technische universität dortmund

MAX PLANCK INSTITUTE FOR DYNAMICS OF COMPLEX TECHNICAL SYSTEMS MAGDEBURG

Future Data Centers for Energy-Efficient Large Scale Numerical Simulations

On the need for a combination of Hardware-oriented Numerics with Unconventional HPC

Markus Geveler*, Martin Köhler**, Dirk Ribbrock*, Jens Saak**,
Gerry Truschkewitz**, Peter Benner**, Stefan Turek*

7th KoMSO Challenge Workshop, Heidelberg
2015 / 10 / 9

markus.geveler@math.tu-dortmund.de
*Institute f. Applied Mathematics, TU Dortmund
**Computational Methods in Systems and Control Theory,
MPI for Dynamics of Complex Technical Systems Magdeburg

Outline

Roughly three parts:

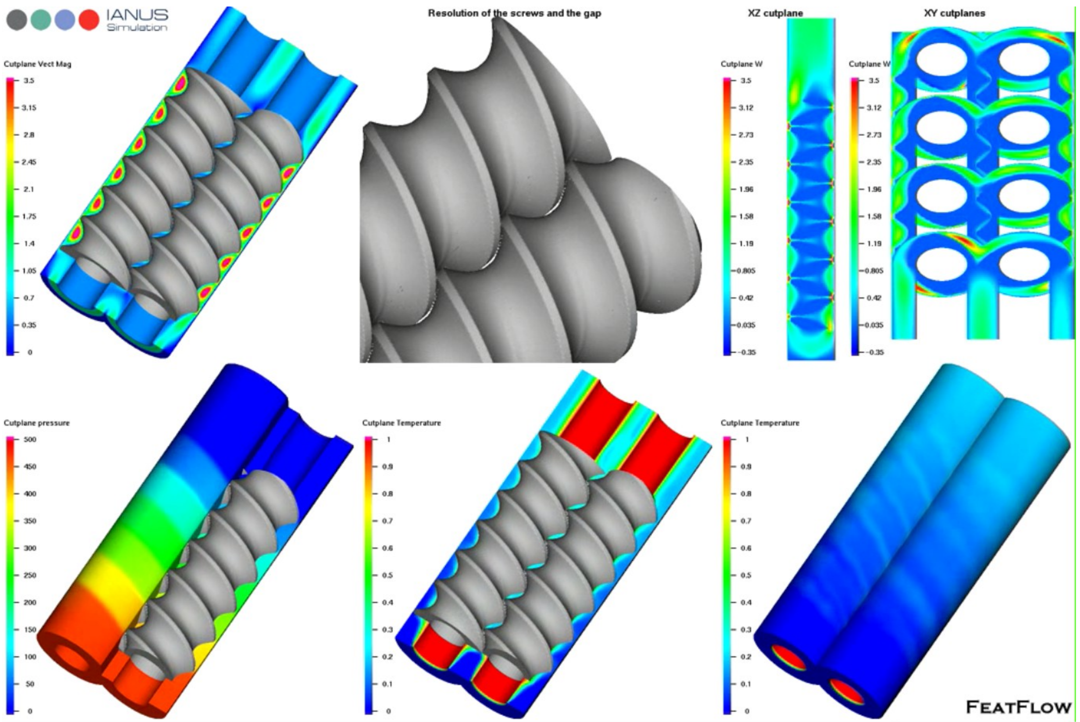
→ **Hardware-oriented Numerics** in *Energiewende* (or: why do Mathematicians build a supercomputer?)
→ Our perspective on Energiewende
→ **Green computing, Hardware-oriented Numerics and Unconventional HPC**
→ Simulation w.r.t. hardware-, numerical-, and energy-efficiency

→ A prototype for future Data Centers
→ Preliminary work with ARM-based clusters
→ the I.C.A.R.U.S experimental cluster based on NVIDIA Tegra K1 and a minimum energy data storage system

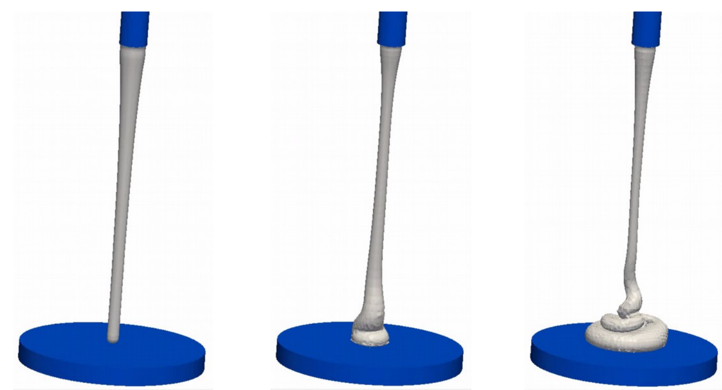
→ Performance engineering for unconventional hardware with focus on energy efficiency in the **FEAT software family**

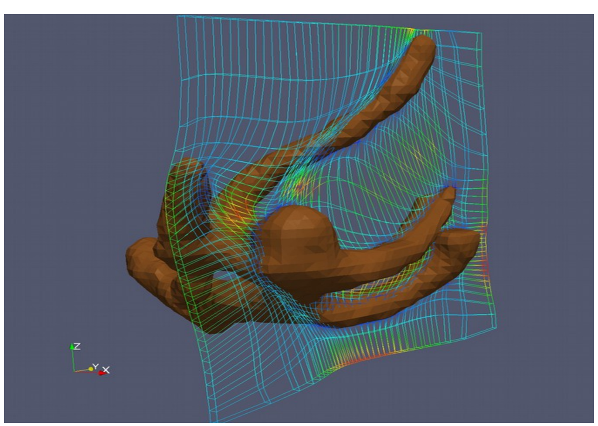
Motivation: where it all leads...

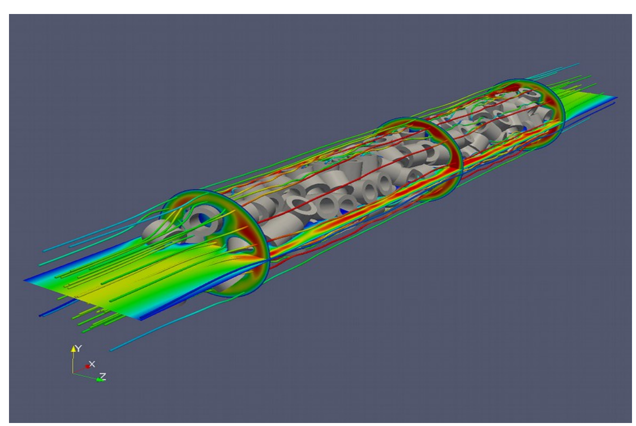
Simulation of technical flows

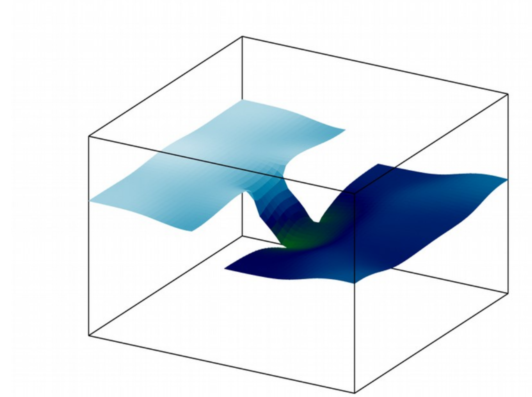


characteristics:
→ high end modelling, numerics
→ huge requirements: computation, storage









Our perspective on *Energiewende*

Applied Mathematics is also on the 'user'-side!

- Energy Production based on renewables and better grids are crucially needed
- But also *energy consumers* have to adapt → **Energy Efficiency (EE) increase is needed ('output up, consumption down')**
- MSO are rightfully considered offering powerful tools for conserving energy in industrial processes
- But: **How energy-efficient can simulation be performed?**

How can the mathematical community increase EE in what we do?

Green HPC and Hardware-oriented Numerics

Current supercomputers / data centers aren't green

RANK	SITE	SYSTEM	CORES	RMAX (TFLOP/S)	RPEAK (TFLOP/S)	POWER (KW)
1	National Super Computer Center in Guangzhou China	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P NUDT	3,120,000	33,862.7	54,902.4	17,808
2	DOE/SC/Oak Ridge National Laboratory United States	**Titan** - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x Cray Inc.	560,640	17,590.0	27,112.5	8,209
3	DOE/NNSA/LLNL United States	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom IBM	1,572,864	17,173.2	20,132.7	7,890
4	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu	705,024	10,510.0	11,280.4	12,660
5	DOE/SC/Argonne National Laboratory United States	**Mira** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM	786,432	8,586.6	10,066.3	3,945
6	Swiss National Supercomputing Centre (CSCS) Switzerland	**Piz Daint** - Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect , NVIDIA K20x Cray Inc.	115,984	6,271.0	7,788.9	2,325
7	King Abdullah University of Science and Technology Saudi Arabia	**Shaheen II** - Cray XC40, Xeon E5-2698v3 16C 2.3GHz, Aries interconnect Cray Inc.	196,608	5,537.0	7,235.2	2,834
8	Texas Advanced Computing Center/Univ. of Texas United States	**Stampede** - PowerEdge C8220, Xeon E5-2680 8C 2.700GHz, Infiniband FDR, Intel Xeon Phi SE10P Dell	462,462	5,168.1	8,520.1	4,510
9	Forschungszentrum Juelich (FZJ) Germany	**JUQUEEN** - BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect IBM	458,752	5,008.9	5,872.0	2,301
10	DOE/NNSA/LLNL United States	**Vulcan** - BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect IBM	393,216	4,293.3	5,033.2	1,972

Mflop/s / W

1902

2143

2177

830

2 – 17 MW of power!

