703365 Sustainability in Computer Science:
Green HPC: Paving the Way for Sustainable Supercomputing

Philipp Gschwandtner, Research Center HPC, 30.10.2023

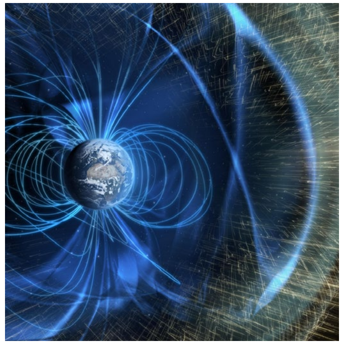
What is Green HPC?

“environmentally friendly”

High Performance Computing

“[…] uses supercomputers and computer clusters to solve advanced computation problems” (Wikipedia)

Green HPC

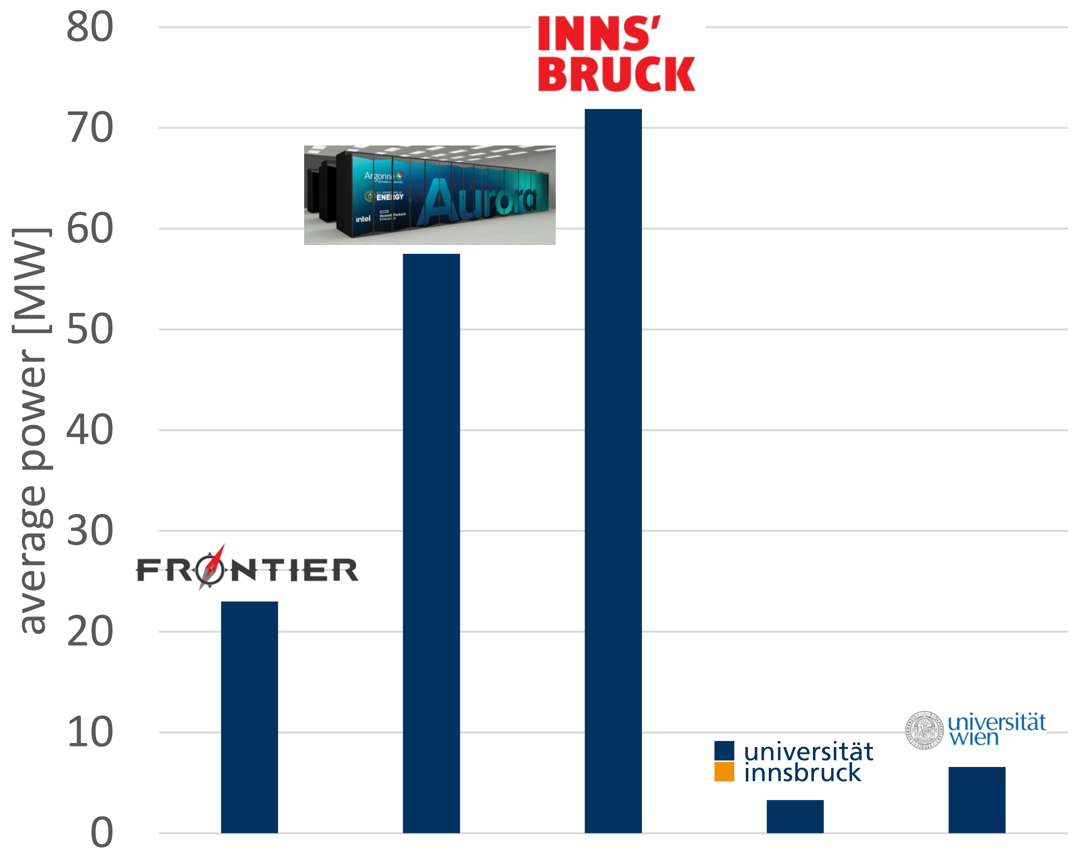


Why is energy consumption relevant in HPC?

- Example: Frontier
  - currently the fastest supercomputer world-wide, located in Oak Ridge, Tennessee, USA
- Extreme computing and storage capacities
  - 8,7 million cores
  - 9,2 petabytes of main memory (RAM)
  - 753 petabytes of storage
  - measured peak performance of ~1,2 exaflops (>10^{18} arithmetic operations per second)
- High electrical and thermal operating requirements
  - 23 megawatts of power
  - the **main limiting design factor** when building supercomputers

How much is 23 megawatts?

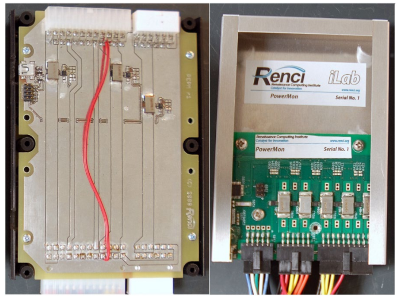
- 100% Frontier supercomputer
- 40% Aurora supercomputer (expected no. 1 in November 2023)
- 32% of Innsbruck
- 7x University of Innsbruck
- 3,5x University of Vienna



To be able to optimize, we need to measure first!

- Need power/energy instrumentation
  - “homemade” examples on the right, good for experimental research but does not scale to large systems (also: fire hazard!)
  - modern supercomputers have this built-in
  - note that there is also a plethora of power/energy models – some better, some worse

Voltech PM1000+



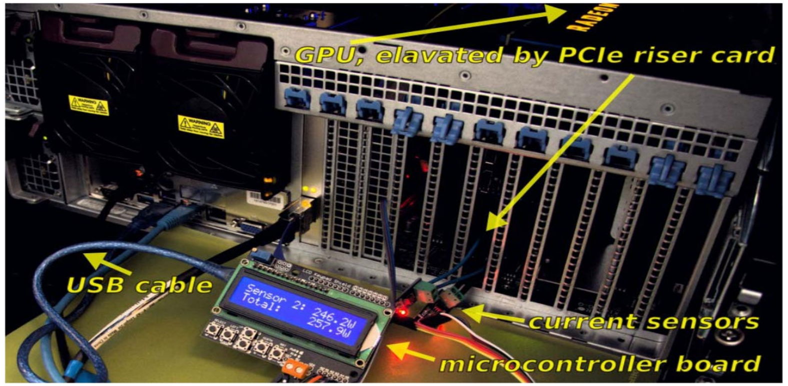
PowerMon2

- Need metrics to be able to evaluate, compare and optimize
  - e.g. Power Usage Effectiveness

$$\mathrm{PUE} = \frac{\text{total facility energy}}{\text{IT equipment energy}}$$

  - e.g. Energy-Delay Product (EDP)

$$\mathrm{EDP} = \text{energy} \times \text{walltime}$$



PowerSensor 2

TOP500 List

- List of the fastest supercomputers world-wide
  - released twice every year
  - high performance linpack (HPL) benchmarking (linear algebra stress-testing)
  - very interesting analyses and statistics around supercomputing and HPC
  - https://www.top500.org
- Currently June 2023 edition
  - Power consumption reported for many (but not all!) systems
- Also: Green500
  - Performance-per-energy ranking

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,699,904	1,194.00	1,679.82	22,703
2	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899
3	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	2,220,288	309.10	428.70	6,016
4	**Leonardo** - BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 SXM4 64 GB, Quad-rail NVIDIA HDR100 Infiniband, Atos EuroHPC/CINECA Italy	1,824,768	238.70	304.47	7,404

