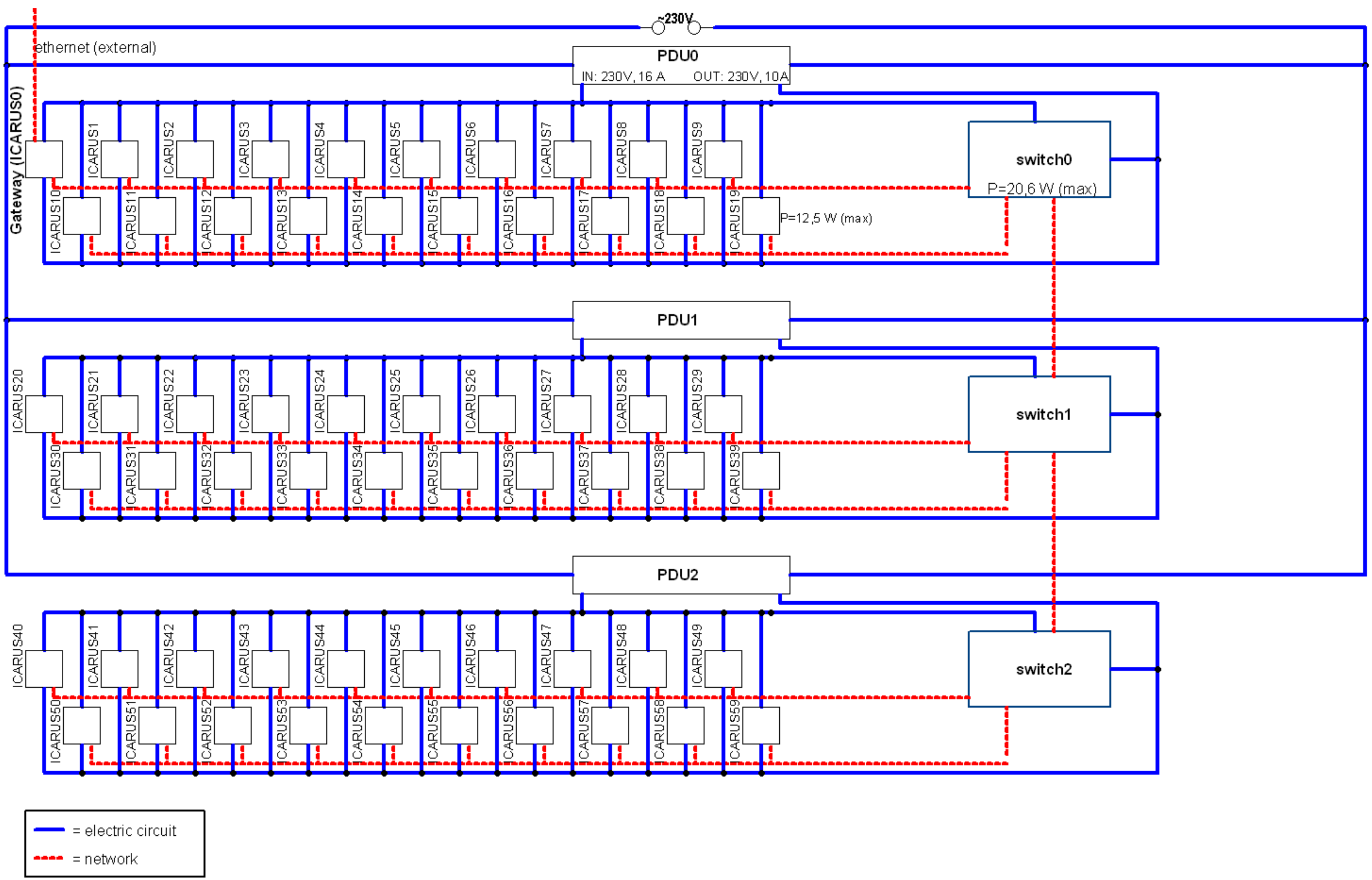
→ **PV capacity: 8kWp**
→ **battery: 8kWh**

→ **Software: FEAT (optimised for Tegra K1): www.featflow.de**

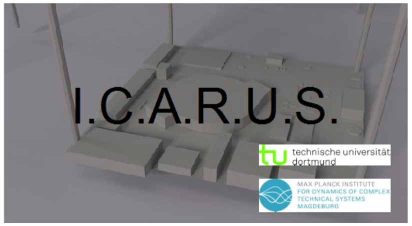
Cluster



Architecture



Housing and power supply



Photovoltaic units and battery rack

→ primary: approx. 16m x 3m area, 8kWp
→ secondary: for ventilation and cooling
→ battery: Li-Ion, 8kWh
→ 2 solar converters, 1 battery converter

