703308 VO High-Performance Computing
Energy in HPC

Philipp Gschwandtner

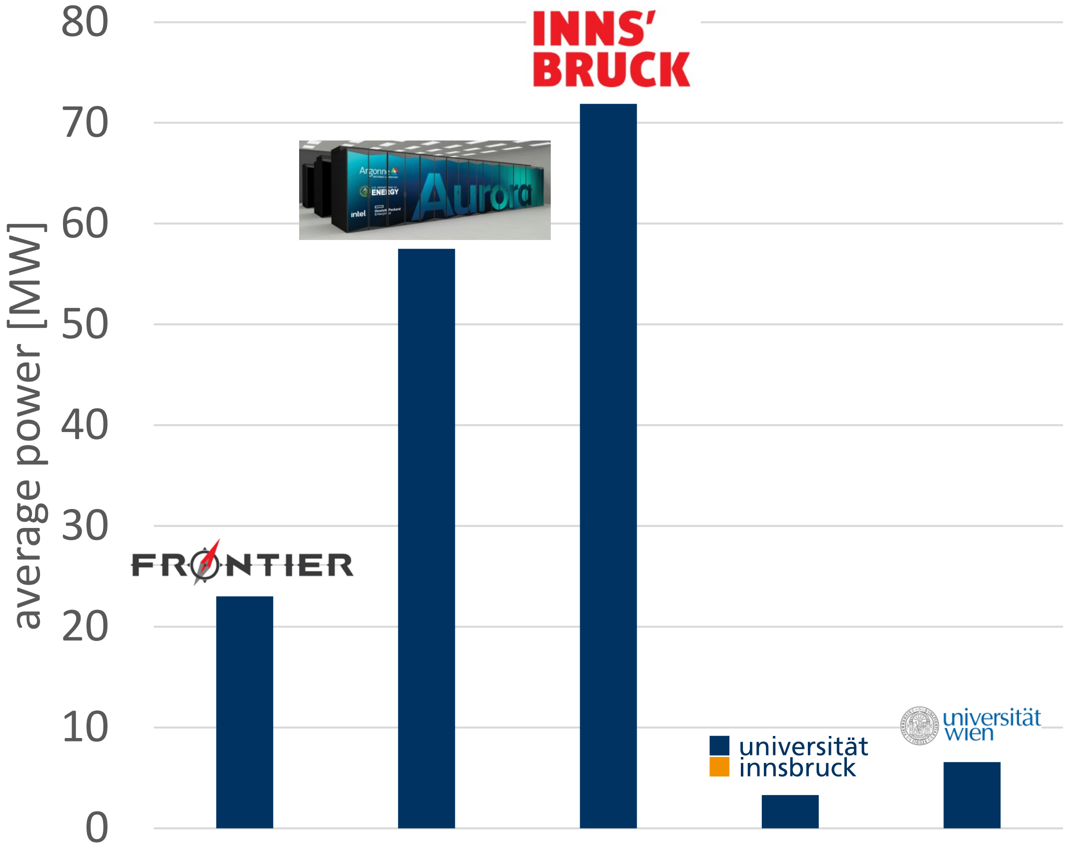
Overview

- motivation / indications
  - supercomputers
  - Green500 & accelerators
- basics
  - power & energy
  - instrumentation, measurement and modeling
  - control mechanisms
- consequences & applications
  - multi-objective optimization, workload co-scheduling, MPI slack optimization, etc.