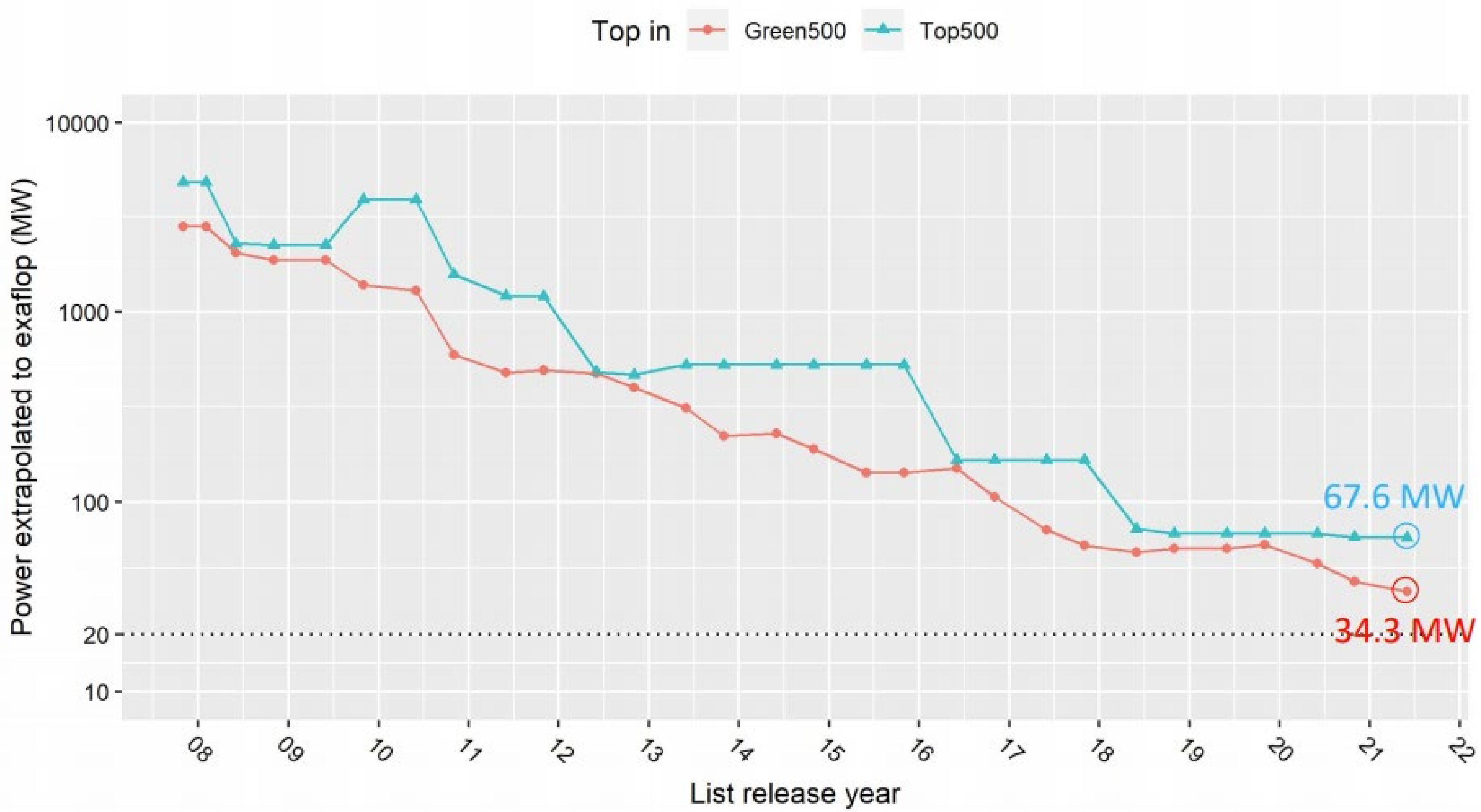
Green500 measurement methodology

- 33 pages of definitions: measurement devices, topology, workload requirements, averaging, etc.

- Level 1 requires to measure
  - The entire "core" phase ≥1 minute, compute-nodes + measure or estimate network interconnect
  - Power and take the average, at least `std::max({2 kW, 10% of the system, 15 nodes})`
- Level 2
  - Level 1 + average power of full run, intermediate measurements (at least 10 averages in core phase)
  - Compute-node subsystem + measure or estimate all other subsystems, at least `std::max({10 kW, 12% of the system, 15 nodes})`
- Level 3
  - Level 2 but measure energy and compute average power consumption
  - Energy measurement resolution: 120 Hz for DC, 5 KHz for AC, entire system (all components, all nodes, no extrapolations!)

Rank	TOP500 Rank	System	Cores	Rmax (PFlop/s)	Power (kW)	Energy Efficiency (GFlops/watts)
1	255	**Henri** - ThinkSystem SR670 V2, Intel Xeon Platinum 8362 32C 2.8GHz, NVIDIA H100 80GB PCIe, Infiniband HDR, Lenovo Flatiron Institute United States	8,288	2.88	44	65.396
2	34	**Frontier TDS** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	120,832	19.20	309	62.684
3	12	**Adastra** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE Grand Equipement National de Calcul Intensif - Centre Informatique National de l'Enseignement Suprieur (GENCI-CINES) France	319,072	46.10	921	58.021
4	17	**Setonix – GPU** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE Pawsey Supercomputing Centre, Kensington, Western Australia Australia	181,248	27.16	477	56.983

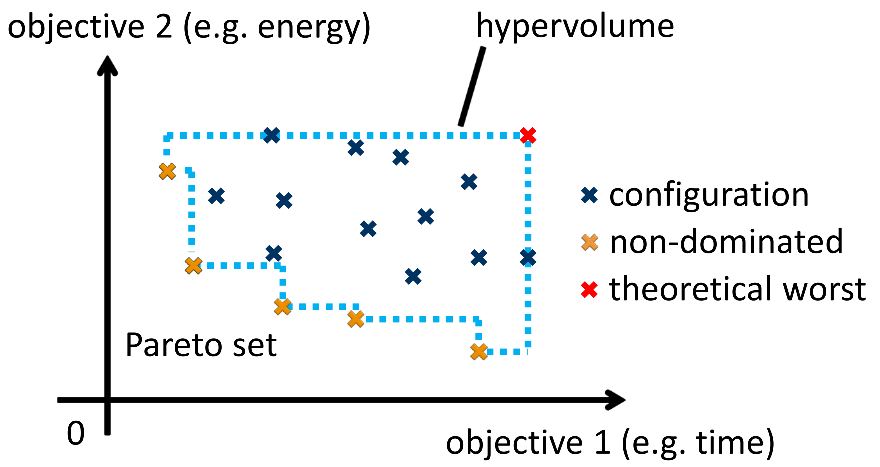
Power consumption projected to 1 exaflop



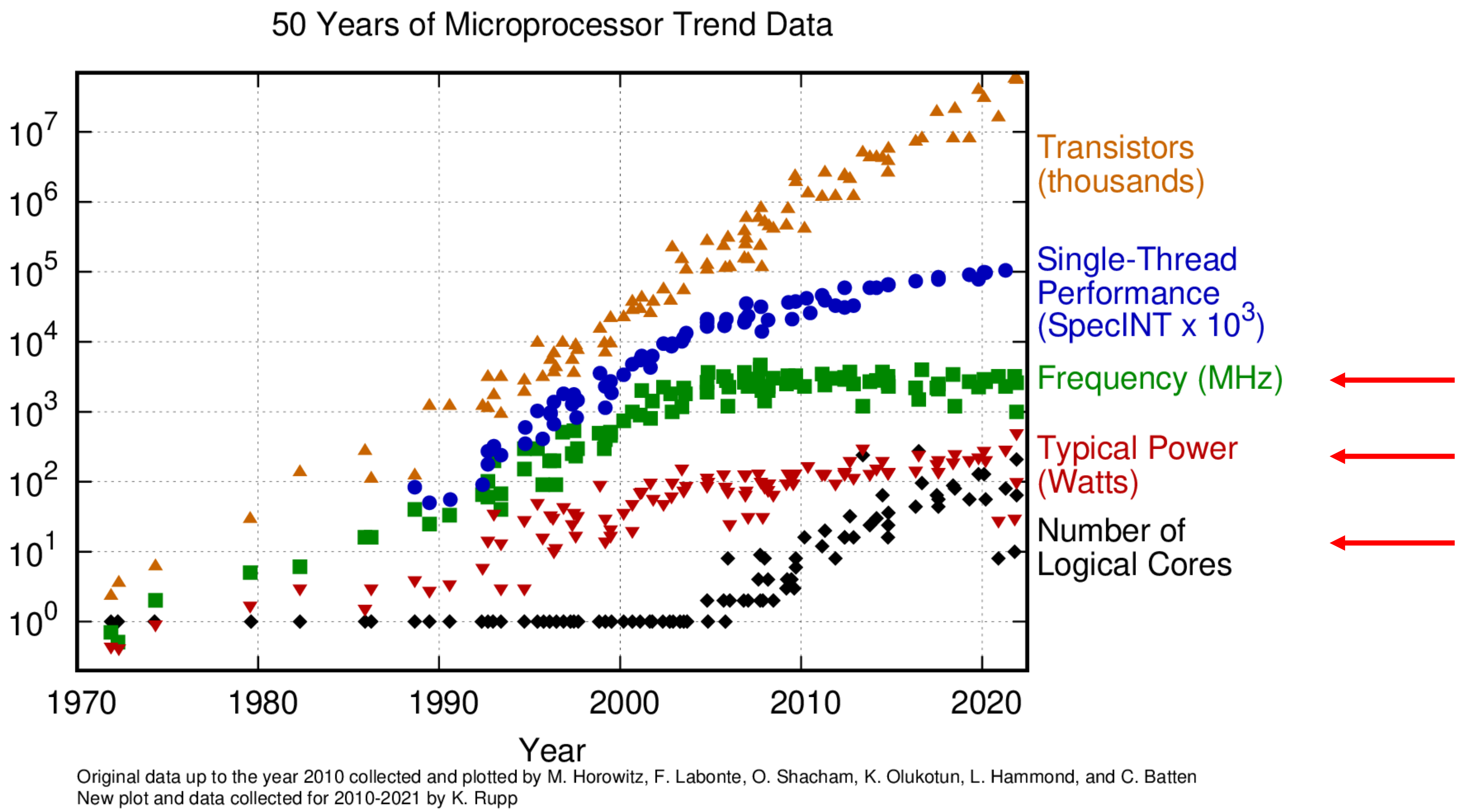
https://www.hpcwire.com/2021/07/15/15-years-later-the-green500-continues-its-push-for-energy-efficiency-as-a-first-order-concern-in-hpc/