Modified overseas cargo container

→ Steel Dry Cargo Container (High Cube) with dimensions 20 x 8 x 10 feet
→ climate isolation (90mm)
→ only connection to infrastructure: network cable

ICARUS compute hardware

We can build better computers

	i5-3470	i5-4690K	Jetson TK1
micro-architecture	Ivy Bridge	Haswell	Cortex-A15 (Tegra K1)
N_{cores}	4	4	4
clock speed	3.20 GHz (turbo 3.60 GHz)	3.50 GHz (turbo 3.9 GHz)	2.3 GHz
L1-cache	4x 32 KB + 4x 32 KB	4x 32 KB + 4x 32 KB	32 KB + 32 KB
L2- / L3-cache	4x 256 KB / 6 MB	4x 256 KB / 6 MB	2 MB / –
memory type	DDR3	DDR3	LPDDR3
peak memory bandwidth	25.6 GByte/s	25.6 GByte/s	14.9 GByte/s
P_{base}	51 W (Intel chipset)	41 W (Intel chipset)	3.9 W (Jetson TK1)
release date	Q2'12	Q2'14	Q2'14

	GTX 660 / Tesla K20x systems	GTX 980 system	Jetson TK1
micro-architecture	Kepler	Maxwell	Kepler
memory type	GDDR5	GDDR5	LPDDR3
peak memory bandwidth	144.2/250 GByte/s	336.5 GByte/s	14.9 GByte/s
peak performance (SP)	1881/3935 GFlop/s	6054 GFlop/s	326 GFlop/s
peak performance (DP)	78/1312 GFlop/s	189 GFlop/s	13 GFlop/s
P_{base}	41/45 W (Intel chipset)	51 W (Intel chipset)	3.9 W (Jetson TK1)
release date	Q3'12	Q2'15	Q2'14

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

Testhardware and measuring

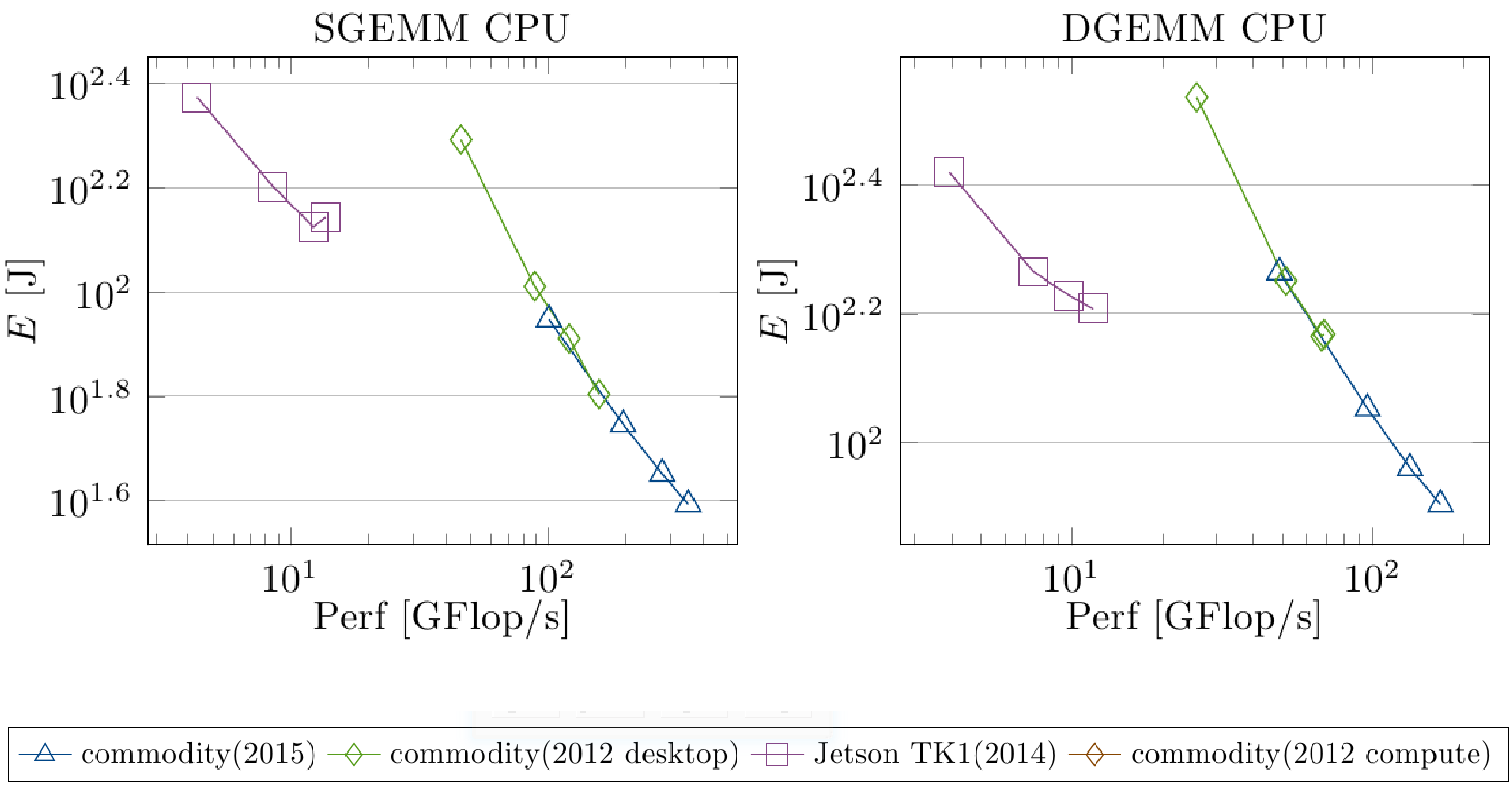
Complete 'box'