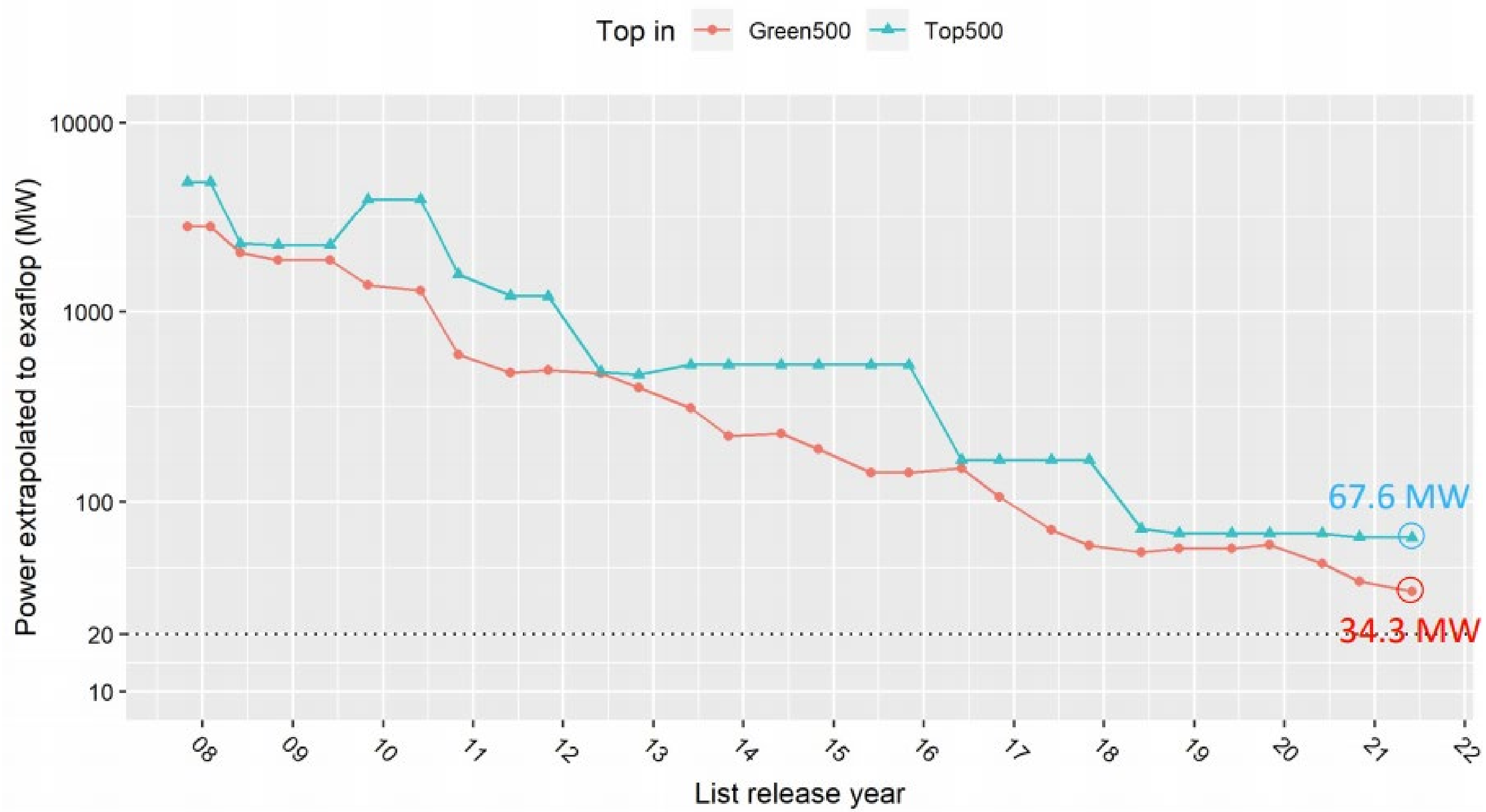
Why is energy consumption relevant in HPC?

- Current No. 1 in Top500: Frontier
  - 23 MW of average power
  - But how much is that?

- 40% Aurora supercomputer (full system!)
- 32% of Innsbruck
- 7x University of Innsbruck
- 3,5x University of Vienna