Avenues of optimization

- Multiple attack points for making HPC more energy-efficient
  - increased parallelism
  - cooling
  - what and how the hardware is used
- When working with energy-efficient HPC, it's always a multi-objective problem
  - optimizing for power and/or energy often means sacrificing (a little bit of) performance
  - e.g. Pareto-optimality

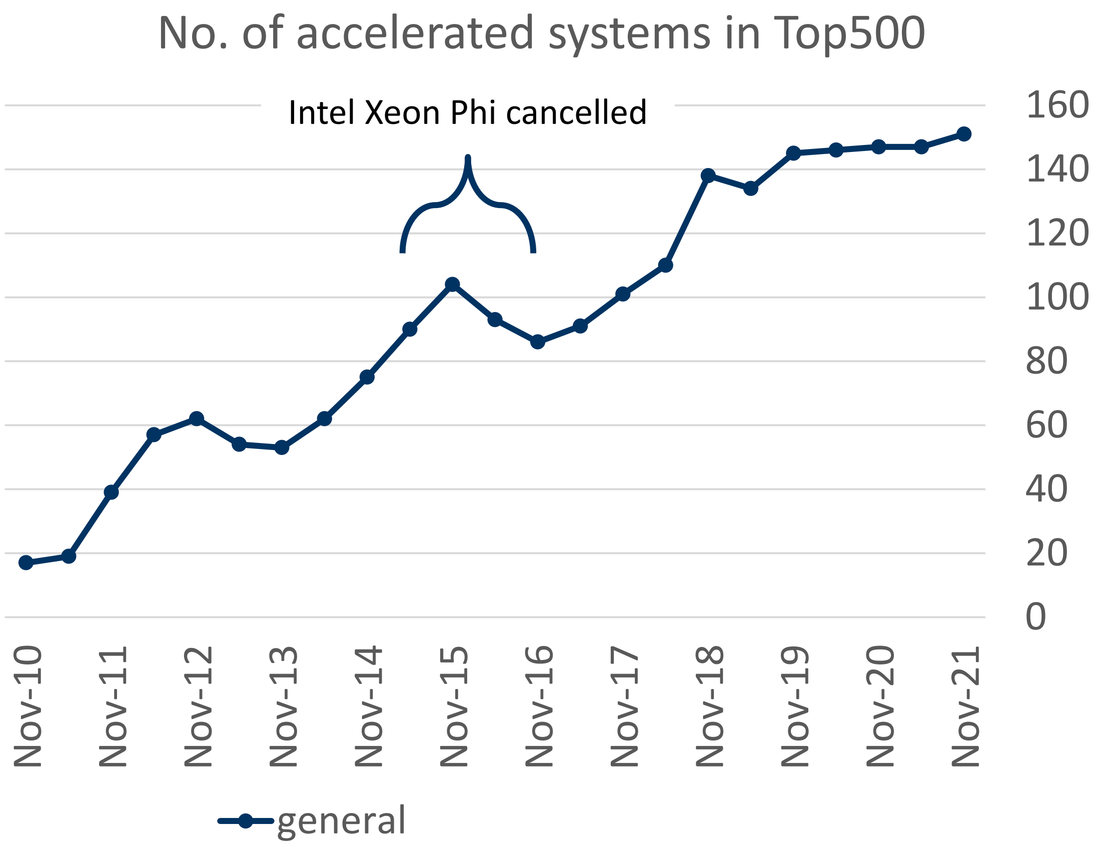


Reducing energy in computing: Parallelism!



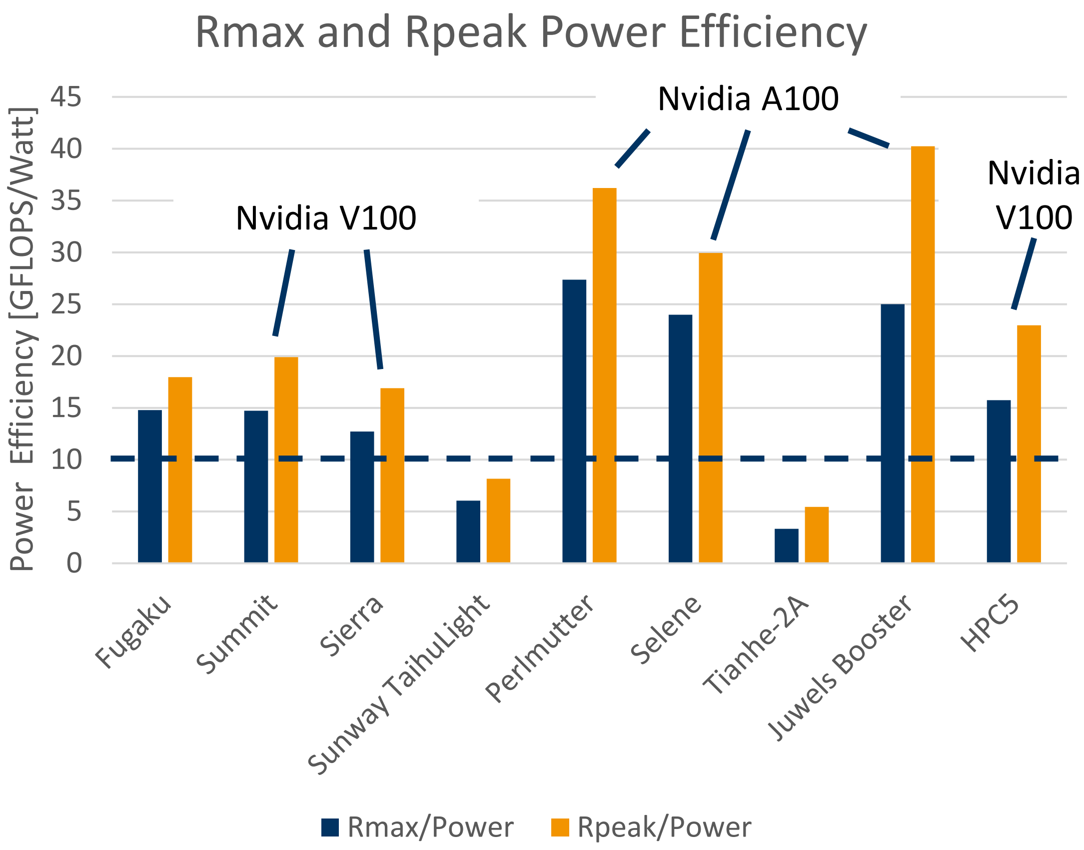
Reducing energy in computing: Accelerators!