Simulation comes at huge cost!

http://www.top500.org/lists/2015/06/

Green HPC and Hardware-oriented Numerics

Greenmost supercomputers are 'unconventional'

	Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)
Japan	1	7,031.58	RIKEN	Shoubu - ExaScaler-1.4 80Brick, Xeon E5-2618Lv3 8C 2.3GHz, Infiniband FDR, PEZY-SC	50.32
Japan	2	6,842.31	High Energy Accelerator Research Organization /KEK	Suiren Blue - ExaScaler-1.4 16Brick, Xeon E5-2618Lv3 8C 2.3GHz, Infiniband PEZY-SC	28.25
Japan	3	6,217.04	High Energy Accelerator Research Organization /KEK	Suiren - ExaScaler 32U256SC Cluster, Intel Xeon E5-2660v2 10C 2.2GHz, Infiniband FDR PEZY-SC	32.59
Germany	4	5,271.81	GSI Helmholtz Center	ASUS ESC4000 FDR/G2S, Intel Xeon E5-2690v2 10C 3GHz, Infiniband FDR, AMD FirePro S9150	57.15
Japan	5	4,257.88	GSIC Center, Tokyo Institute of Technology	TSUBAME-KFC - LX 1U-4GPU/104Re-1G Cluster, Intel Xeon E5-2620v2 6C 2.100GHz, Infiniband FDR, NVIDIA K20x	39.83
USA	6	4,112.11	Stanford Research Computing Center	XStream - Cray CS-Storm, Intel Xeon E5-2680v2 10C 2.8GHz, Infiniband FDR, Nvidia K80	190.00
USA	7	3,962.73	Cray Inc.	Storm1 - Cray CS-Storm, Intel Xeon E5-2660v2 10C 2.2GHz, Infiniband FDR, Nvidia K40m	44.54
USA	8	3,631.70	Cambridge University	Wilkes - Dell T620 Cluster, Intel Xeon E5-2630v2 6C 2.600GHz, Infiniband FDR, NVIDIA K20	52.62
Germany	9	3,614.71	TU Dresden, ZIH	Taurus GPUs - Bull bullx R400, Xeon E5-2680v3 12C 2.5GHz, Infiniband FDR, Nvidia K80	58.01
USA	10	3,543.32	Financial Institution	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband, NVIDIA K20x	54.60

http://www.green500.org/lists/green201506

Accelerators rule the field, unconventional design is leading, Germany could potentially do better

Green HPC and Hardware-oriented Numerics

Greenmost supercomputers are 'unconventional'

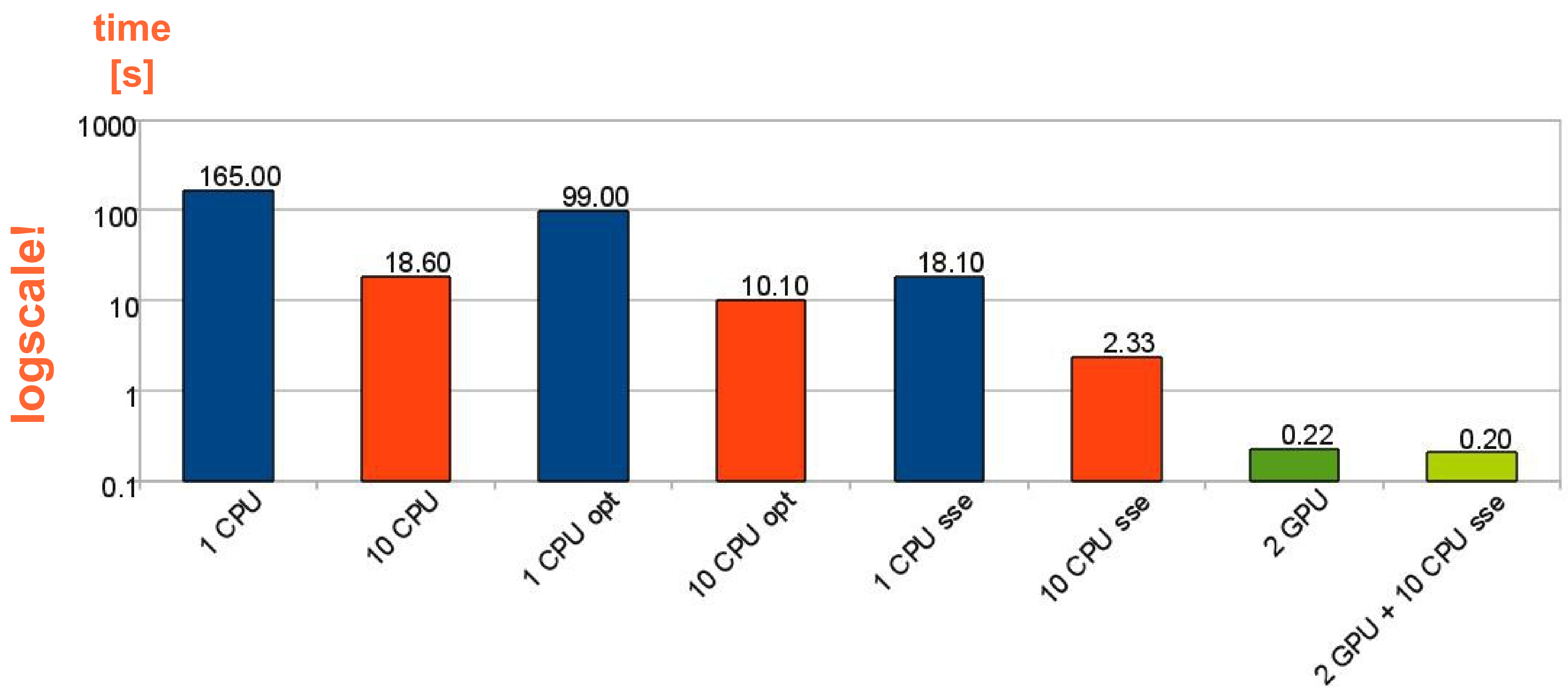
Top500 rank	Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)
160	1	7,031.58	RIKEN	Shoubu - ExaScaler-1.4 80Brick, Xeon E5-2618Lv3 8C 2.3GHz, Infiniband FDR, PEZY-SC	50.32
392	2	6,842.31	High Energy Accelerator Research Organization /KEK	Suiren Blue - ExaScaler-1.4 16Brick, Xeon E5-2618Lv3 8C 2.3GHz, Infiniband, PEZY-SC	28.25
366	3	6,217.04	High Energy Accelerator Research Organization /KEK	Suiren - ExaScaler 32U256SC Cluster, Intel Xeon E5-2660v2 10C 2.2GHz, Infiniband FDR, PEZY-SC	32.59
215	4	5,271.81	GSI Helmholtz Center	ASUS ESC4000 FDR/G2S, Intel Xeon E5-2690v2 10C 3GHz, Infiniband FDR, AMD FirePro S9150	57.15
	5	4,257.88	GSIC Center, Tokyo Institute of Technology	TSUBAME-KFC - LX 1U-4GPU/104Re-1G Cluster, Intel Xeon E5-2620v2 6C 2.100GHz, Infiniband FDR, NVIDIA K20x	39.83
	6	4,112.11	Stanford Research Computing Center	XStream - Cray CS-Storm, Intel Xeon E5-2680v2 10C 2.8GHz, Infiniband FDR, Nvidia K80	190.00
	7	3,962.73	Cray Inc.	Storm1 - Cray CS-Storm, Intel Xeon E5-2660v2 10C 2.2GHz, Infiniband FDR, Nvidia K40m	44.54
	8	3,631.70	Cambridge University	Wilkes - Dell T620 Cluster, Intel Xeon E5-2630v2 6C 2.600GHz, Infiniband FDR, NVIDIA K20	52.62
	9	3,614.71	TU Dresden, ZIH	Taurus GPUs - Bull bullx R400, Xeon E5-2680v3 12C 2.5GHz, Infiniband FDR, Nvidia K80	58.01
	10	3,543.32	Financial Institution	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband, NVIDIA K20x	54.60

http://www.green500.org/lists/green201506

Still not developed under the premise of EE, power source not included in thinking yet

Hardware-oriented Numerics

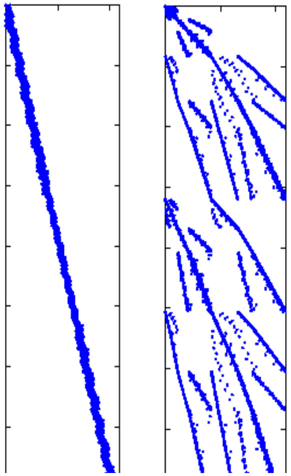
(I) : Hardware Efficiency: apply 'classical' roofline models until optimal



→ good, but: sole concentration on HE will not do the job

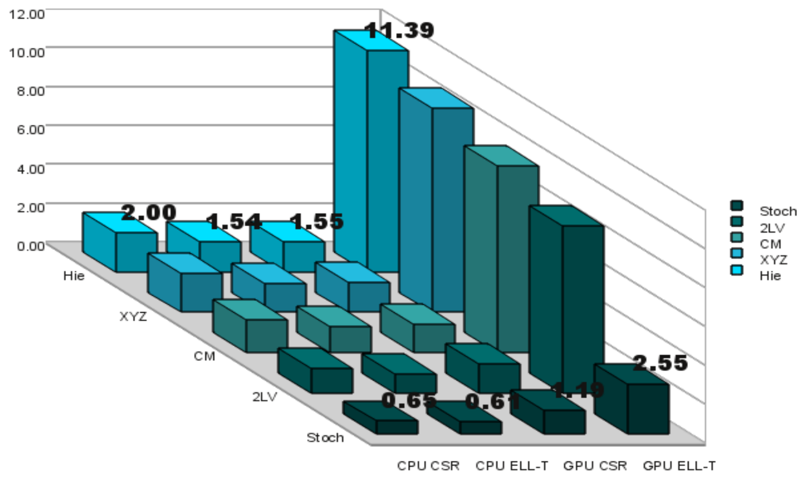
Hardware-oriented Numerics

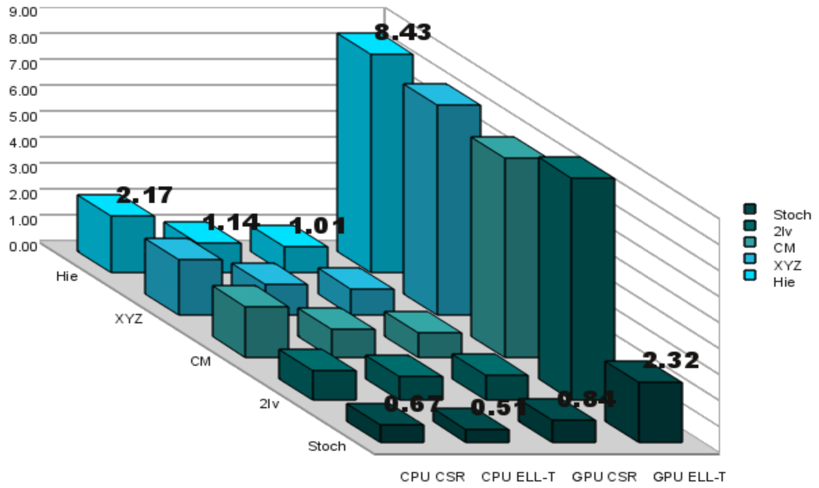
(I) Hardware Efficiency: kernel-based optimisation: SpMV



- one of the most prominent kernels in solving PDEs with high-end FEM
- memory access matters a lot
- hardware efficiency considerations start early: DOF numbering
- hardware-efficiency requires different matrix storage
- FE space matters

Performance [Gflop/s]