→ measure power at
AC-converter (inlet)
→ all power needed for the node

→ Note: not a chip-to-chip comparison, but a system-to-system one

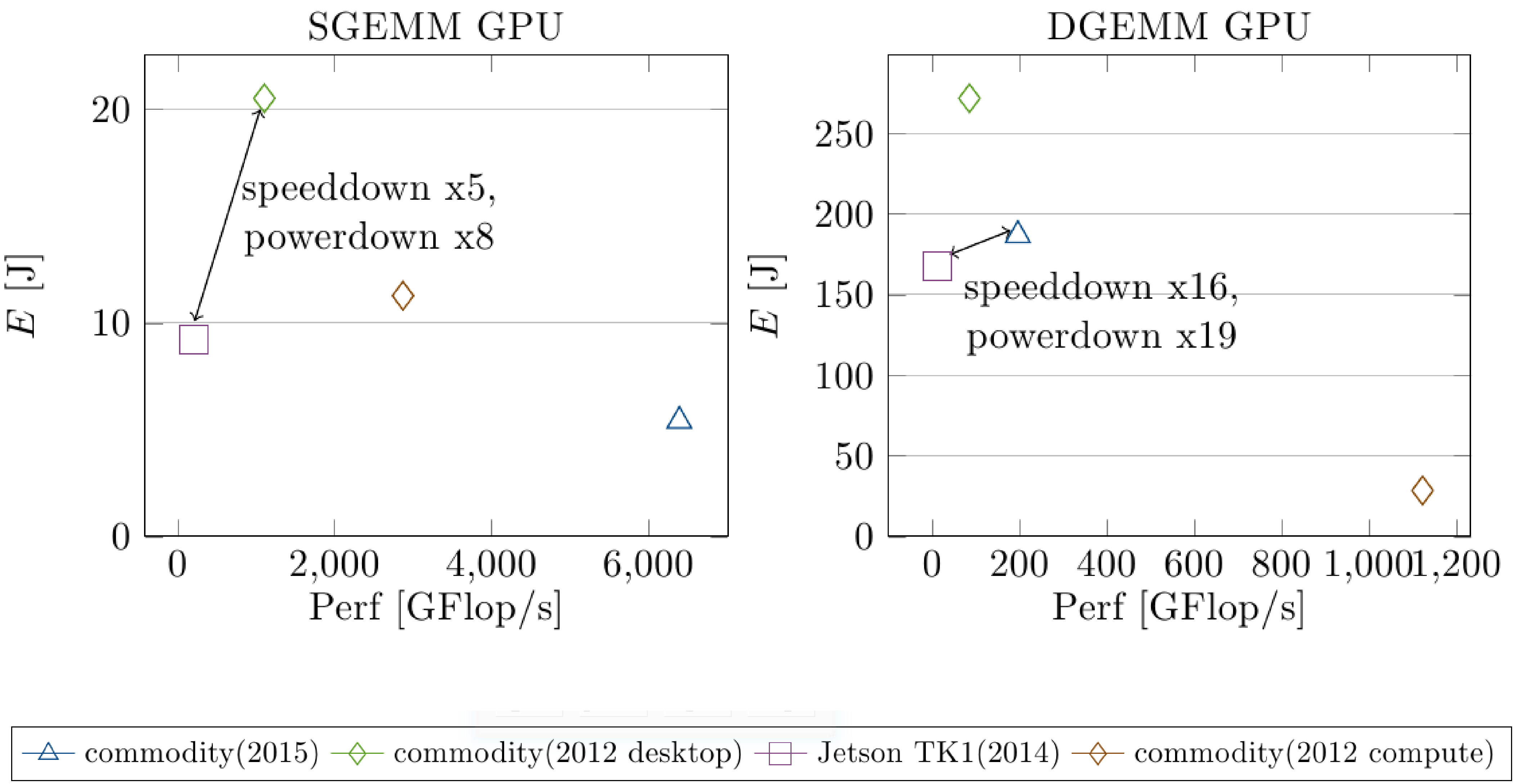