- Accelerator market share in HPC has been steadily increasing and will likely continue to do so
  - Why? Distributed memory clusters with accelerators provide some of the best cost- and energy-efficiency in HPC
  - All 10 out of the top 10 entries in the November 2022 “Green 500” list are accelerator clusters (9/10 GPUs)

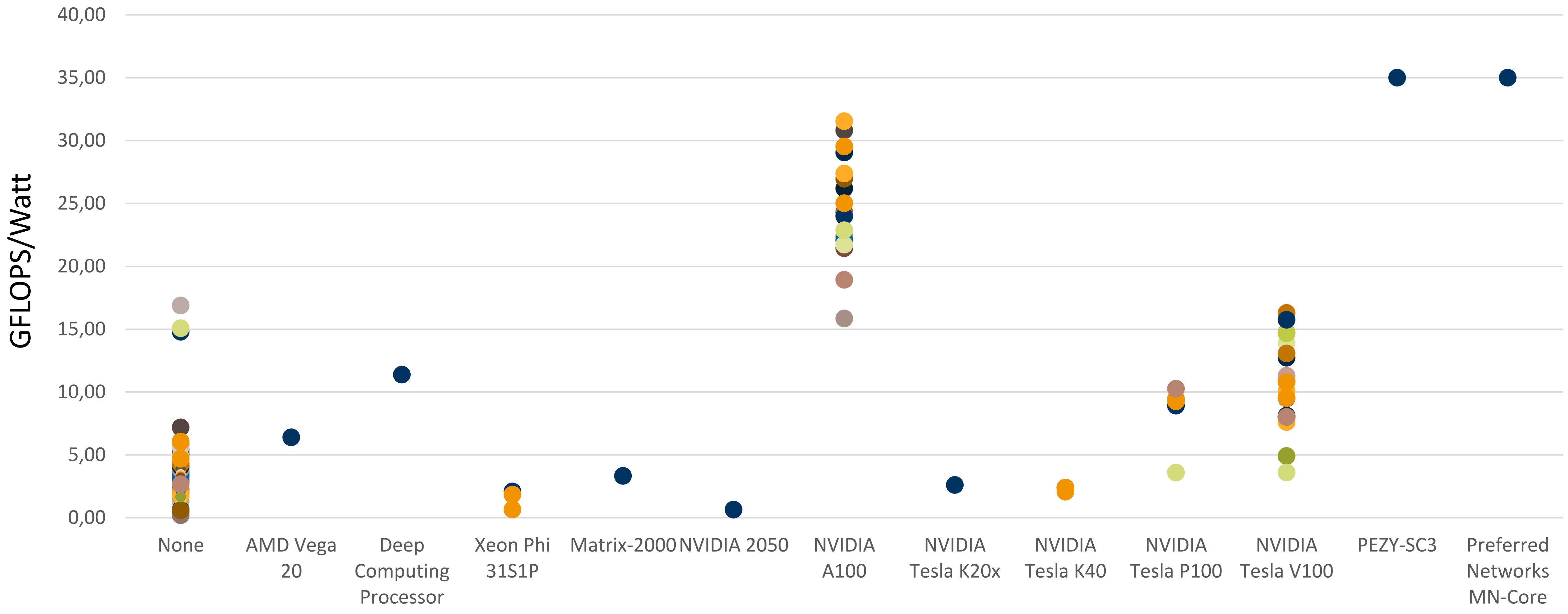


A closer look at power efficiency

- 8 of the top 10 systems above 10 GFLOPS/Watt use accelerators (2021 data)
- Exceptions:
  - Fugaku: ARM-based, no accelerators
  - Tianhe-2A: Matrix 2000 accelerators (128 core RISC CPUs)

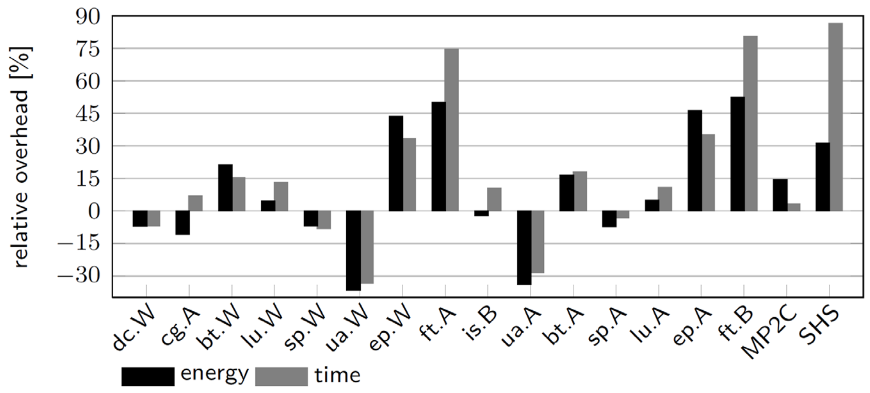


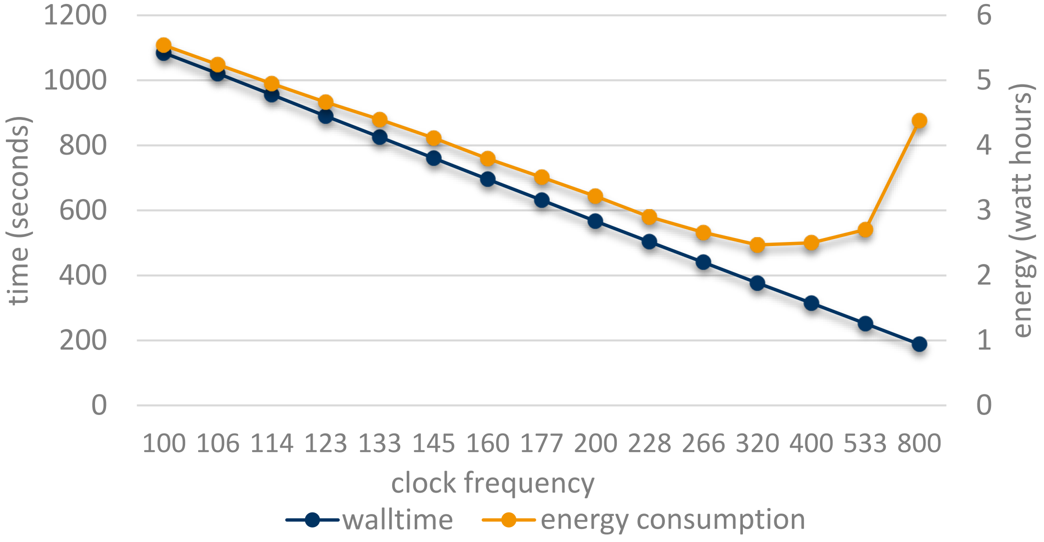
Power efficiency of all TOP500 systems (2021)



Reducing energy in computing: Tuning!

- Lots of research in software means for reducing energy consumption
- Top figure: effects of instruction mix on energy consumption of an IBM POWER7 CPU (using GCC vs. IBM XL compilers)
  - Result: In general, IBM XL produces more efficient binaries, but not always!
- Bottom figure: Dynamic Frequency and Voltage Scaling (DVFS) of the Intel SCC (experimental many-core CPU)
  - reduce clock frequency to save power and often also energy, effect heavily depends on workload
  - used in most CPUs these days (laptops, desktops, server, smartphones, etc.)
  - used on supercomputers (e.g. Energy Aware Runtime)