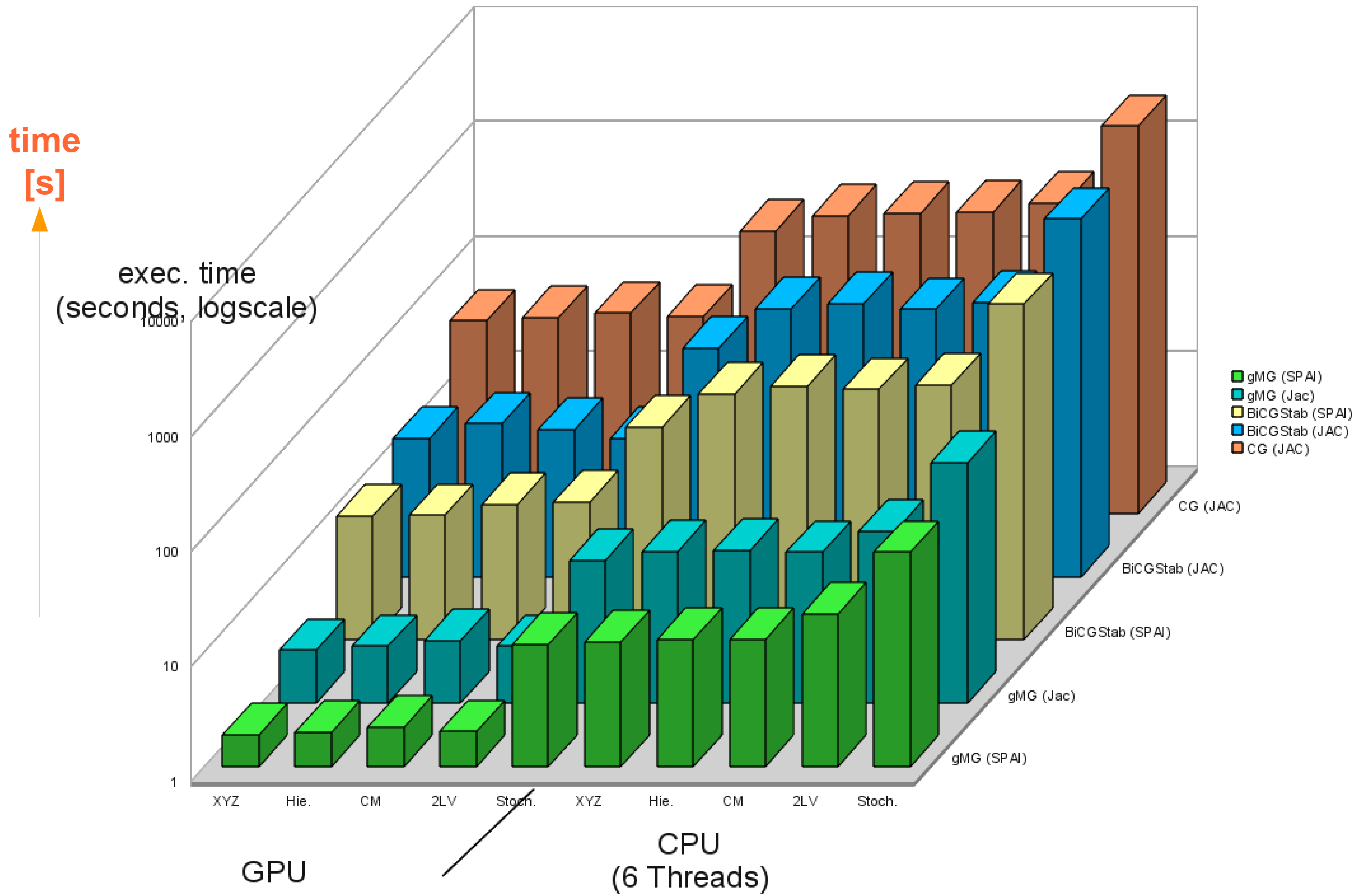
Q1

Q2

→ good, but: sole concentration on HE will not do the job

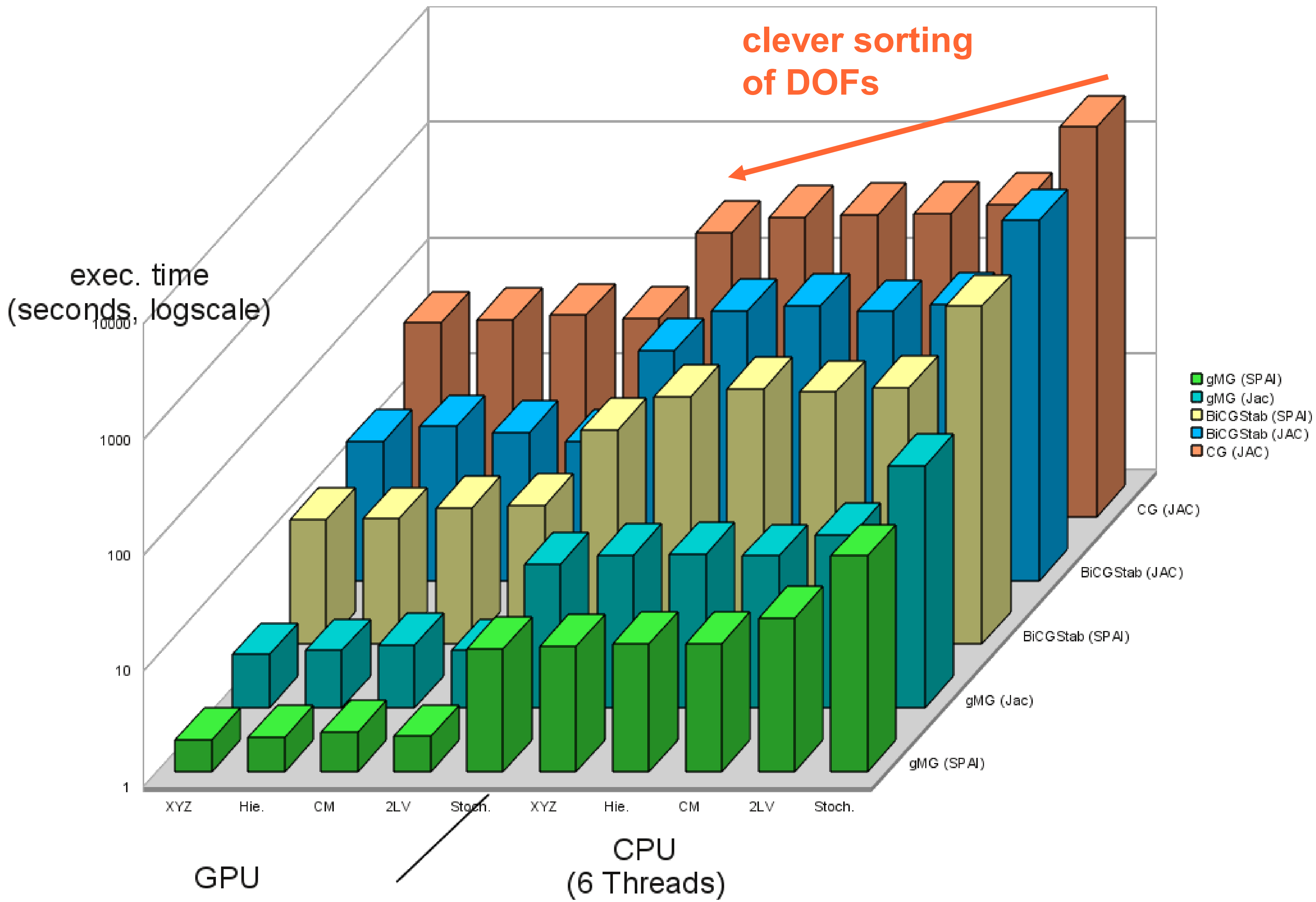
Hardware-oriented Numerics

(II) Numerical Efficiency



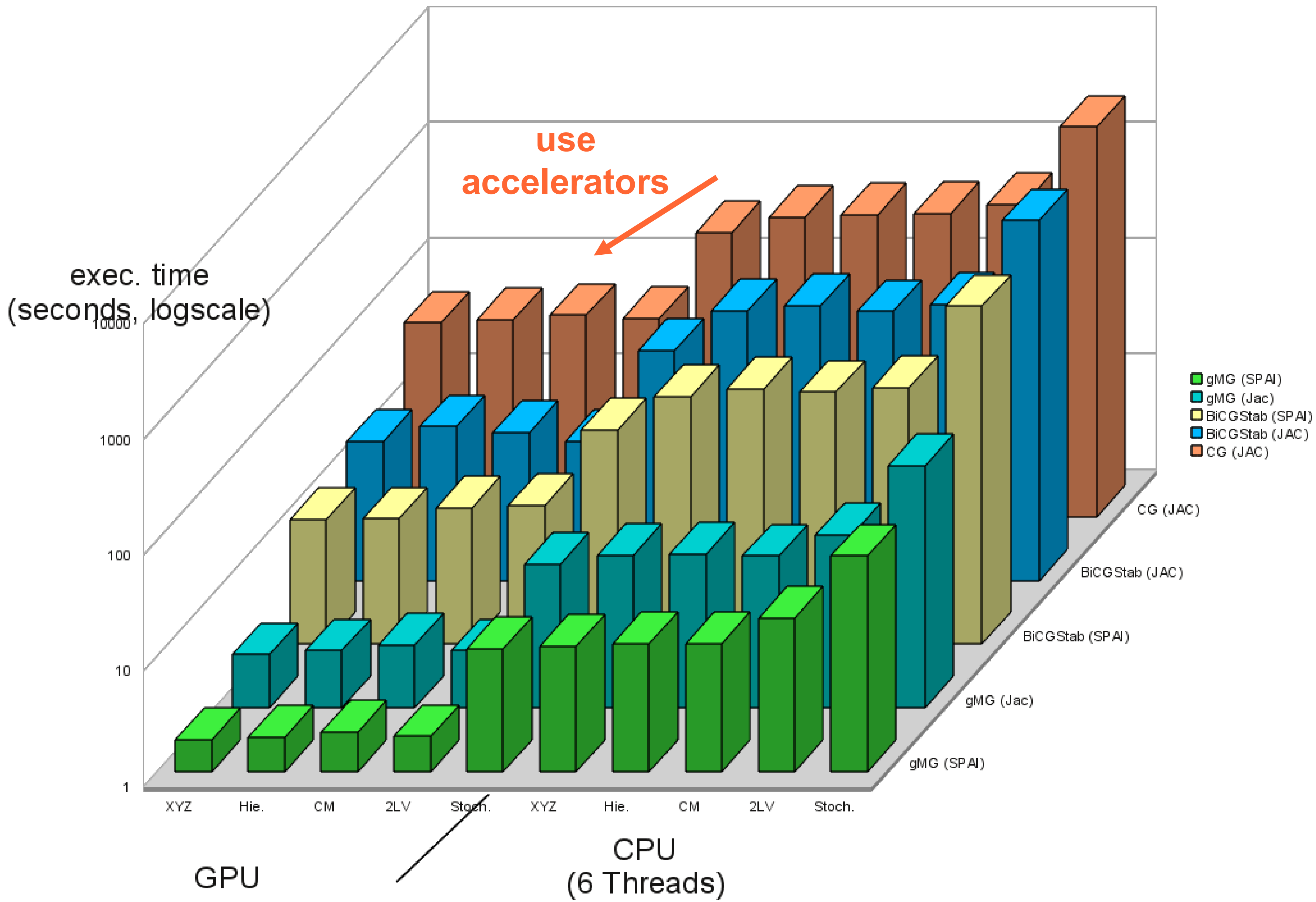
Hardware-oriented Numerics

(II) Numerical Efficiency



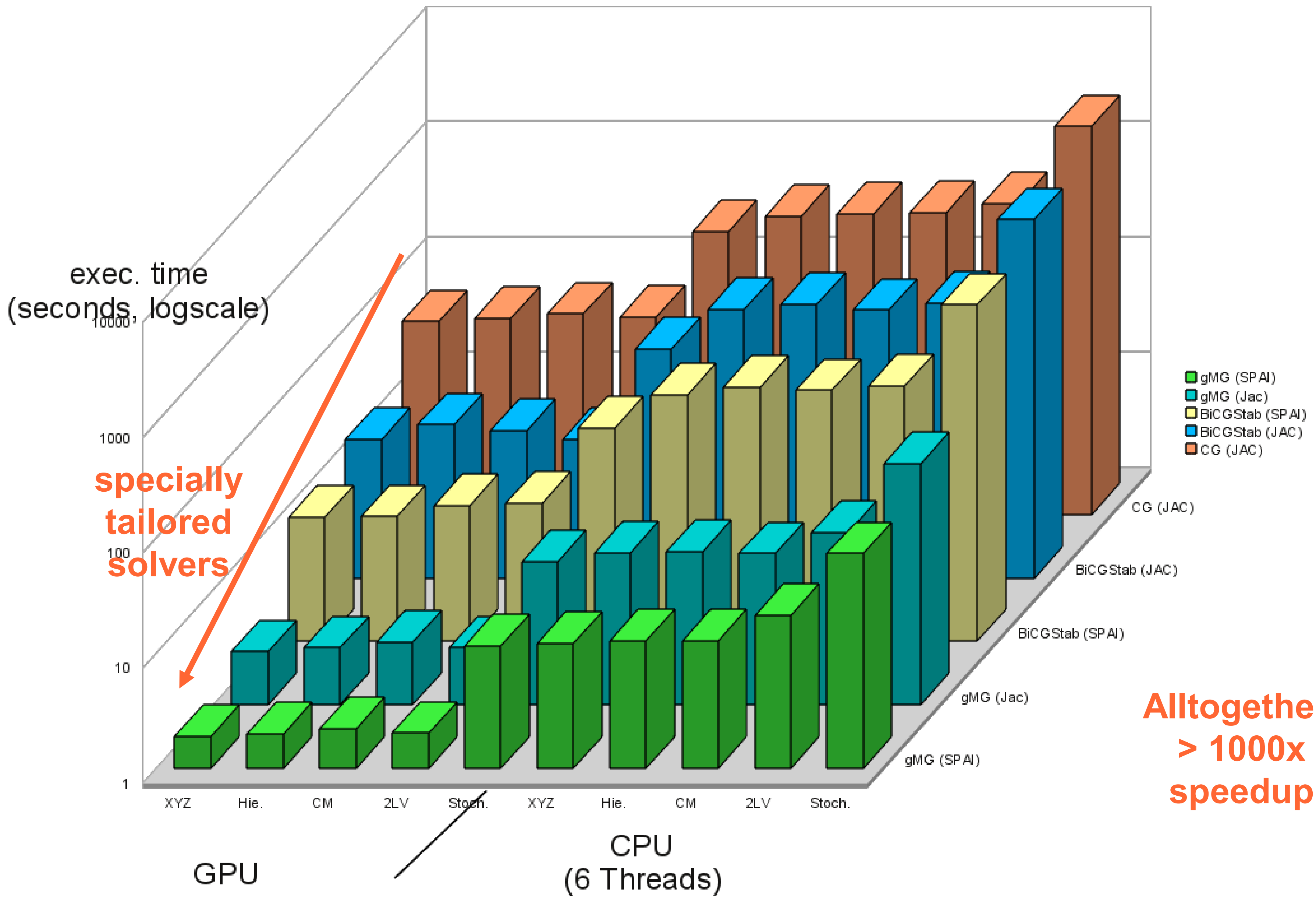
Hardware-oriented Numerics

(II) Numerical Efficiency



Hardware-oriented Numerics

(II) Numerical Efficiency

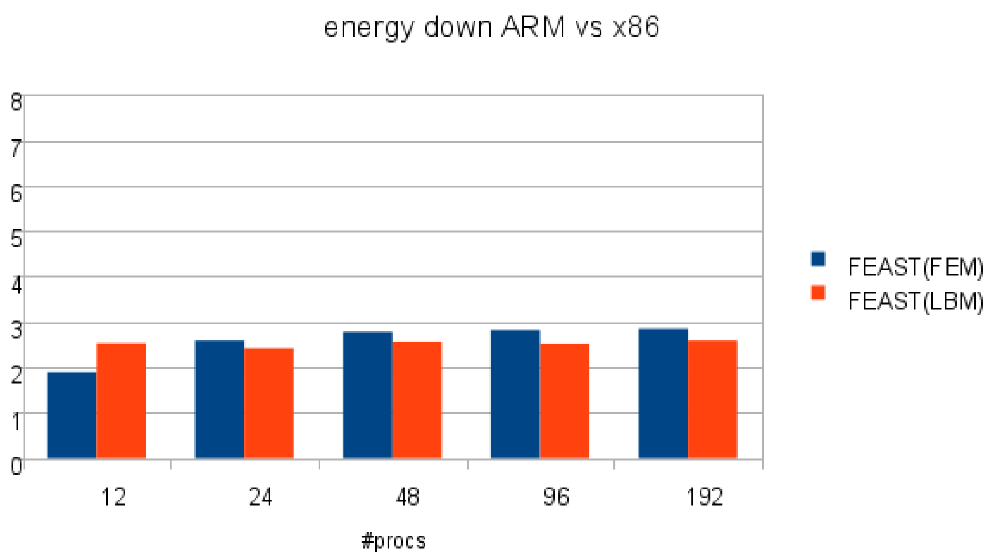


Alltogether: > 1000x speedup

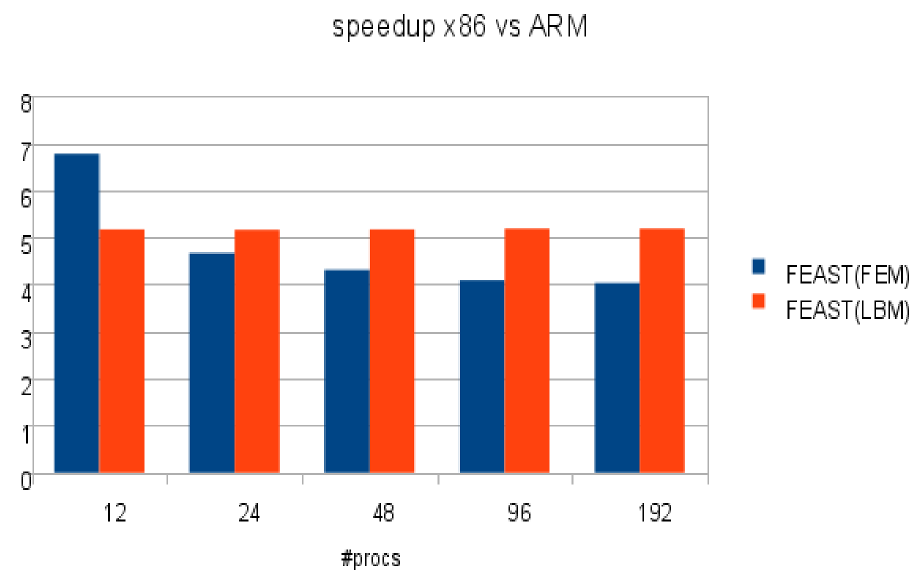
Hardware-oriented Numerics

(III) *Energy* Efficiency (?)

- energy consumption/efficiency is one of the major challenges for future supercomputers
- we can not afford to go all 'macho-flops' any more
- 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 (like powering down your x86)