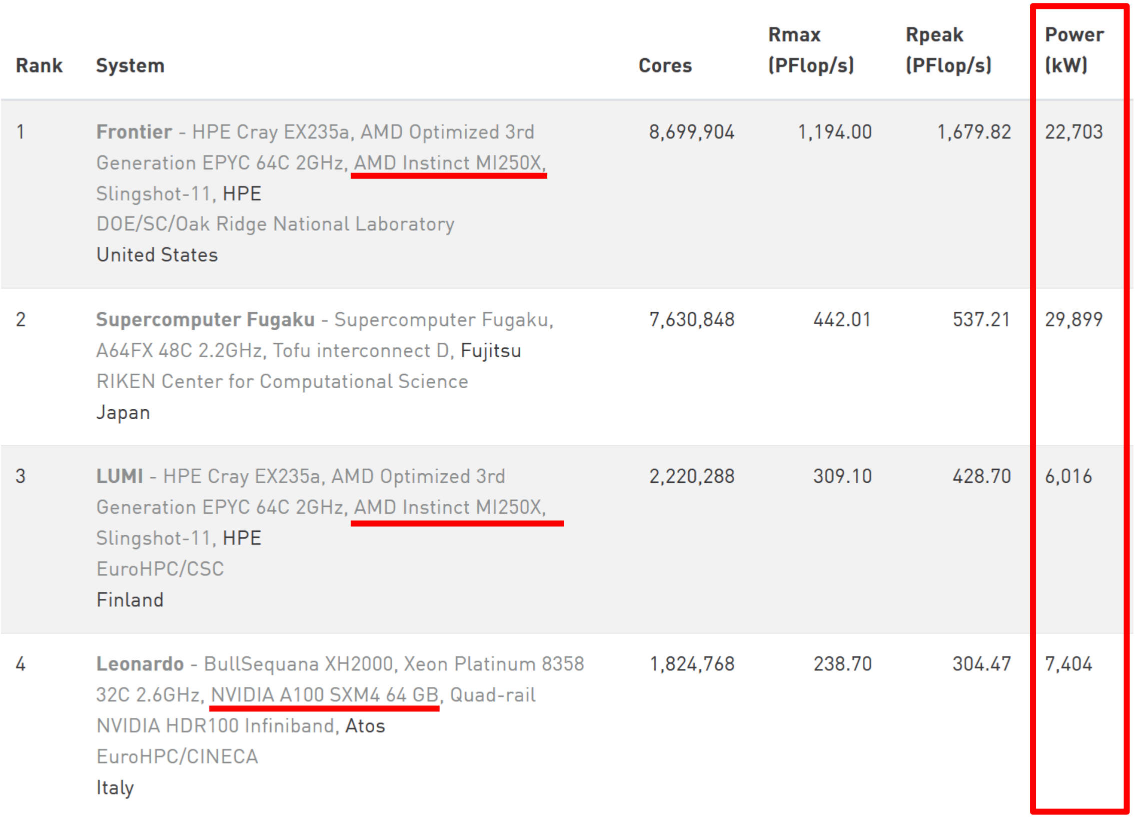


Power consumption projected to 1 exaflop

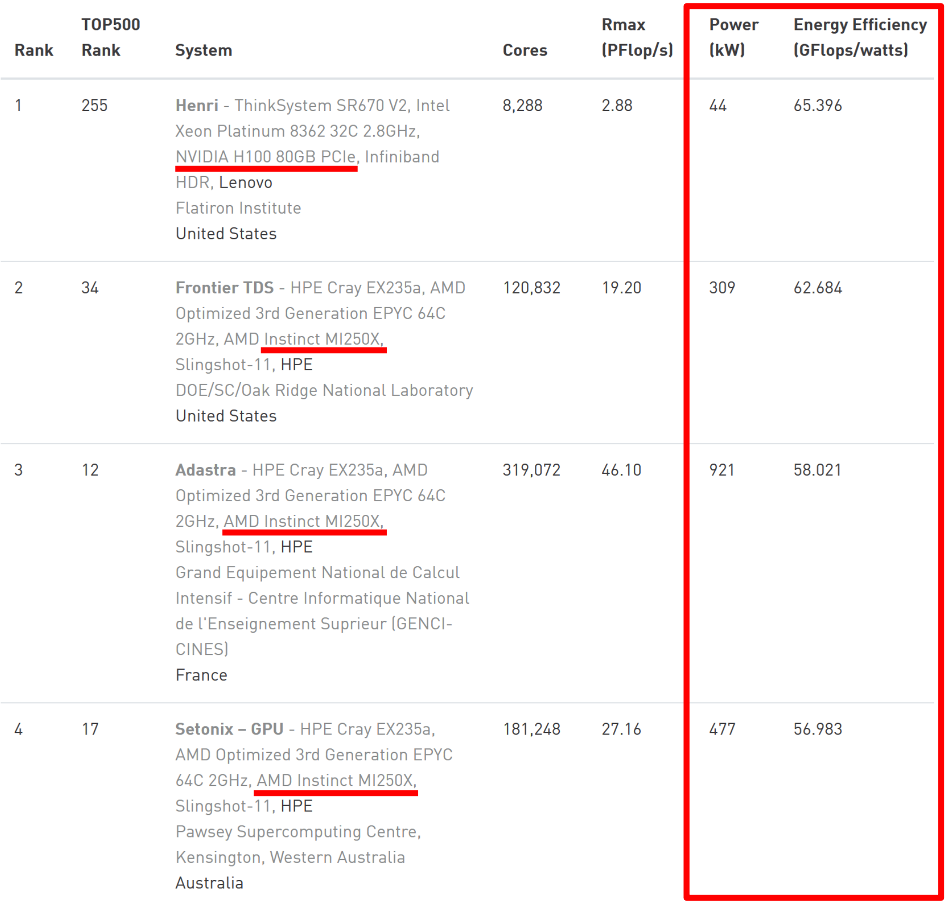


https://www.hpcwire.com/2021/07/15/15-years-later-the-green500-continues-its-push-for-energy-efficiency-as-a-first-order-concern-in-hpc/