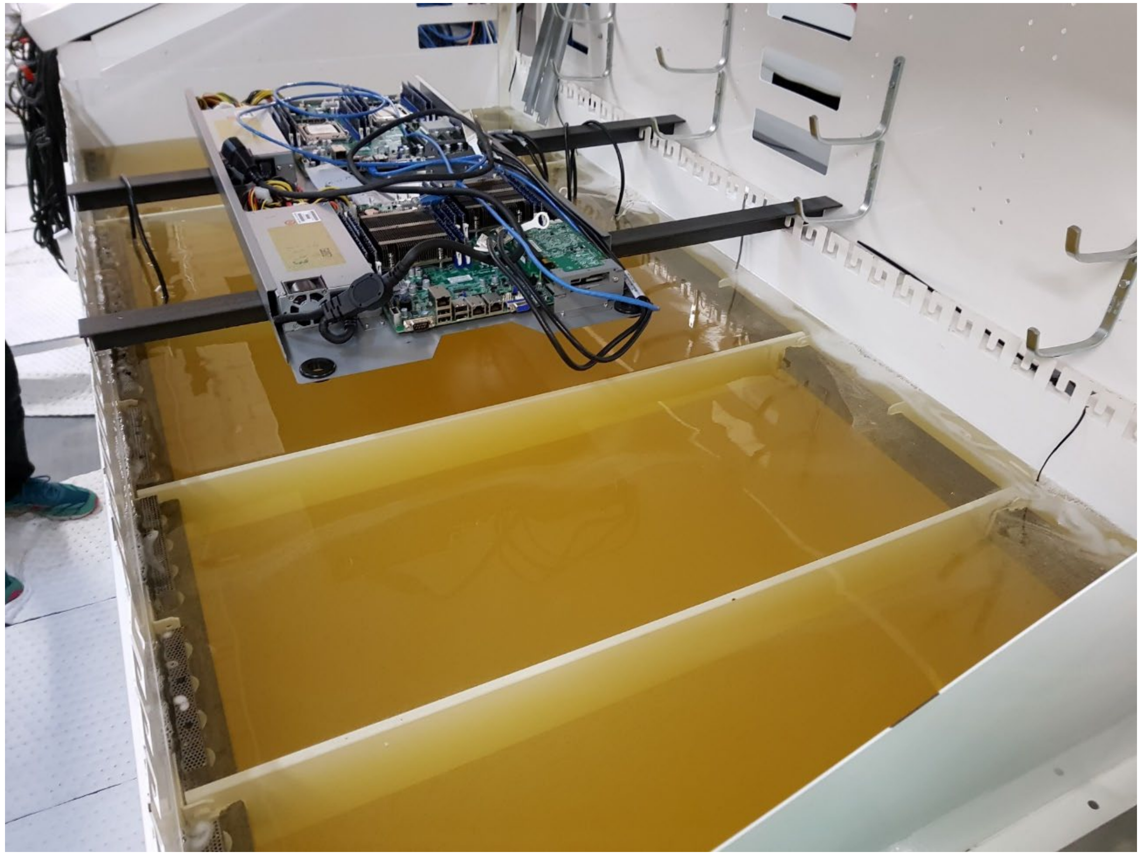




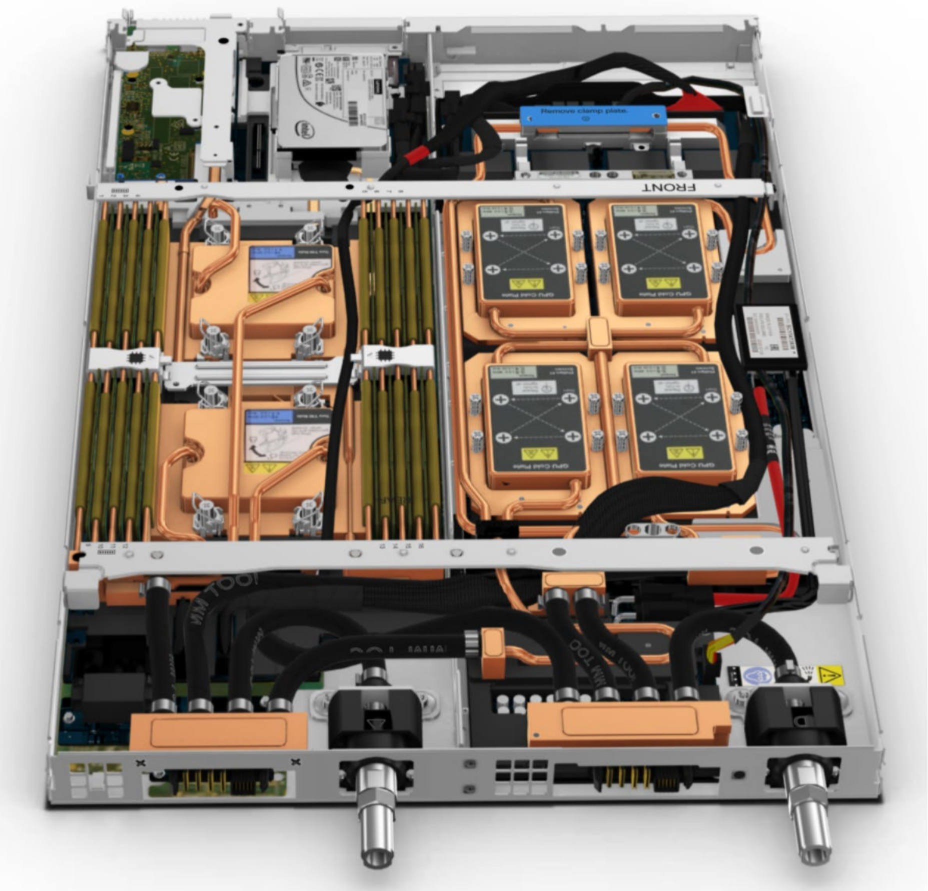
Reducing energy in cooling: use oil!

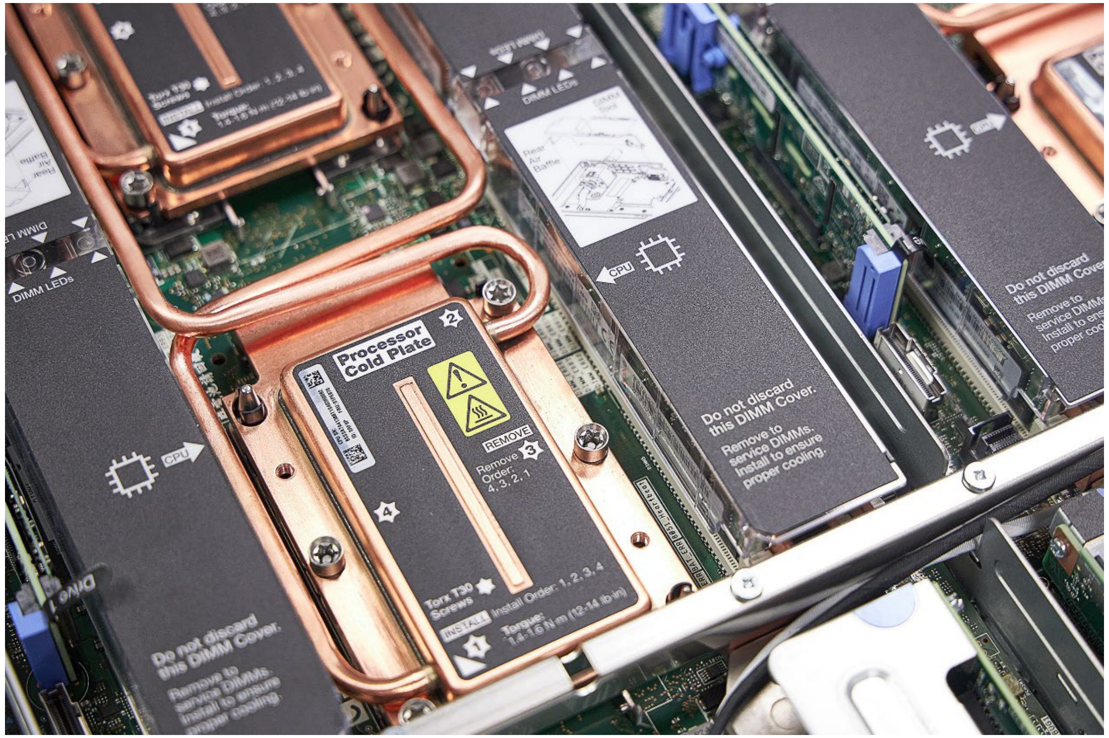
- VSC-3, fastest supercomputer in Austria in 2014
  - ranked 85th world-wide
  - 32.768 cores
  - 450 kW
  - mechanical PUE of **1.02**!
    - compare to VSC-2 (water-cooled): mPUE of 1.18
    - VSC-4 (water-cooled): 1.05

“Immersion cooling”