~3x less energy



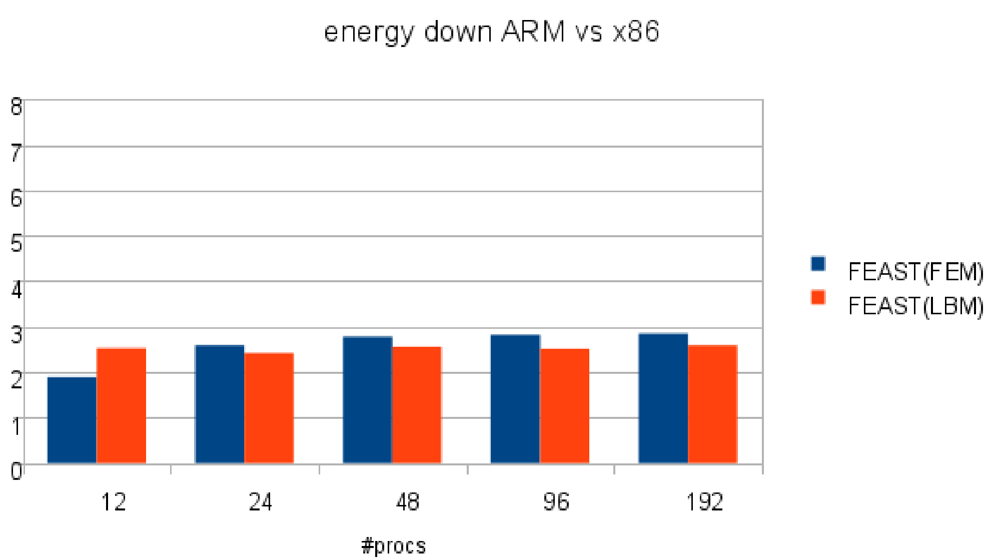
but: also ~5x more time!

Hardware-oriented Numerics

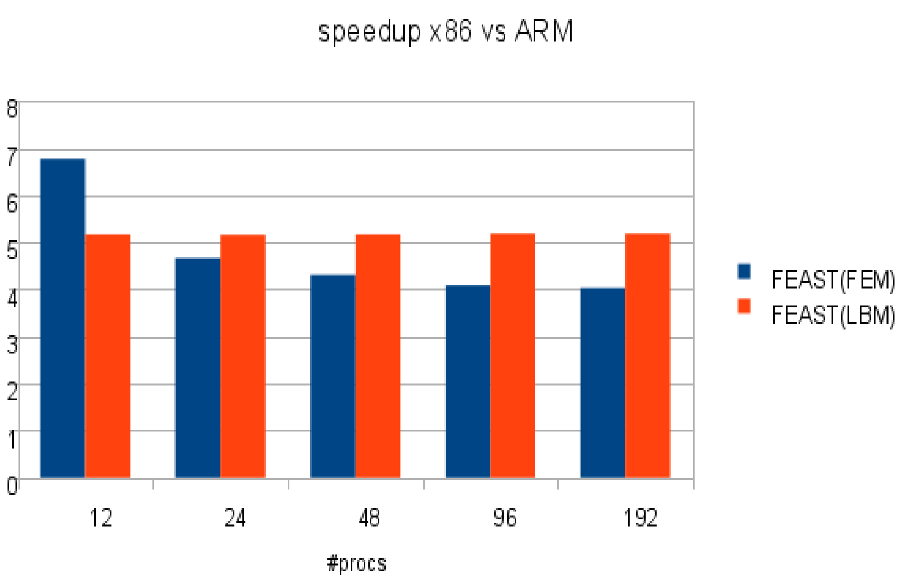
(III) *Energy* Efficiency (?)

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

→ More performance per Watt!



~3x less energy



but: also ~5x more time!

Hardware-oriented Numerics

(III) *Energy* Efficiency: technology of ARM-based SoCs since 2012

Something has been happening in the mobile-computing hardware evolution:

[one word in advance: there are many more SoC designs (like from TI, Qualcomm, …)]

→ 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

More importantly: 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

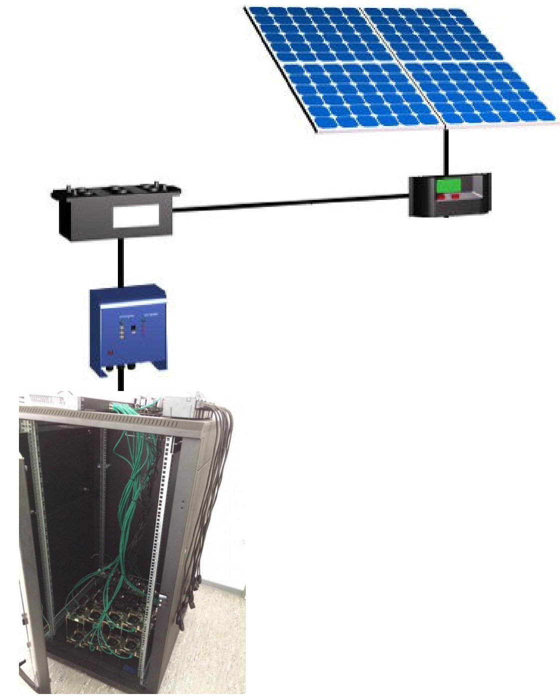
→ 2.5x higher EE promised

→ interesting for Scientific Computing! Higher EE than commodity!