TOP500 & Green500



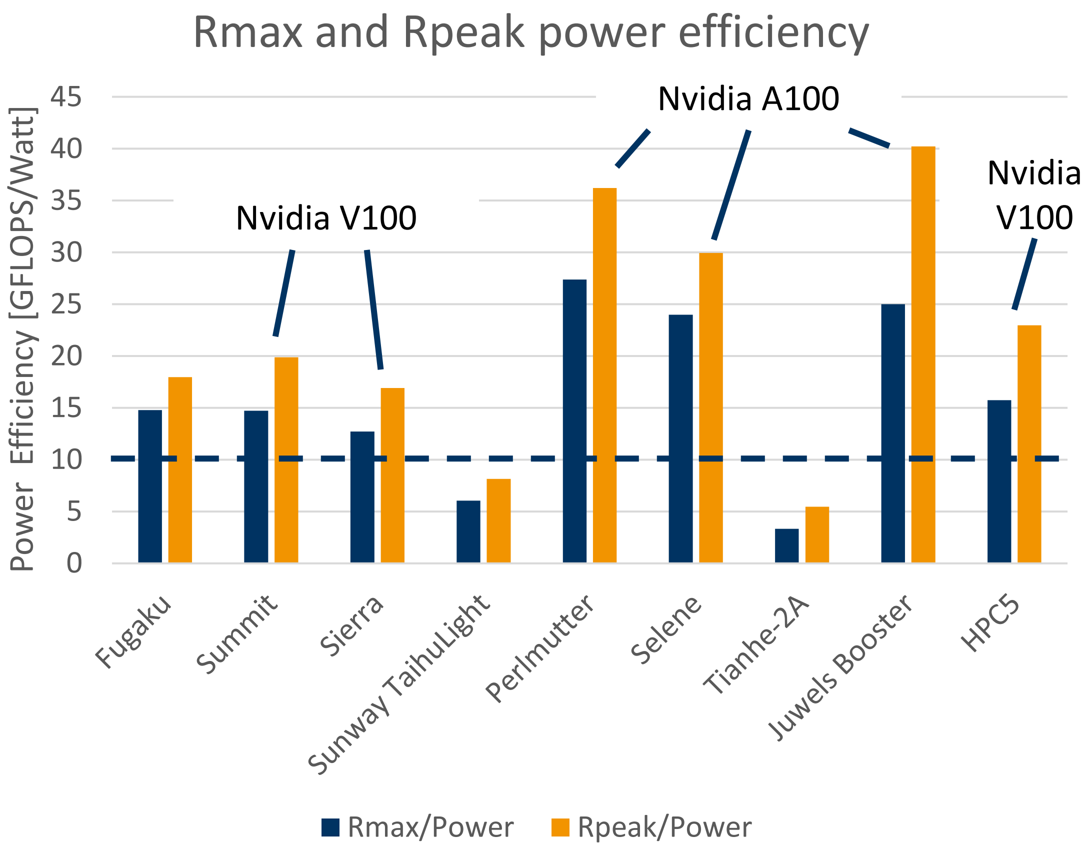
Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,699,904	1,194.00	1,679.82	22,703
2	**Supercomputer Fugaku** - Supercomputer 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