Sample results

Compute-bound, CPU



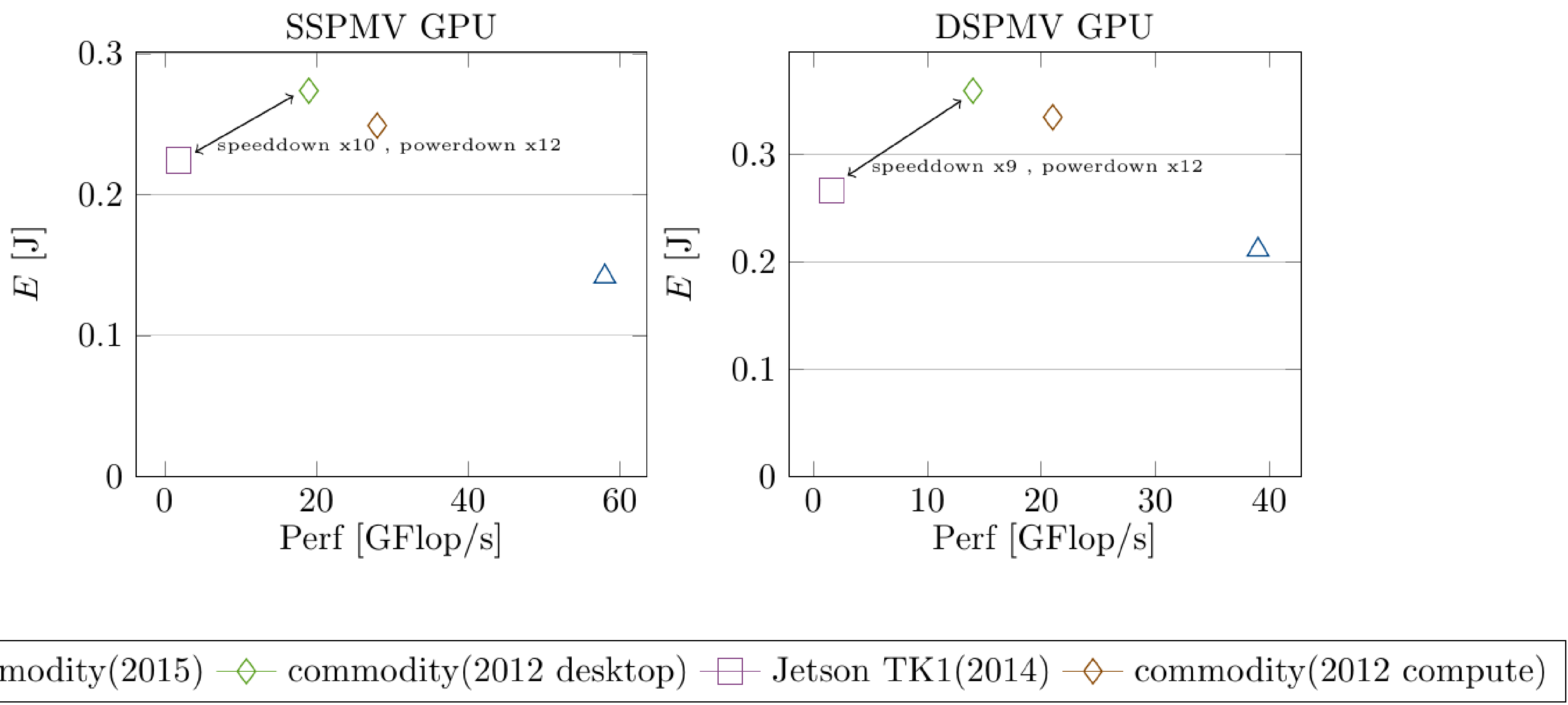
Sample results

Compute-bound, GPU



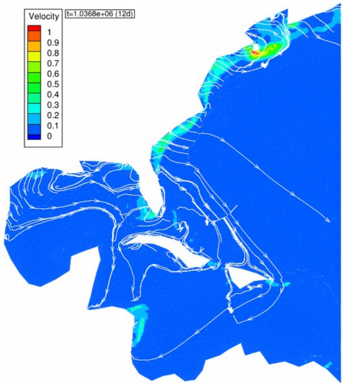