Unconventional HPC for EE

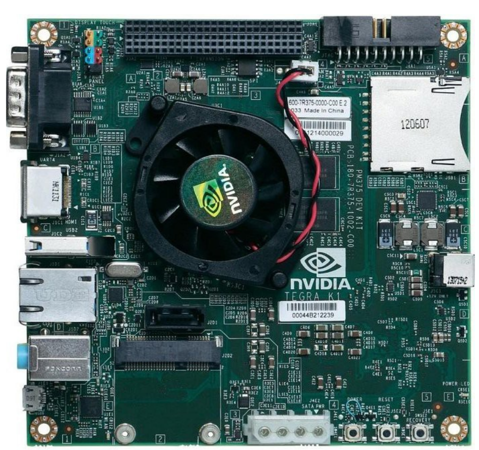
Bring together the two pillars of Energiewende for HPC

- Renewable power source
- Energy Efficiency



→ Design the hardware for EE!
→ Design the software for the hardware by using HWON!

x60



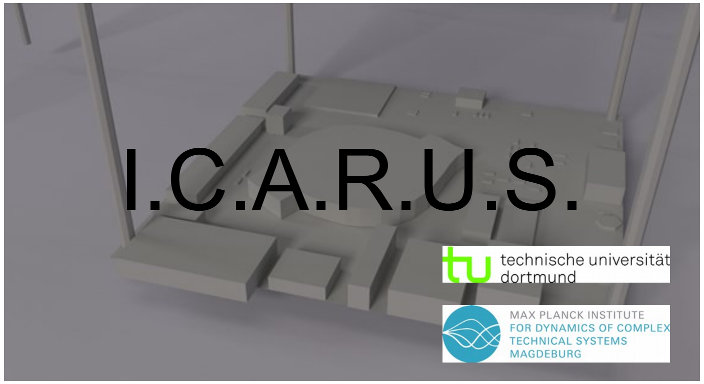
x1

x1

A 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 (photo-voltaic)
  → **specially tailored numerics and simulation software**: **high end Mathematics**
- **no future spendings due to energy consumtion**
- **SME-class resource: <80K€**
- **Scalability, modular design**
- (simplicity)
- (maintainability)
- (safety)
- ...

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

Whitesheet

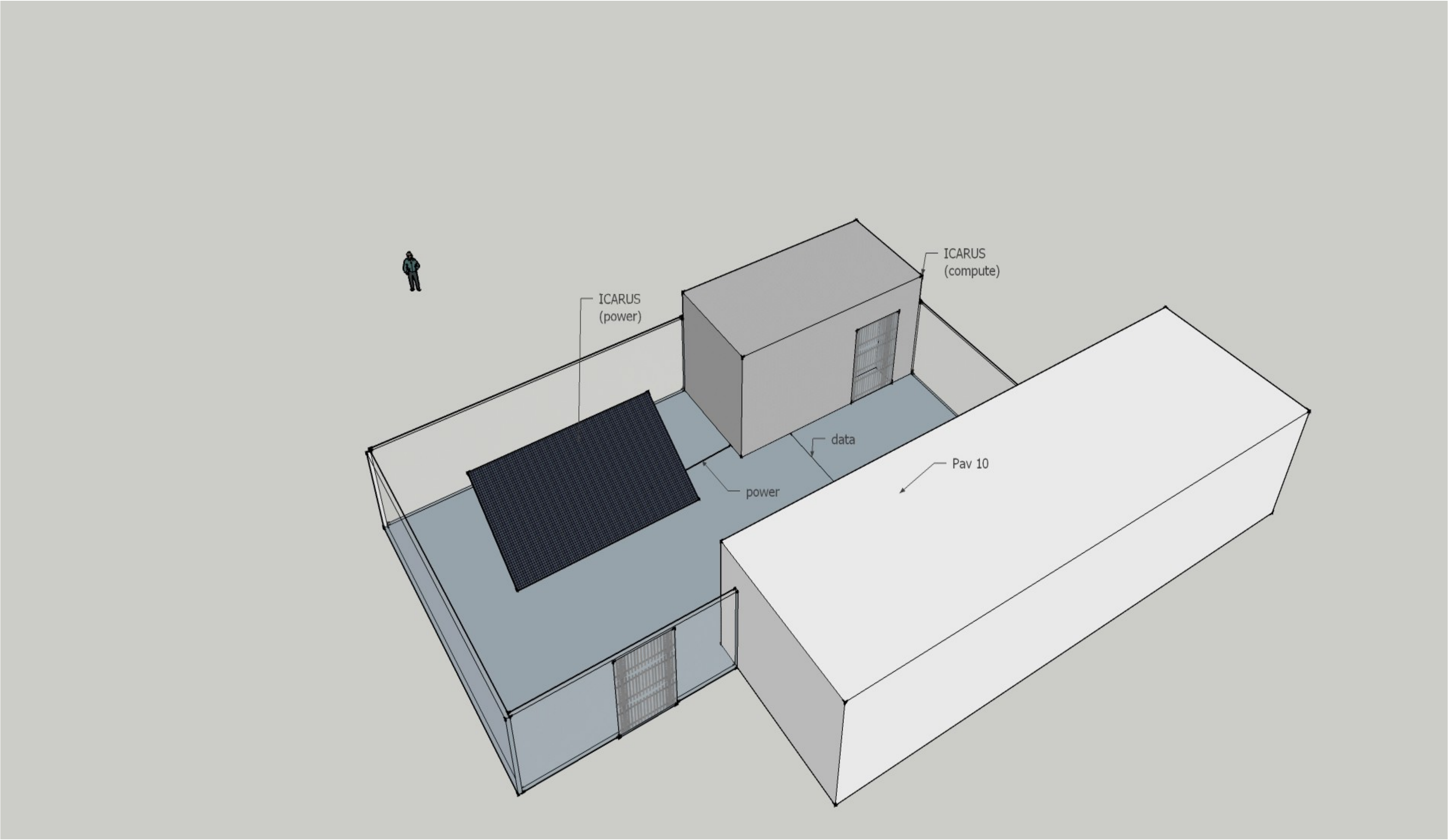
→ **nodes:** 60 x NVIDIA Jetson TK 1
→ #cores (ARM Cortex-A15): 240
→ #GPUs (Kepler, 192 cores): 60
→ RAM/core: 2GB LPDDR3
→ switches (GiBit Ethernet): 3xL1, 1xL2
→ **cluster theoretical peak perf: ~20TFlop/s SP**
→ **cluster peak power (including cooling/heating): < 2kW, provided by PV**

→ **storage:** 10+1 BananaPI Boards comprising:
  → 1 TB Western Digital Eco HDD
  → 2 Dual Core ARM (1 GHz,1 GB RAM)
  → GigabitEthernet networking
  → SATA
  → plus 16 GB eMMC internal (OS) and 128 GB SD swap / scratch per node

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

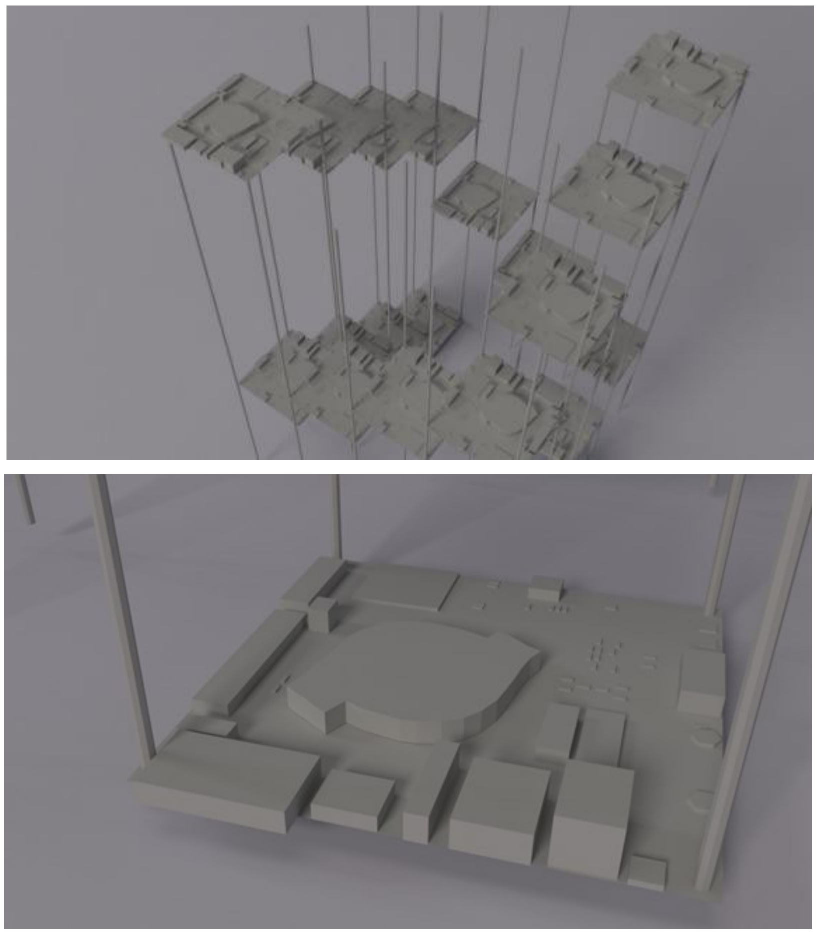
I.C.A.R.U.S construction site

solar modules delivering 6kWp, cluster built into container



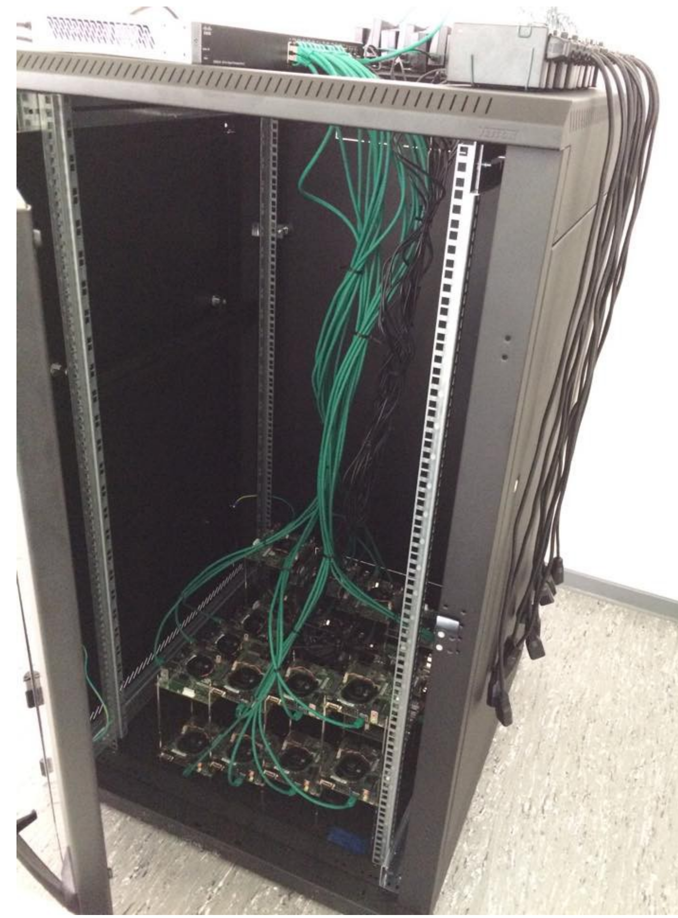
Progress

Two student projects Technomathematik @TU Do (Bachelor's and Master's levels) are on board (18 students)



...from heat- and airflow optimisation computer models...