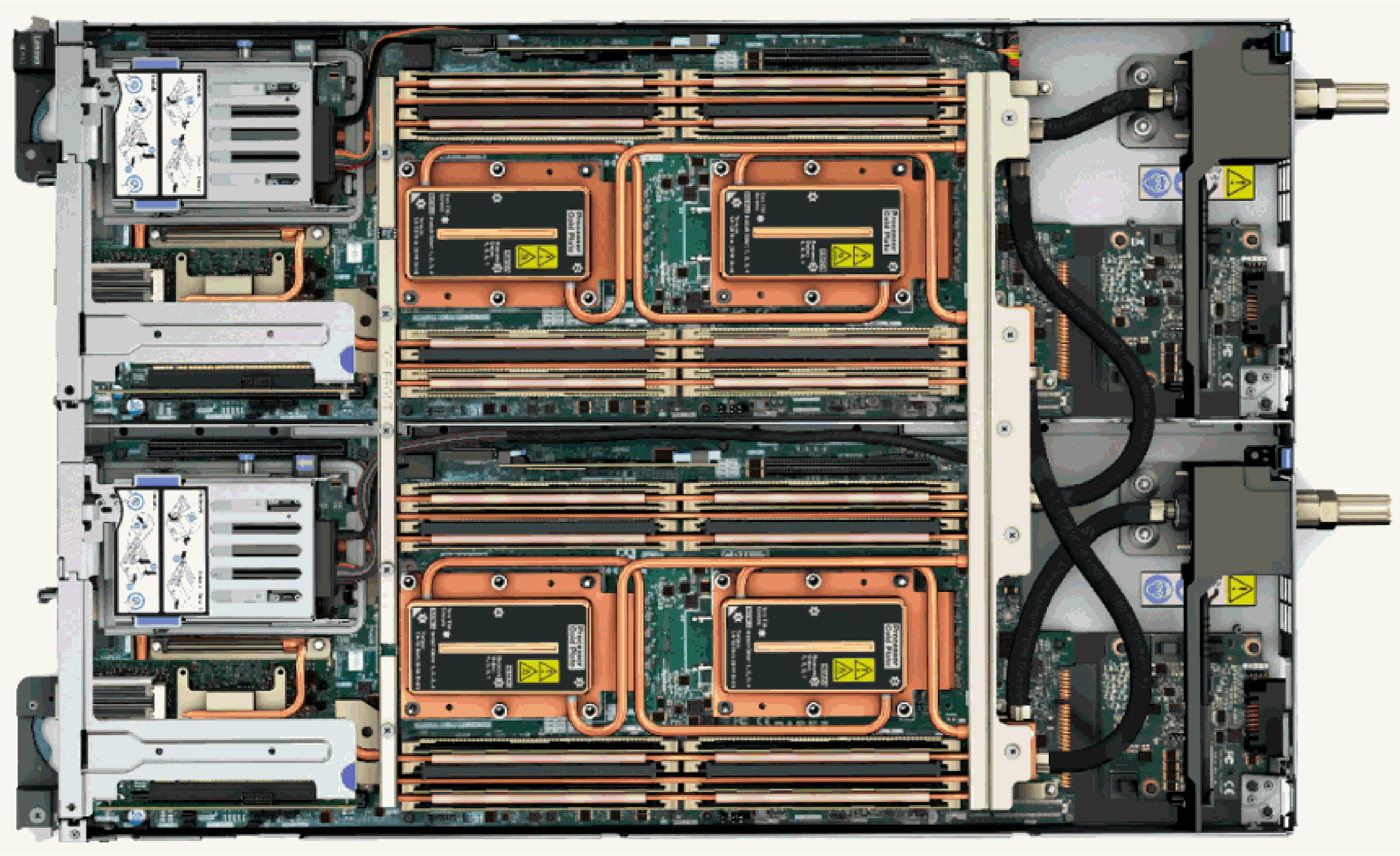


Supermuc-NG (Lenovo SD650 nodes, direct water cooling)



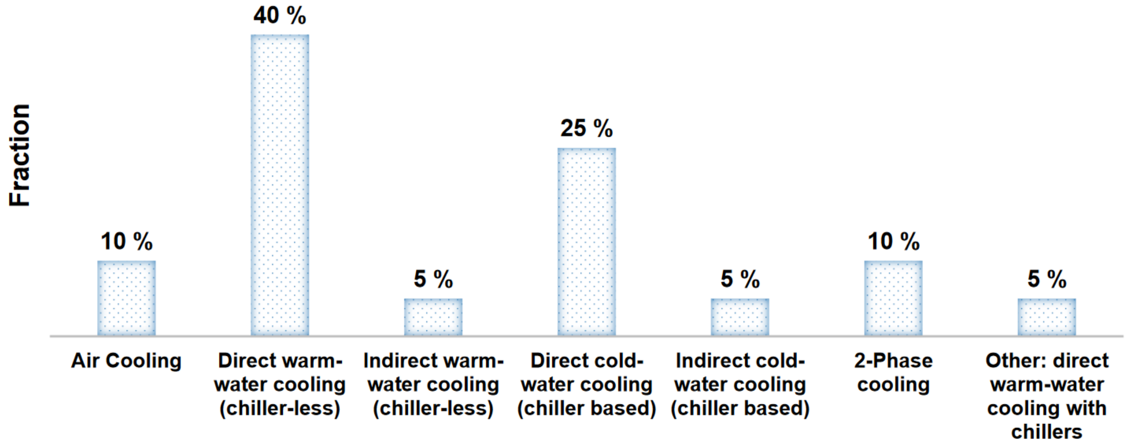


Lenovo SD650 direct water cooling



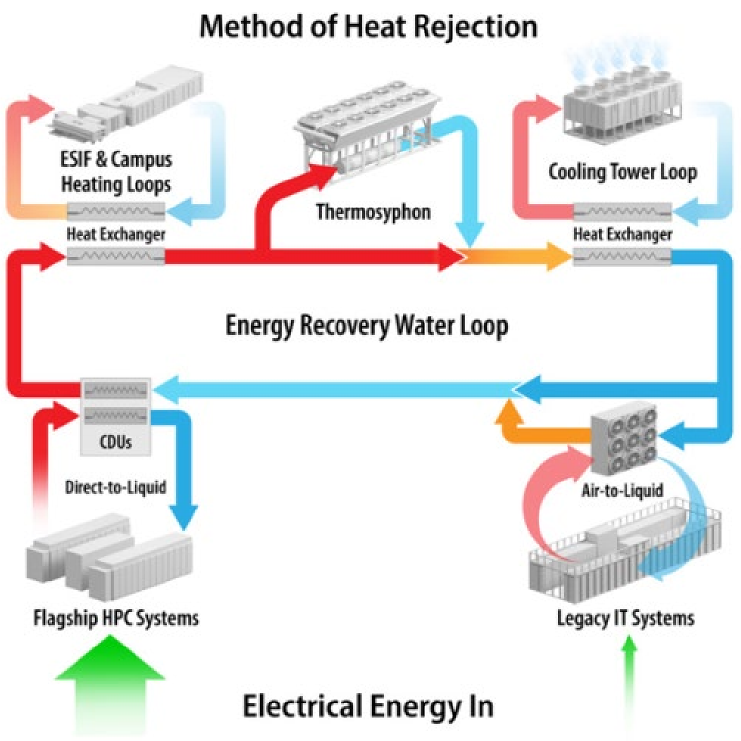
Cooling technologies

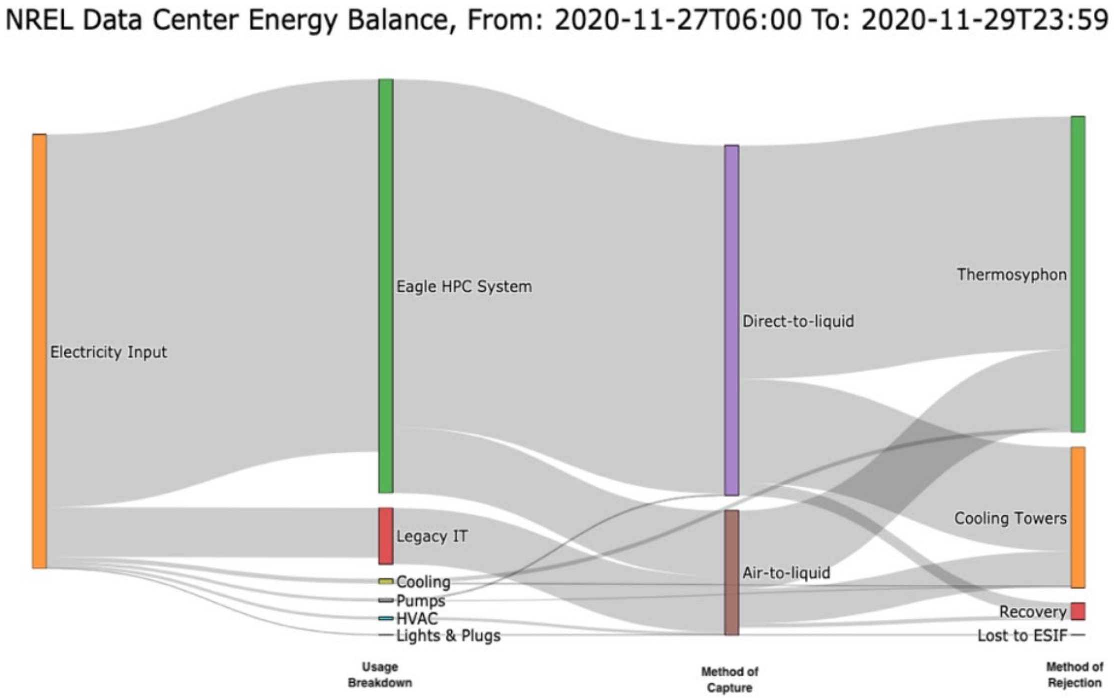
- Air cooling
  - easy to build and maintain, inefficient
- Direct water cooling
  - warm: difficult to build and maintain, very efficient, only for cooler climates (“free air cooling”)
  - cold: difficult to build and maintain, semi-efficient, for warmer climates
- Indirect cooling
  - cool hardware with air, cool air with water