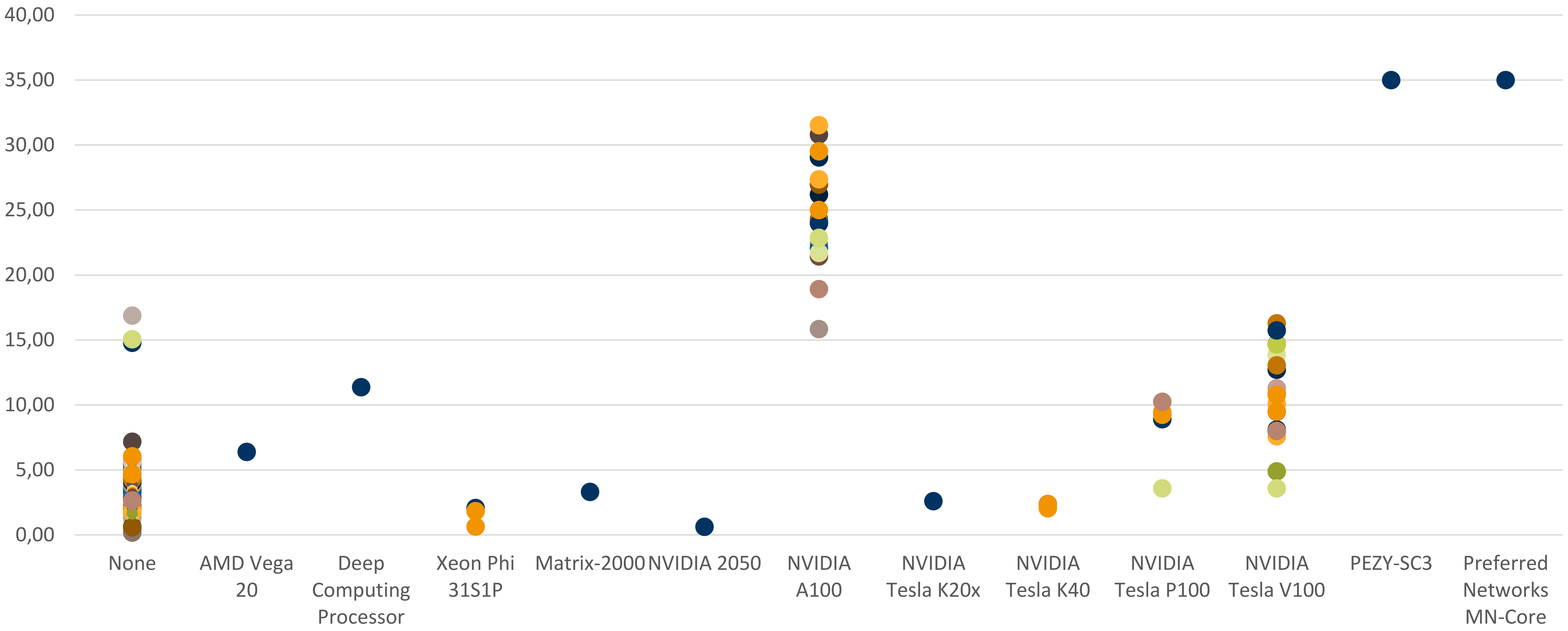
Rank	TOP500 Rank	System	Cores	Rmax (PFlop/s)	Power (kW)	Energy Efficiency (GFlops/watts)
1	255	**Henri** - ThinkSystem SR670 V2, Intel Xeon Platinum 8362 32C 2.8GHz, NVIDIA H100 80GB PCIe, Infiniband HDR, Lenovo Flatiron Institute United States	8,288	2.88	44	65.396
2	34	**Frontier TDS** - HPE Cray EX235a, 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

Why are we using accelerators?

- All top 10 systems above 10 GFLOPS/Watt use accelerators
  - Nov ‘21 data
- Exceptions:
  - Fugaku: ARM-based, no accelerators
  - Tianhe-2A: Matrix 2000 accelerators (128 core RISC CPUs)