Sample results

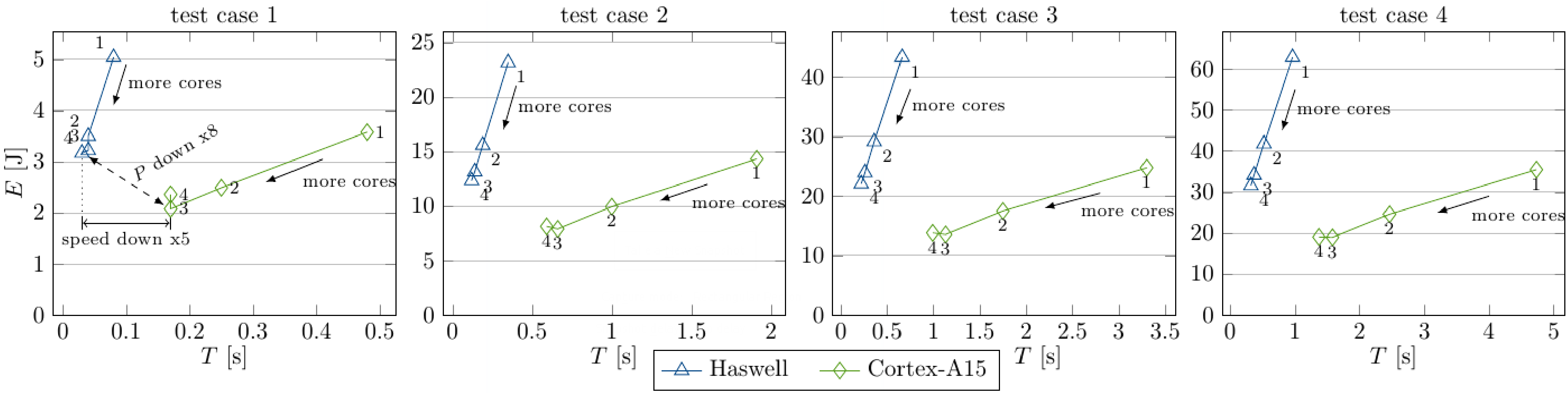
Memory bandwidth-bound, GPU



Sample results

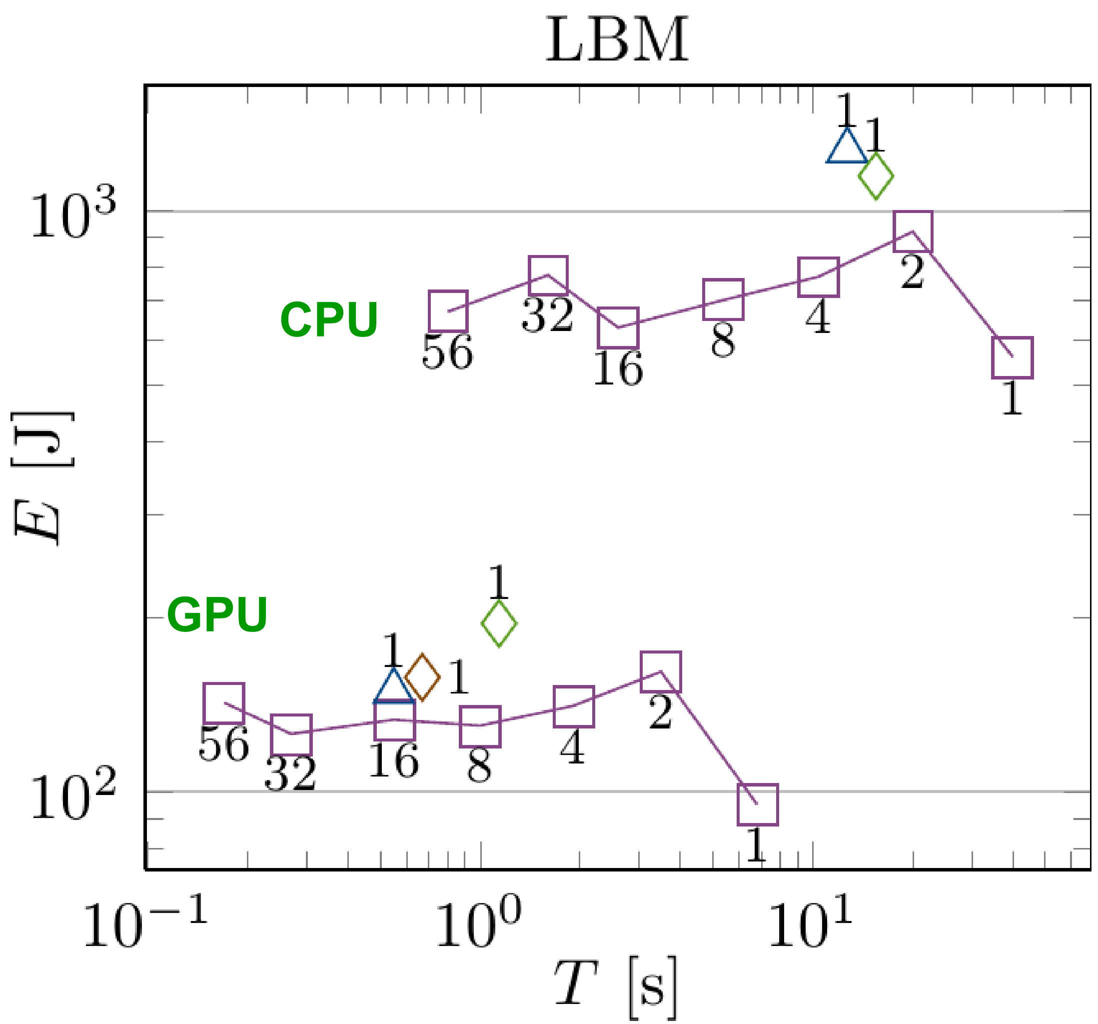
Applications, DG-FEM, single node