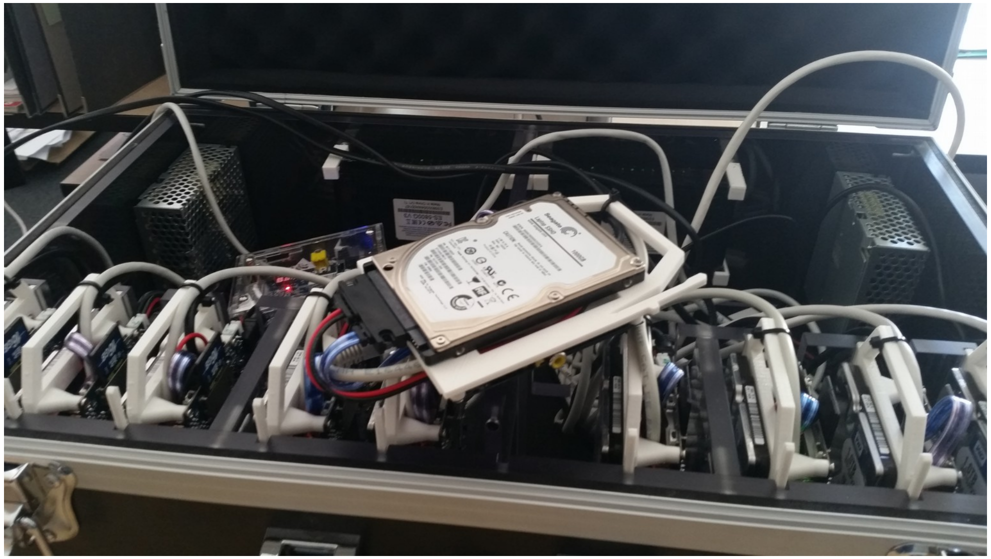
...and first test configurations...



...to fully operational rack with all hand-made compounds.

Progress

Storage subsystem completely operational @ MPI Magdeburg

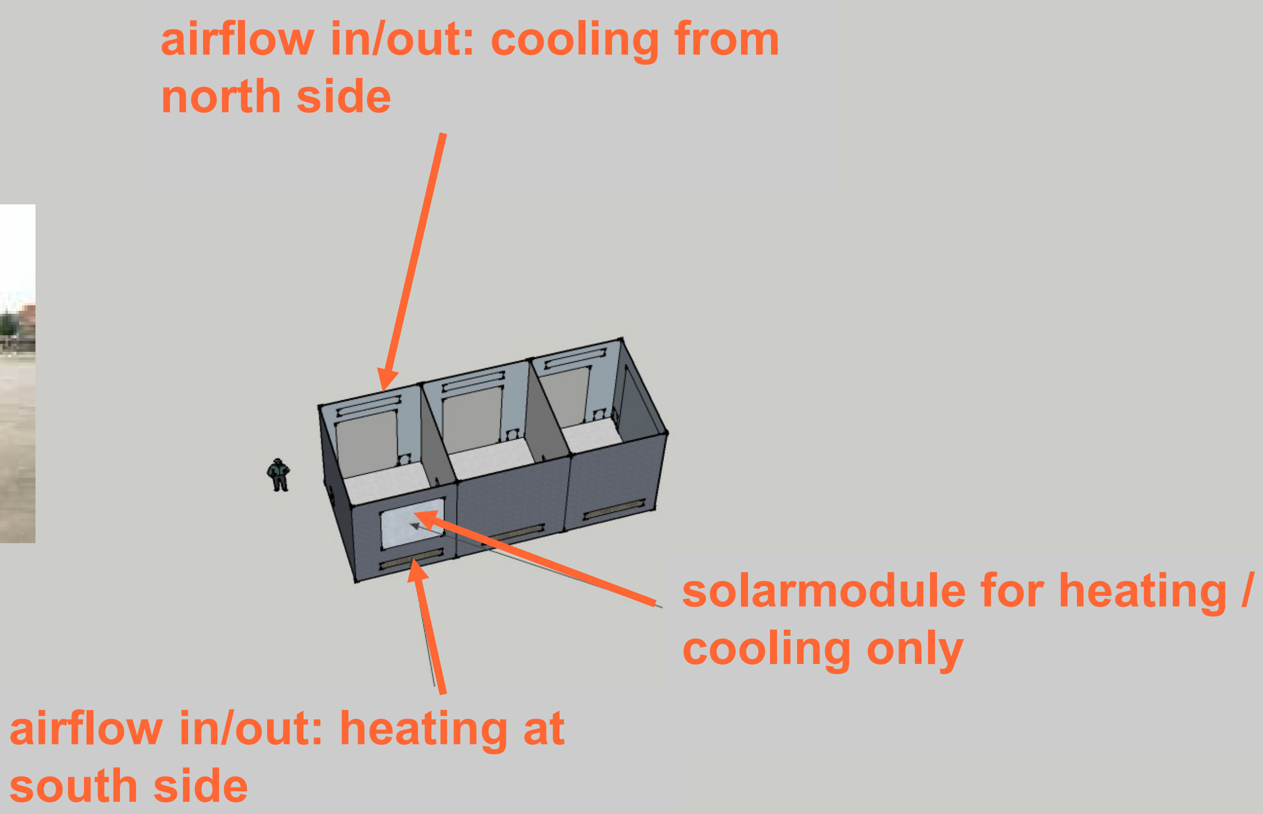


...fully portable, self-contained max. 10 TB storage...

...all storage BananaPi elements on self-made, 3D-printed mounts.

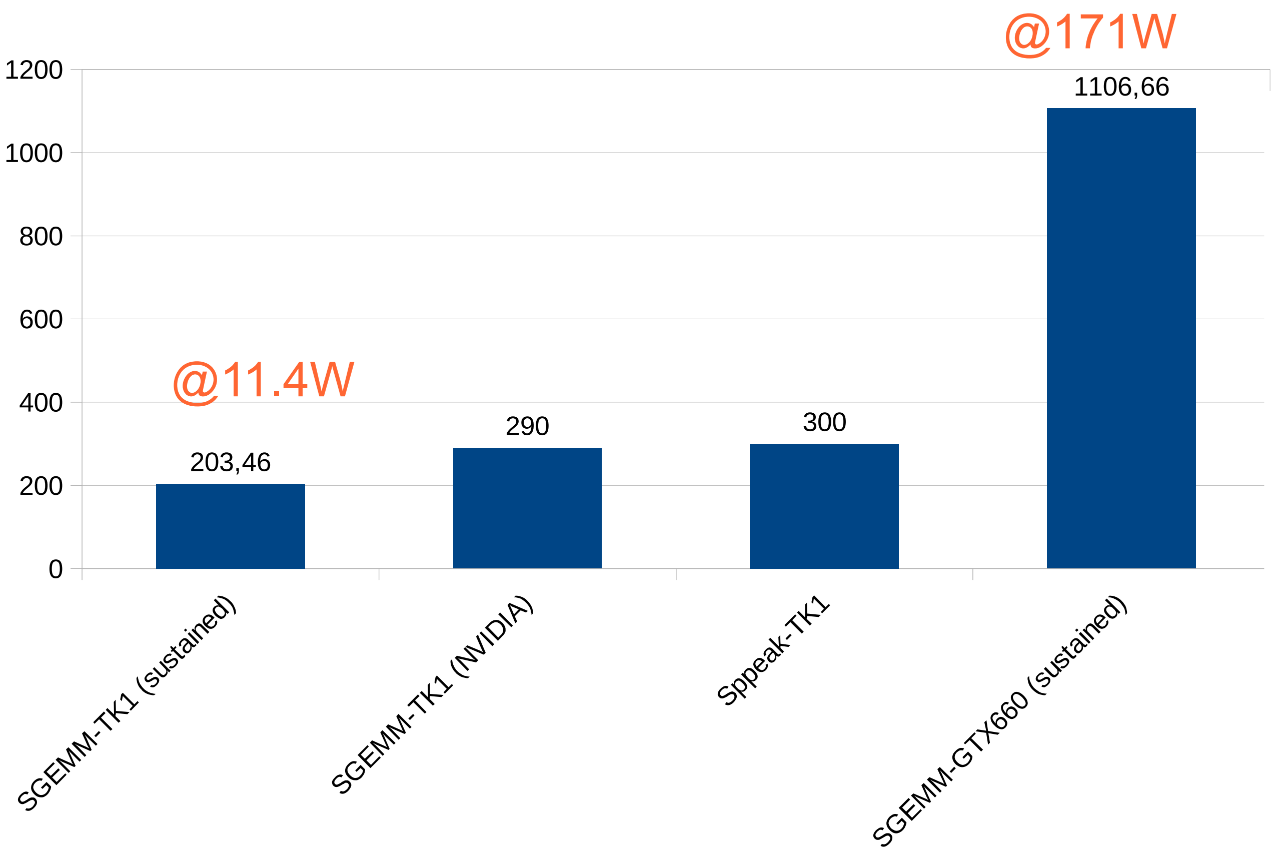
Progress

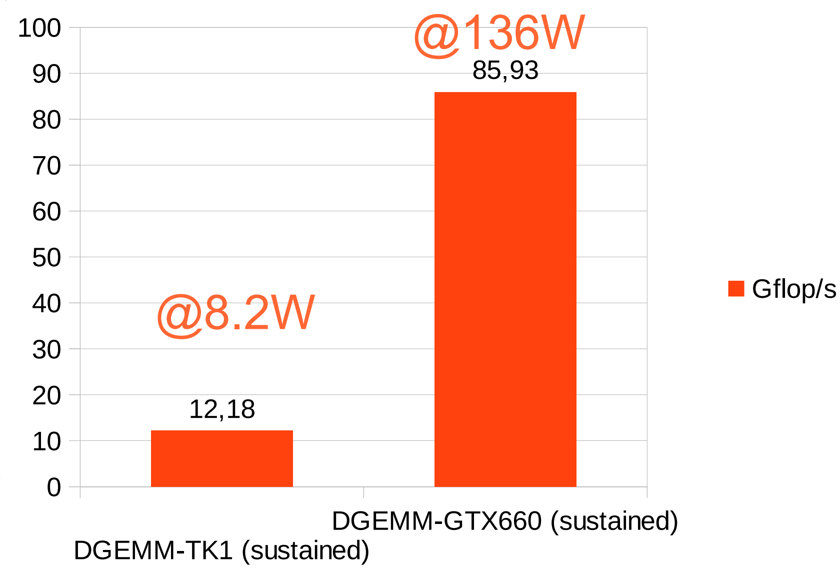
Housing and PV under construction in Dortmund