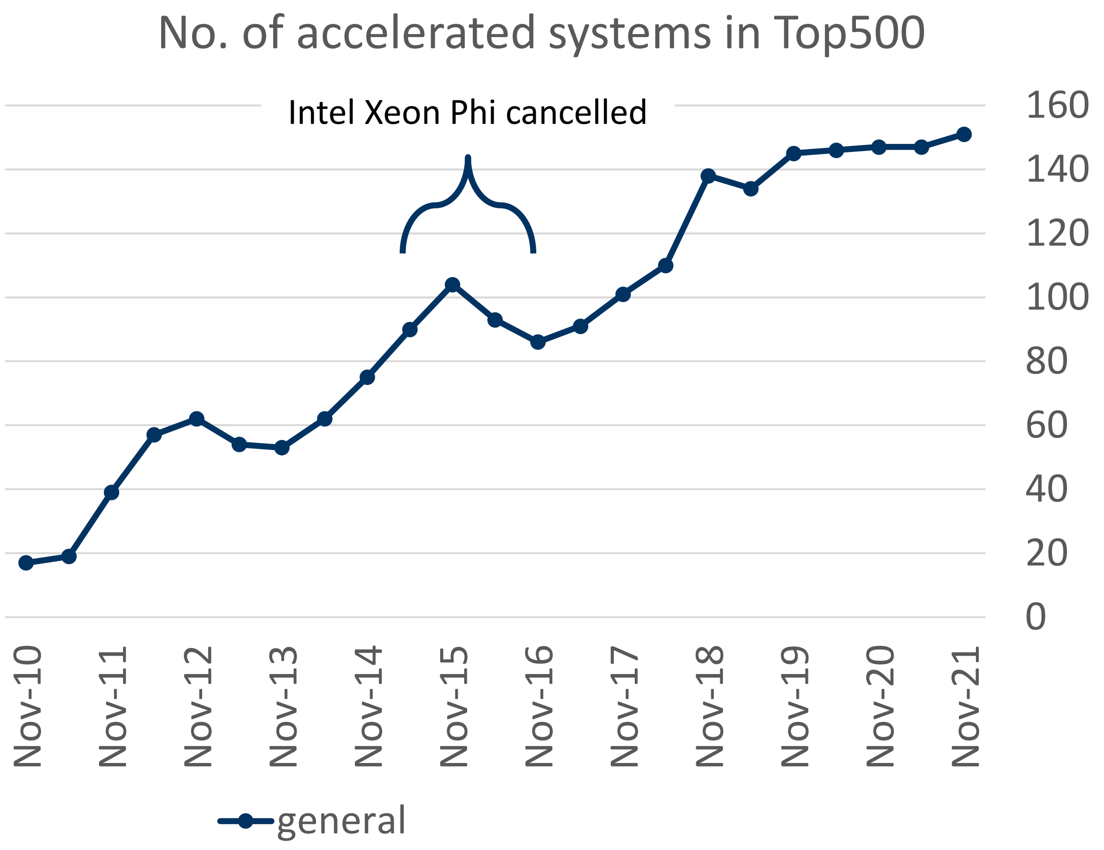
Power Efficiency of all Top 500 Systems



Market share

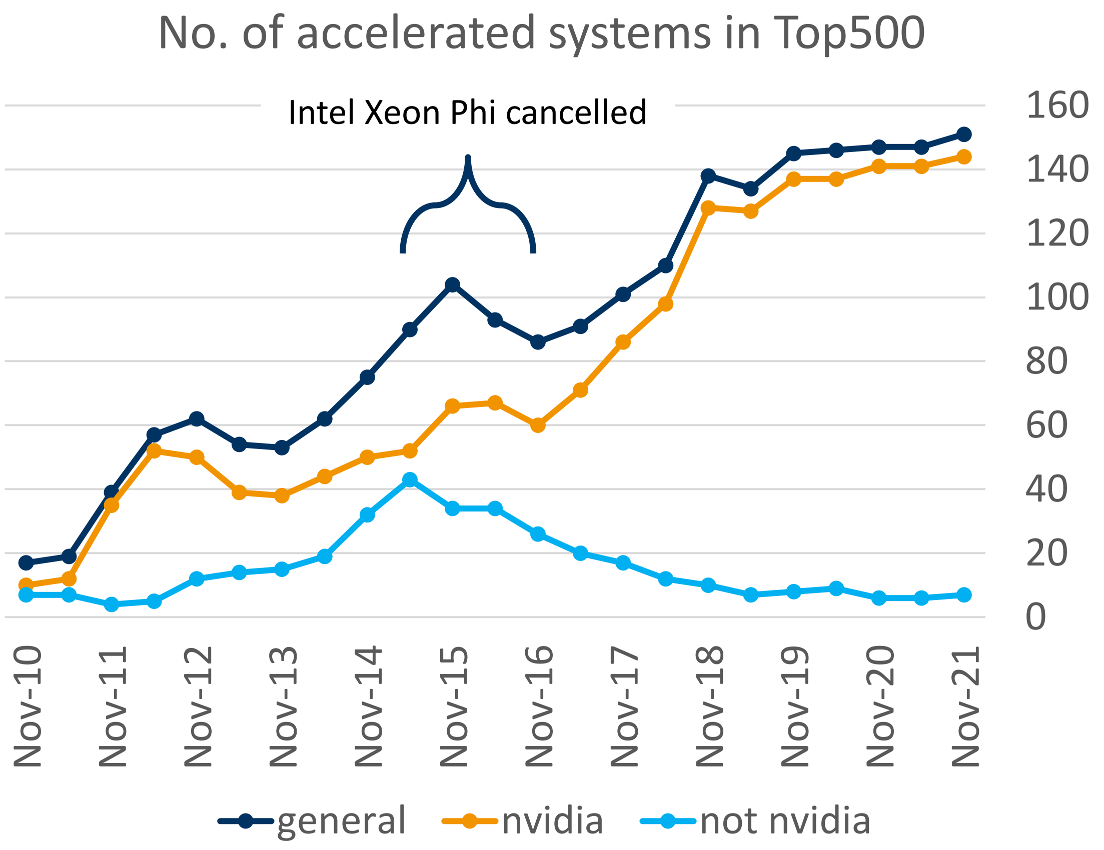
- Accelerator market share in HPC has been steadily increasing and will likely continue to do so
  - 10 out of top 10 on Green500 list (Nov. 2022)