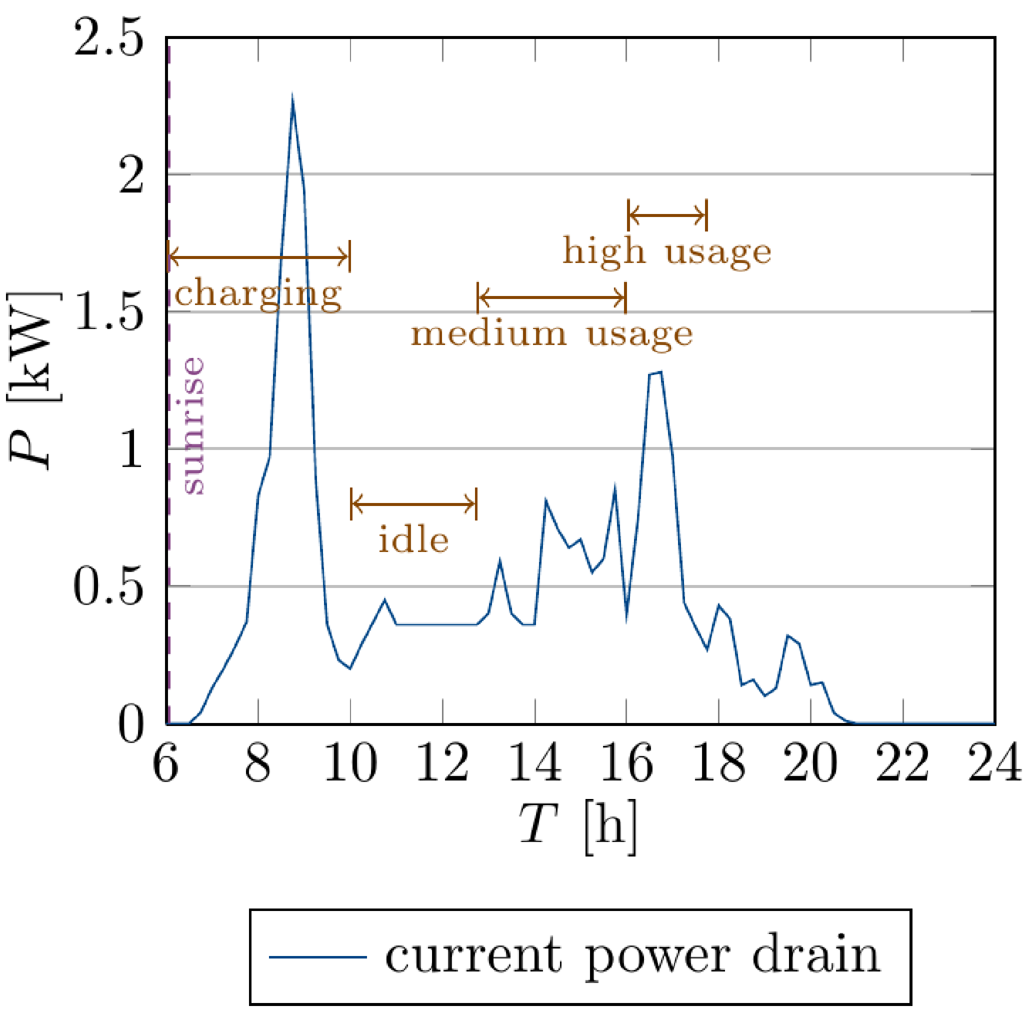
Sample results

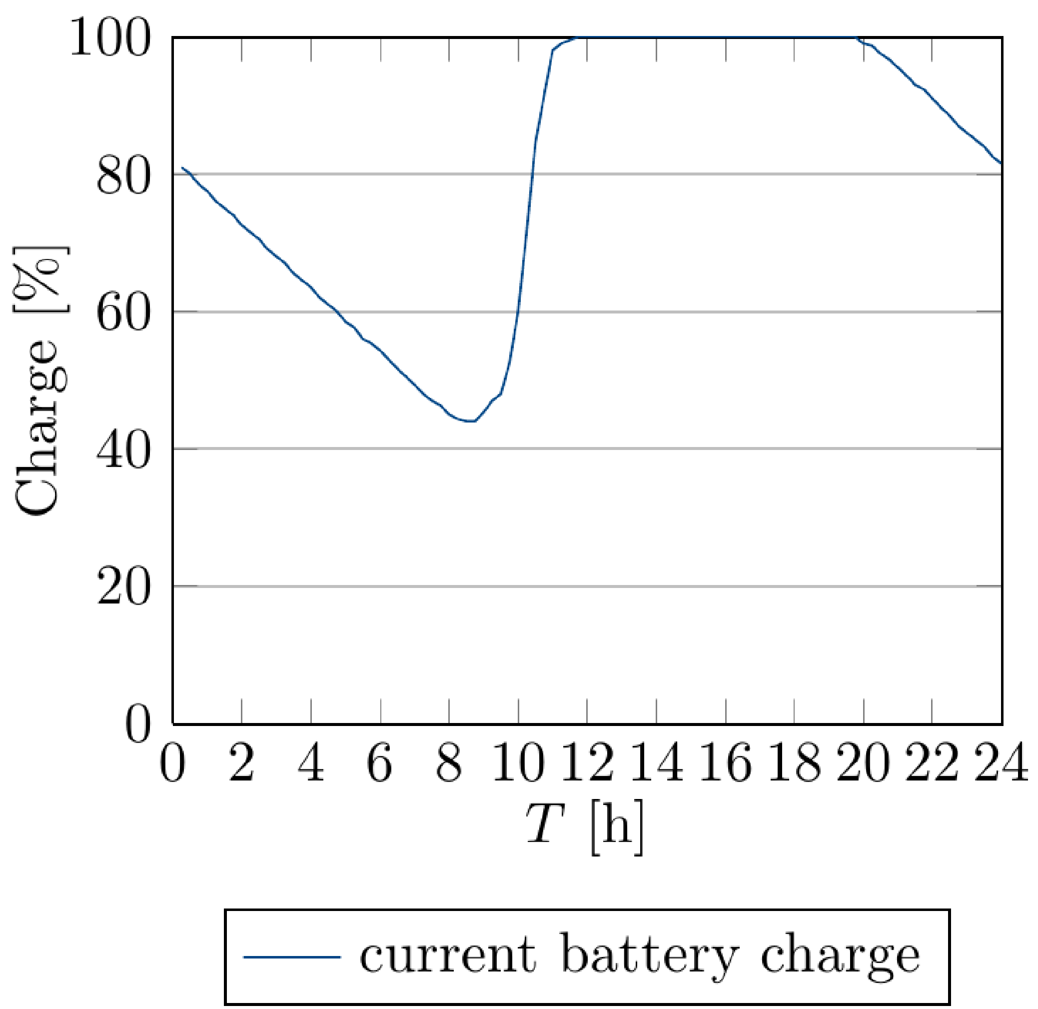
Applications, LBM, full cluster



Sample results

Power supply





More experiences so far

Reliability

→ cluster operation since March 2016
→ 61 Jetson boards
    → 57 working permanently
    → 4 with uncritical fan failure