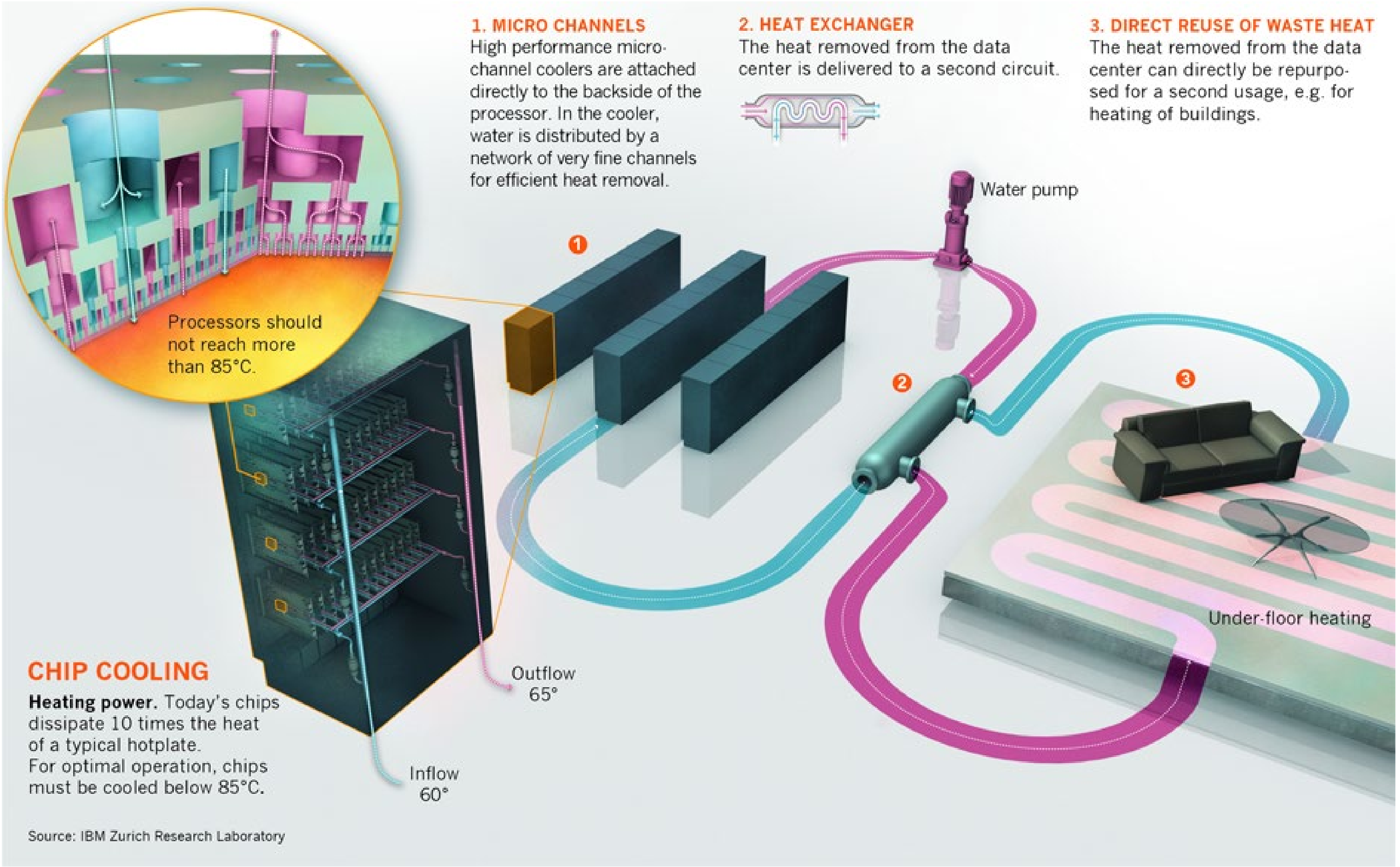
2020 survey among tier-0 and tier-1 HPC sites in Europe

ESIF data center, NREL (PUE of 1.06)

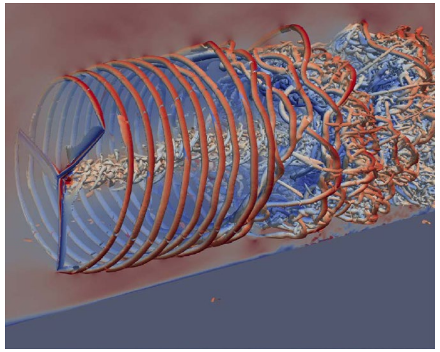


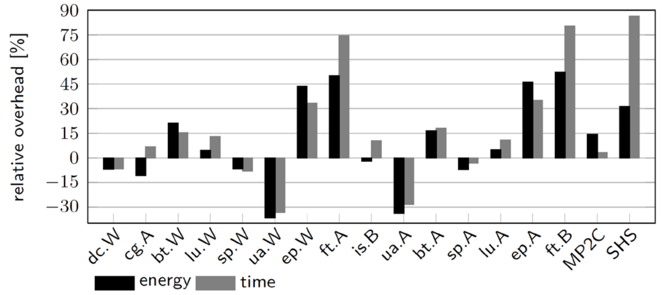


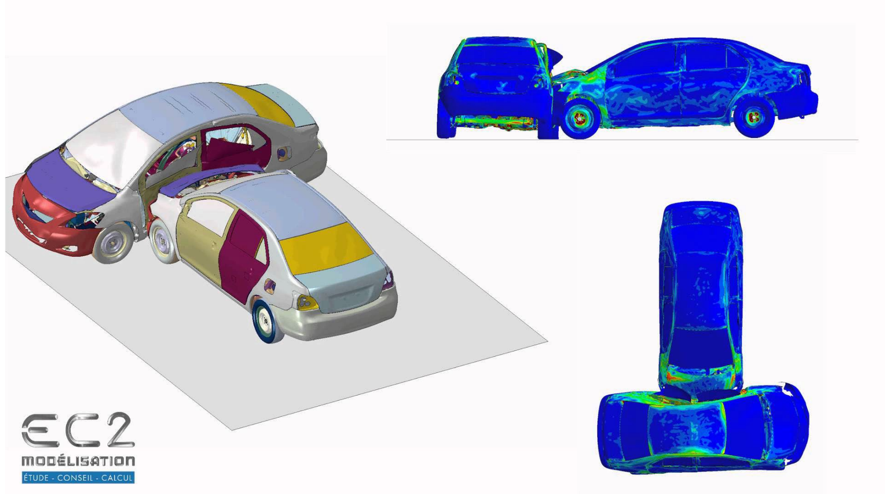
How can we recycle any remaining energy consumption?