Power consumption and performance of basic kernels

S/DGEMM on the GPUs

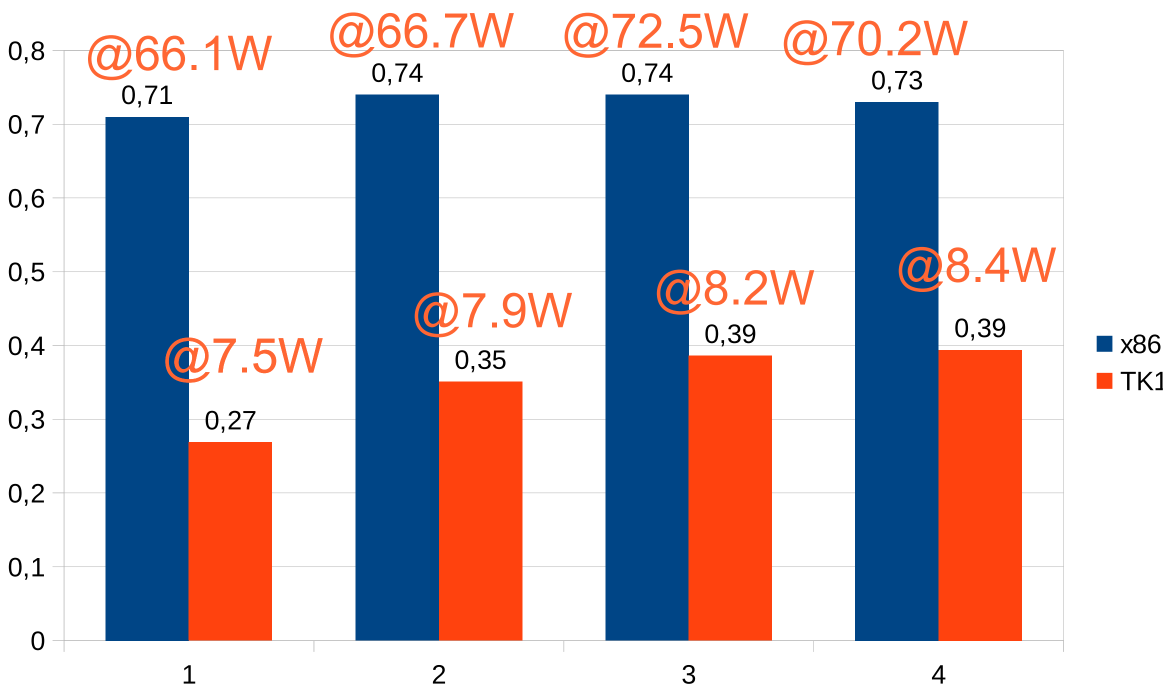


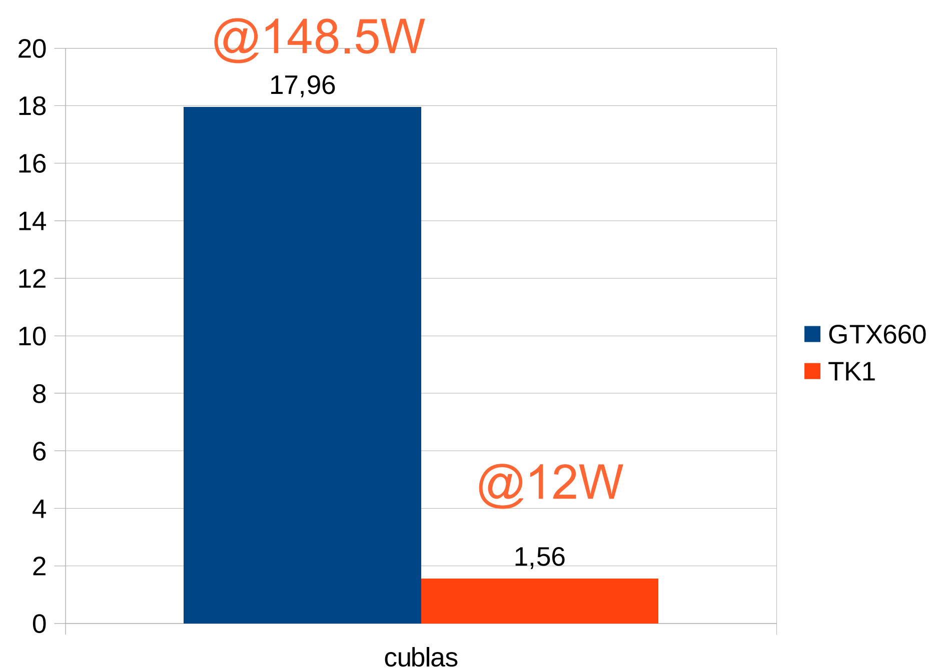


→ TK1 Kepler: **17.85 Gflop/s/W** SP, **1.49 Gflop/s/W** DP
→ GTX660: **6.47 Gflop/s/W** SP, **0.631 Gflop/s/W** DP
→ why SP matters: we can use **mixed precision methods** on a node
→ (Jetson) TK1 is 2-3 times better in this metric

Power consumption and performance of basic kernels

SAXPY (triad) **(float from now: mixed precision)**





→ core occupancy can be seen in power consumption
→ Cortex-A15: **0.05 Gflop/s/W**
→ IvyBridge: **0.01 GFLop/s/W**
→ TK1-Kepler: **0.13 Gflop/s/W**
→ GTX660: **0.12 Gflop/s/W**

Energy cost

SAXPY (triad) E[Ws]

embedded vs commodity **GPU**: x3

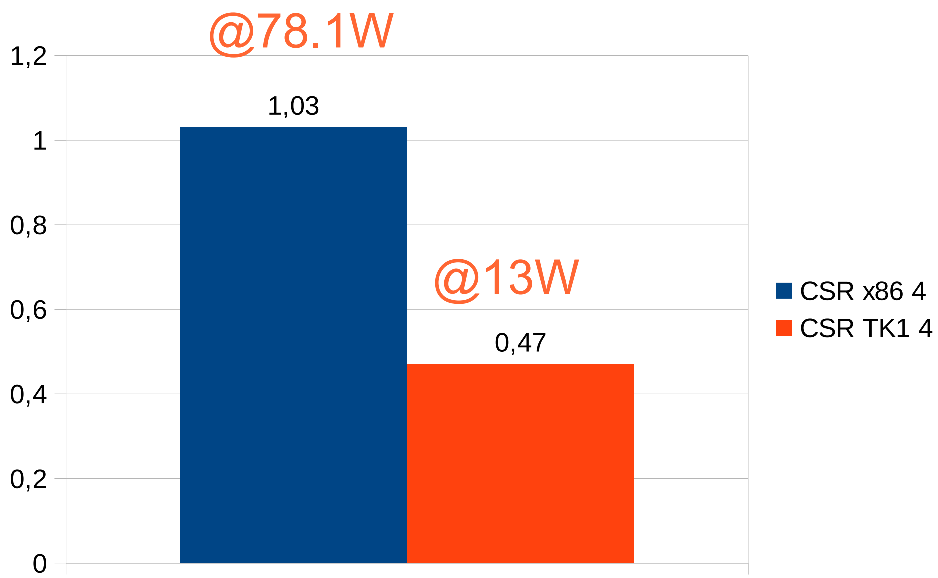
#DOFs	534144			
	TK1 CPU 4	TK1 GPU	IvyBridge 4	GTX660
WCT	0,01	0,0027	0,0057	0,00059
P	9,1	11	70,2	148,5
E	0,09	0,03	0,40	0,09
E / DOF	1,75E-007	5,56E-008	7,49E-007	1,64E-007

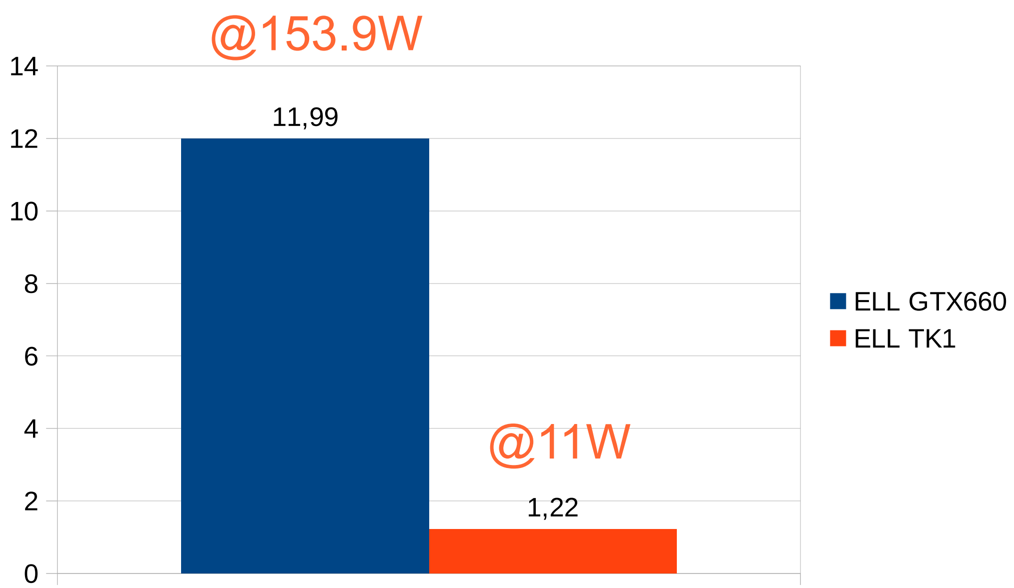
embedded vs commodity **CPU: x3 - x4**

→ Note: Perf.-Engineering for EE is complicated: **higher performance per Watt, less energy consumption, larger WCT, larger speeddown than E-down** at the same time

Power consumption and performance of basic kernels

SpMV SP





→ Cortex-A15: **0.036 Gflop/s/W**
→ IvyBridge: **0.013 GFLop/s/W**
→ TK1-Kepler: **0.11 Gflop/s/W**
→ GTX660: **0.077 Gflop/s/W**

Multigrid

Poisson Problem, 8x10E6 unknowns, 4/4 smoother steps, CSR/ELL, DP

			CPU			
		#iters	WCT	speeddown	P	P-down
Ivy + GTX660	Jac	10	6.58		88.90	
	SPAI	6	4.10		87.80	
Jetson TK1	Jac	10	15.90	2.42	8.10	10.98
	SPAI	6	10.10	2.46	8.10	10.84

All based on SpMV: coarse grid solver: PCG, smoother: Richardson, grid transfer:

$$(P_{2h}^{h})_{ij} = \varphi_{2h}^{(j)}(\xi_h^{(i)})$$

$$R_h^{2h} = (P_{2h}^{h})^T$$

GPU			
WCT	speeddown	P	P-down
0.55		151.50	
0.37		150.30	
4.70	8.55	9.40	16.12
2.80	7.57	9.50	15.82