→ uptime: since first startup (almost), exept for maintenance

Temperature

→ on warm days (31 degrees Celsius external, 50% humidity):
    → 33 degrees Celsius ambient temperature in container
    → 35% relative humidity
    → 39 – 43 degrees Celsius on chip in idle mode
    → approx. 68 degrees Celsius at load
→ monitored by rack PDU sensors

Conclusion and outlook

ICARUS

→ built-in power source
→ built-in ventilation, cooling, heating
→ no infrastructure-needs (except for area)
→ Jetson boards are 'cool': with the currently installed hardware, no additional cooling needed
→ versatile: compute (or other-) hardware can be exchanged easily

→ off-grid resource: can be deployed in areas with weak infrastructure:
→ in developing nations/regions
→ as secondary/emergency system

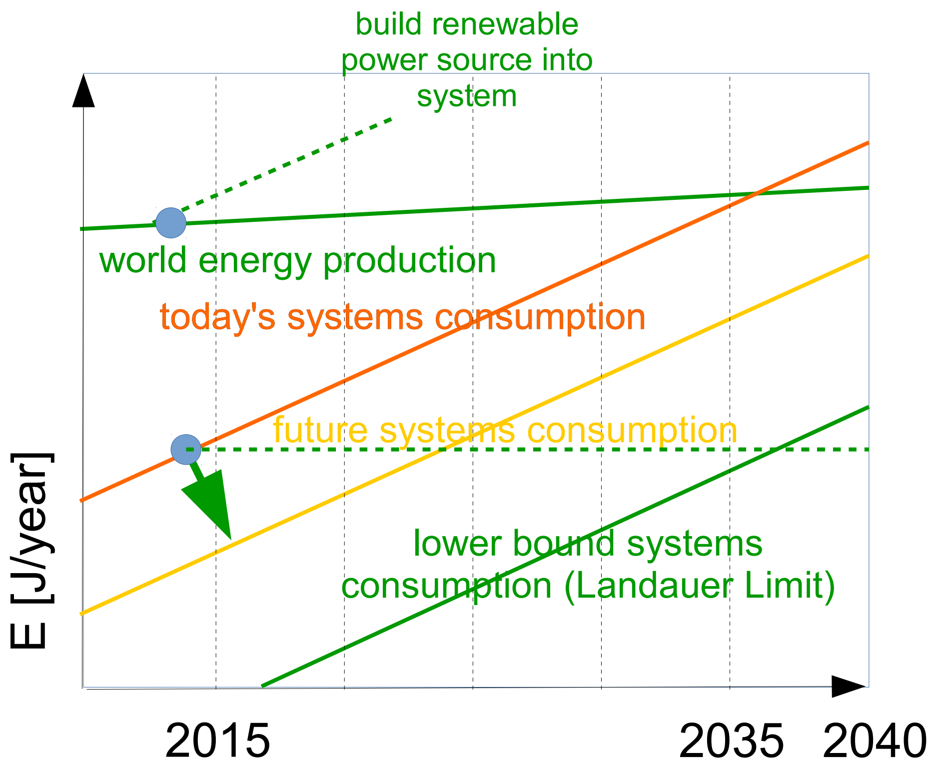
Conclusion and outlook

To be very clear

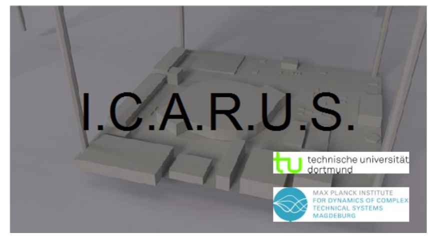
→ We do not see a future where Jetson TK 1 or similar boards are the compute nodes
→ We wanted to show, that we can build compute hardware differently and for a certain kind of application, it works
→ We showed, that system integration of renewables and compute tech is possible

The future

→ Tegra X1, …, or other
→ commodity GPUs (?)
→ collect data all year (weather!)
→ learn how to improve all components



Thank you



www.icarus-green-hpc.org

This work has been supported in part by the German Research Foundation (DFG) through the Priority Program 1648 ‘Software for Exascale Computing’ (grant TU 102/48).

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