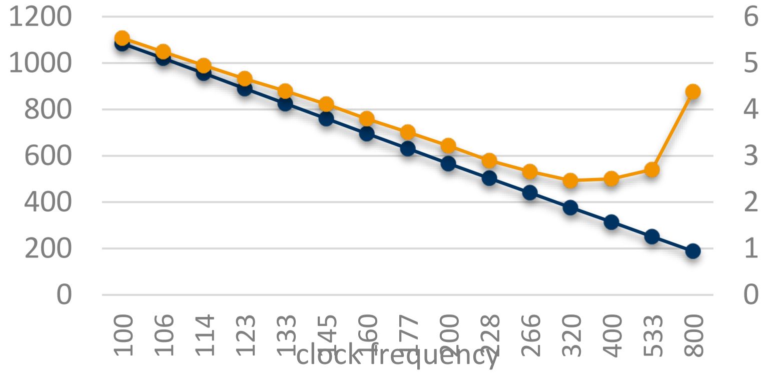


HPC for Sustainability









Open issues

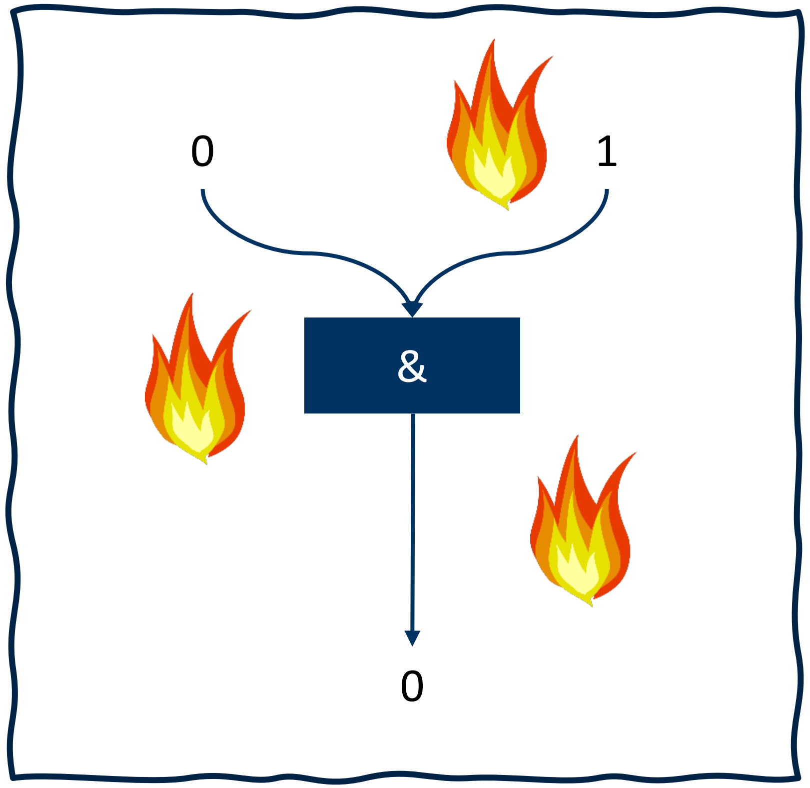
- There is more than just energy and power
  - Carbon Usage Effectiveness (CUE)
  - Water Usage Effectiveness (WUE)
  - Space Usage Effectiveness (SpUE)
- There are too many metrics and many are inaccurate
  - Power Usage Effectiveness (PUE)
    - Partial PUE (pPUE)
  - Energy Reuse Effectiveness (ERE)
  - Energy Reuse Factor (ERF)
- The metrics are often flawed
  - e.g. PUE cannot be used to compare HPC sites in different climate zones
- There are diverging interests
  - Operator: minimize power/energy, maximize workload throughput
  - User: minimize wall time
  - Taxpayer/politicians: minimize costs

Future developments and ideas

- High-bandwidth memory (HBM)
  - Memory and computational units physically as close together as possible, minimize data transport distance
- Fabrication size reduction
  - Research in new designs and materials (away from silicon) to decrease below ~2 nm threshold
- Near-threshold voltage computing
  - operate CPUs below power safety limits, accept computational errors and mitigate in software (e.g. iterative solvers)
- Special purpose hardware
  - Accelerators (scientific computing, AI, etc.)
  - FPGAs
  - Custom hardware designs for domain-specific problems
- Optical computing
  - Use photons instead of electrons
  - Various approaches in research, not clear yet if viable alternative

Reversible computing and Landauer principle: the future?

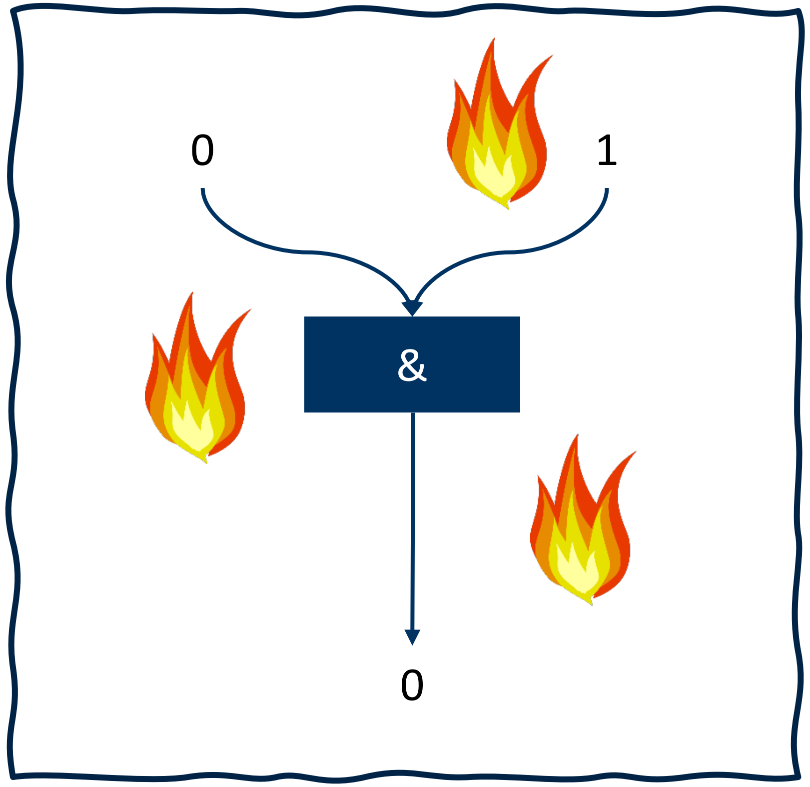
- There's a lower theoretical limit ("Landauer limit") to energy consumption of computation
  - Irreversible computation (e.g. logical AND) erases information, hence must be accompanied by corresponding entropy increase (=heat) in a closed system
    - because thermodynamics ¯_(ツ)_/¯
  - Landauer limit is approx. 0.0175 eV or 2.805 * 10^{-21} J at room temperature
  - We're currently still several orders of magnitude away from that…



Reversible computing and Landauer principle: the future?

- Koomey's Law: The number of computations per joule doubles every 1.57 years
  - Coupled with Landauer limit: no more energy efficiency increase after 2080…
  - Also applies to quantum computing

- Solution: reversible computing
  - In theory, computing without losing information doesn't need to increase entropy, hence no heat



Today's takeaway

- There's a lot of research and engineering going on
  - in sustainability for HPC
  - in sustainability with HPC
- Power/heat are the main limiting factors in HPC
  - almost everything uses water cooling these days
  - waste heat is recycled as much as possible and "freely cooled" afterwards (no active chillers)
  - short-term developments quite clear, long-term future very unclear

- How to reach me/us
  - philipp.gschwandtner@uibk.ac.at
  - https://dps.uibk.ac.at/~philipp
  - https://uibk.ac.at/fz-hpc

universität innsbruck

FZ HPC

Image sources

- Green HPC: https://www.hpcwire.com/2021/07/15/15-years-later-the-green500-continues-its-push-for-energy-efficiency-as-a-first-order-concern-in-hpc/, https://www.chemistryworld.com/features/oil-spill-cleanup/3008990.article, Marcel Ritter (UIBK), https://twitter.com/maven2mars/status/984440044659159040, https://www.nasa.gov/ames/image-feature/nasa-highlights-simulations-at-supercomputing-conference-like-aircraft-landing-gear
- TOP500 Trend: https://www.top500.org/statistics/perfdevel/
- Lenovo SD650 Water Cooling Images and Animation: https://lenovopress.lenovo.com/lp0636-thinksystem-sd650-direct-water-cooled-server-xeon-sp-gen-1
- Cooling Technology Survey: https://events.prace-ri.eu/event/1186/attachments/1587/2924/Shoukourian.pdf
- ESIF Data Center: https://www.nrel.gov/docs/fy21osti/79712.pdf
- IBM Research Energy Reuse: https://www.zurich.ibm.com/st/energy_efficiency/zeroemission.html
- Wind turbine: https://www.nrel.gov/docs/fy22osti/81212.pdf