The storage system by MPI Magdeburg

Results

MAX PLANCK INSTITUTE
FOR DYNAMICS OF COMPLEX
TECHNICAL SYSTEMS
MAGDEBURG

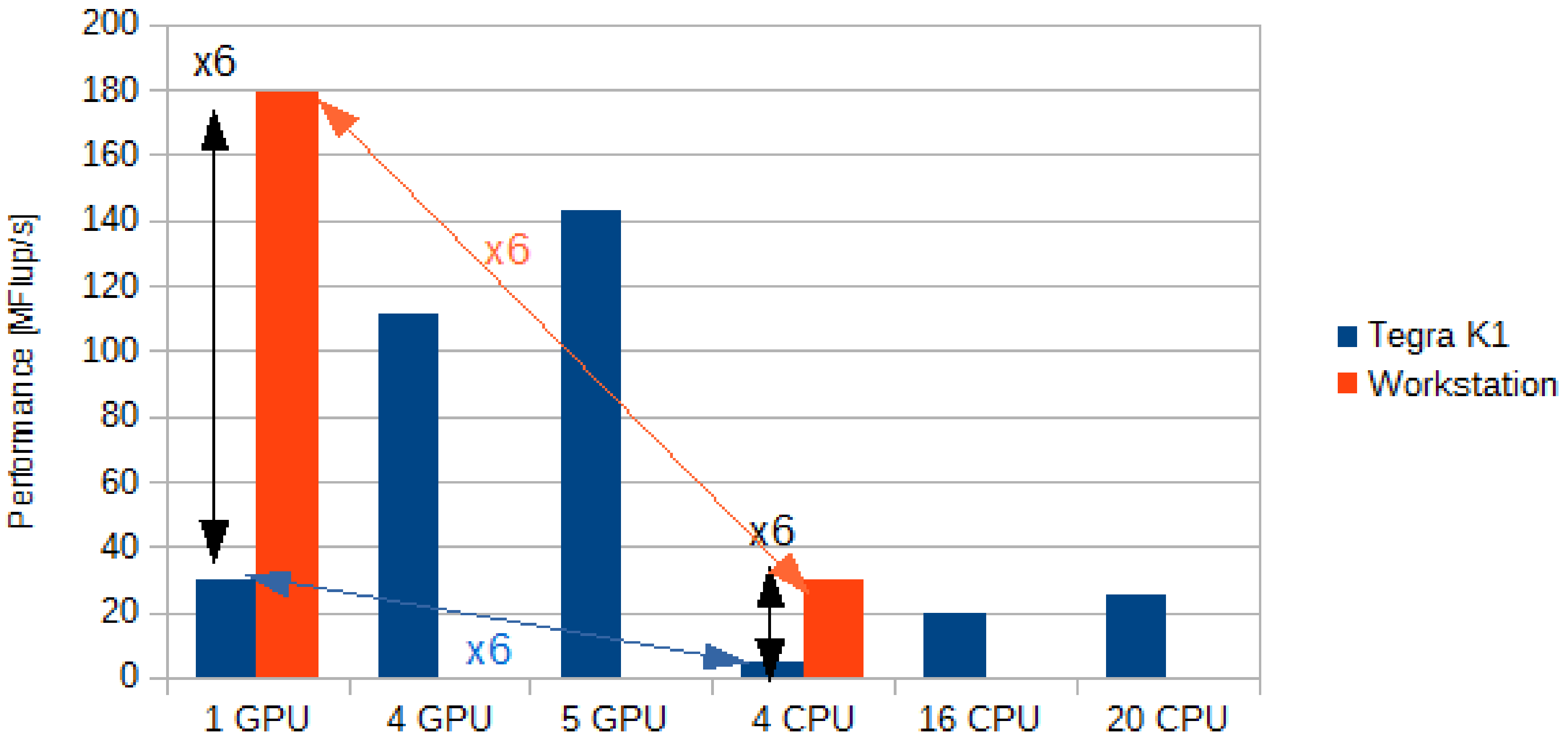
→ **Power dissipation at idle** (HDDs 'idle', no data access): 23W
→ **Maximum power dissipation at full operation:** 37W
→ **Average power dissipation: 30 Watt , 0.003W/GB (vs. 500W, 0.01W/GB commodity)**

→ **Configuration as RAID 0+1:**
- → 20 volumes with 250 GB each, 2x5 RAID 0 with 500 GB each + same as mirror (RAID 1) => 2.5 TB usable
- → max. write rate (single threaded): **55MB/s** **(vs. 130MB/s commodity)**
- → max. read rate (single thread): **71MB/s** **(vs. 130MB/s commodity)**

→ **Configuration as RAID 0:**
- → 20 volumes with 250 GB each, 2x10 RAID 0 with 500 GB each + same as mirror (RAID 0) => 5 TB usable
- → max. write rate (single threaded): **90MB/s** **(vs. 140MB/s commodity)**
- → max. read rate (single thread): **69MB/s** **(vs. 140MB/s commodity)**

Going multi-node

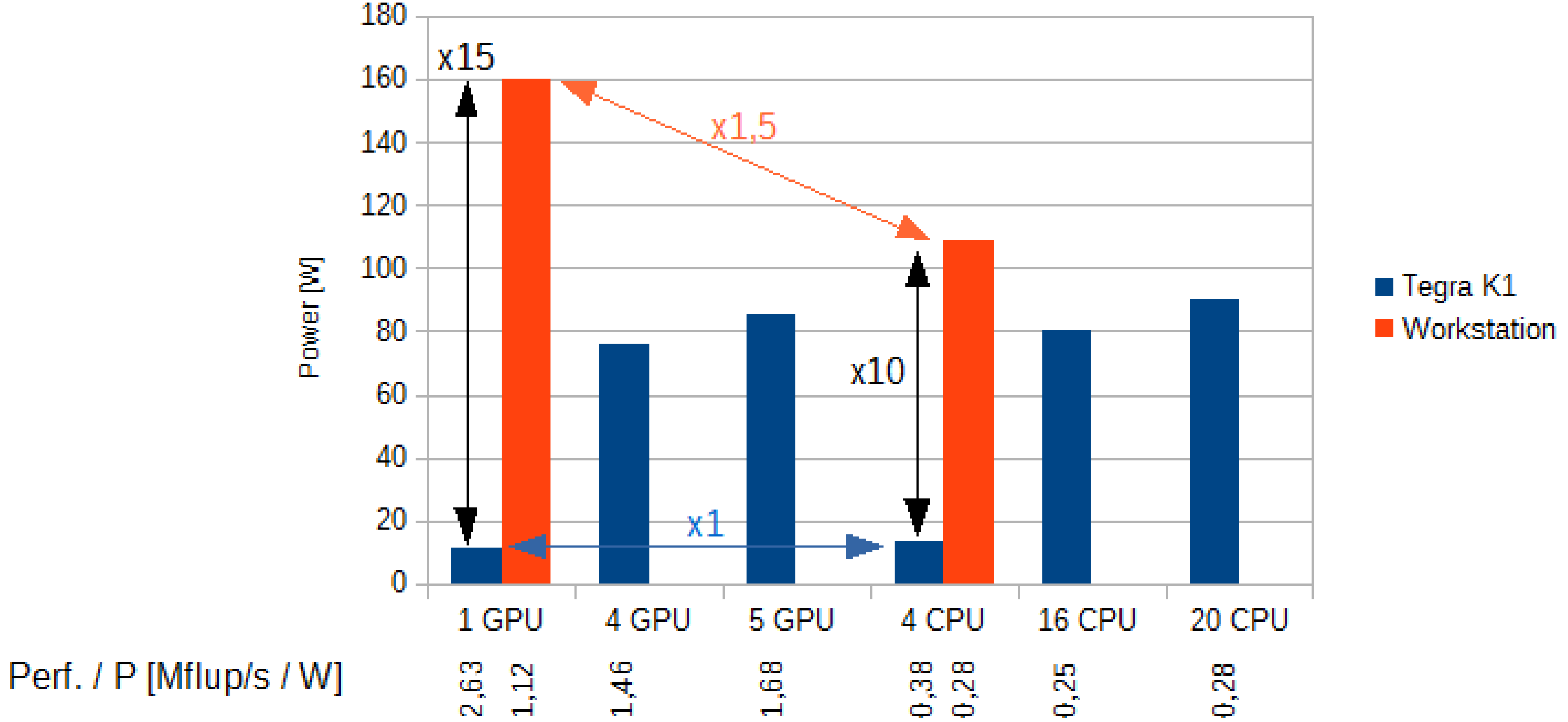
Flow solver on I.C.A.R.U.S. (Tier-0), FEAT software family



→ result: with 7 Jetson boards, we can beat this GPU,
even taking the whole storage cluster (30W average)

Going multi-node

Flow solver on I.C.A.R.U.S. (Tier-0), FEAT Software Family

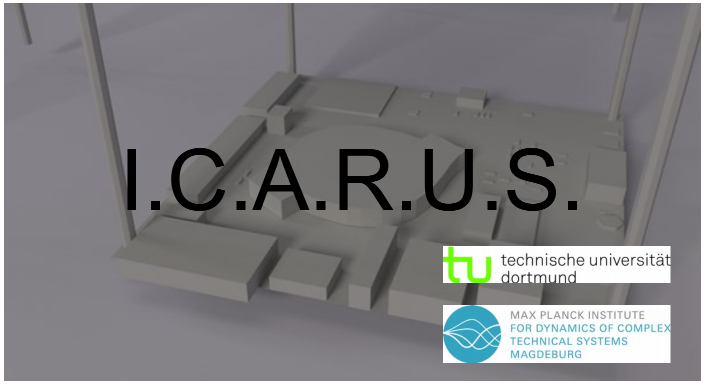


→ result: with 7 Jetson boards + switch, we can beat this GTX GPU
even taking the whole storage cluster (30W average)

→ this would take 123W **(153W with storage)** (switches, storage increase baseline)

→ **EE can be transported to the cluster level when combining UCHPC, HWON**

Conclusion



- EE requires us to rethink simulation
  from the energy-consumers' point of view

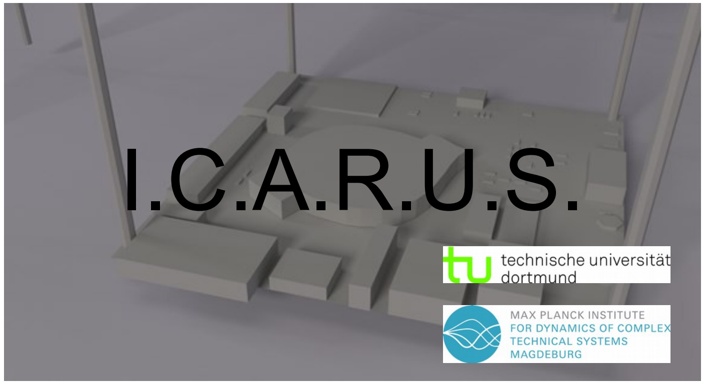
- HWON is threefold now: EE comes into play
- → smaller power dissipation alone is not the deal
  → performance modelling/-engineering of software for EE is needed

- Hardware-/Software Co-Design can be a starting point:
  → Embedded tech has a different history than commodity hardware
  → Energy Efficiency is just starting to arrive in HPC
  → System Integration with state-of-the-art PV tech (or other renewables) is promising

- The I.C.A.R.U.S. computer and its housing/energy-source plus the FEAT software together offers a valuable ressource aiming at SMEs/University departments

Bringing together high-end Mathematics / HWON with Unconventional HPC can ease the energy consumption of simulation.

Thank you



- Stefan Turek, Peter Benner (scientific supervision)
- Markus Geveler (system design)
- Dirk Ribbrock (system administration)
- Martin Köhler, Jens Saak, Gerry Truschkewitz (storage system design)

This work was also supported (in part) by the German Research Foundation (DFG) through the Priority Programme 1648 `Software for Exascale Computing' as well as grant TU 102/50-1.

I.C.A.R.U.S. hardware is financed by MIWF NRW under the lead of MERCUR.

Thanks to all students of the student projects Technomath 2015/2016 @ Tu Dortmund.

Thanks to Mechanical Workshop @ MPI Magdeburg.