- Application developers **need** to use accelerators to get high performance on modern systems



Market share cont'd

- Problem: No market competition
  - Nvidia is predominant
  - originally also Cell processor (Playstation 3!) and Intel Xeon Phi
  - disappeared since ~2015
  - right now (2023): AMD comeback, new Intel attempts
- Nvidia encourages using CUDA
  - vendor lock-in
- There are alternatives!
  - SYCL, ROCm/HIP, OpenMP, etc.



Basics

Disclaimer

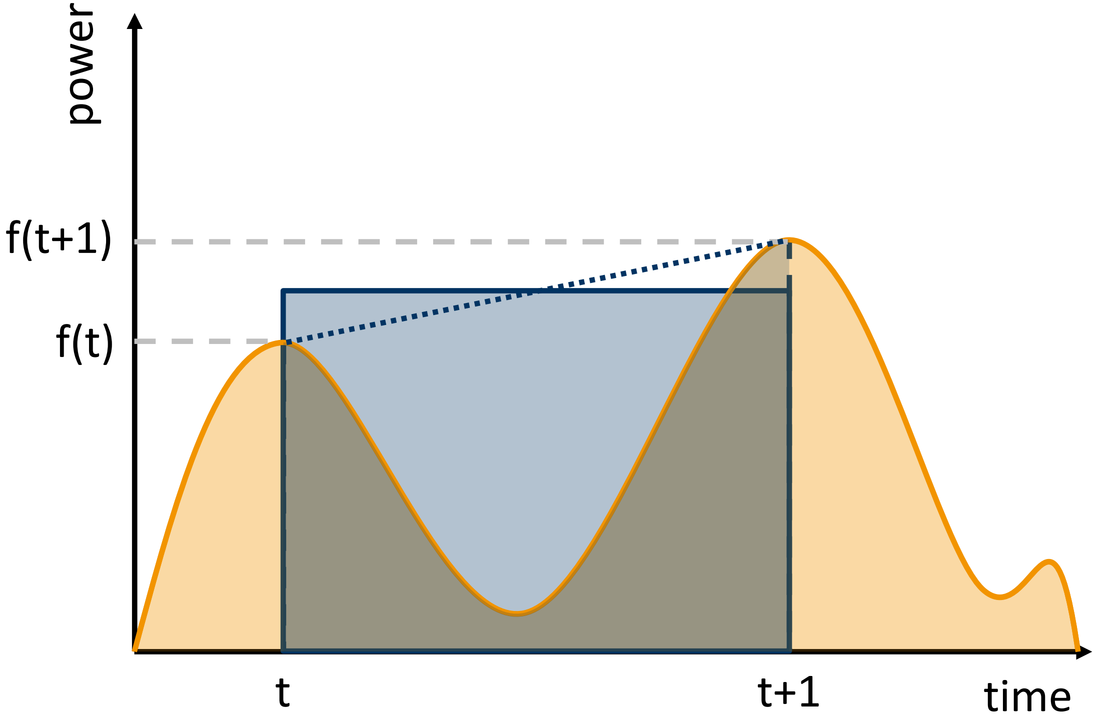
- Yes, according to thermodynamics, there is no energy “consumption”
  - We are just efficiently converting electricity to heat
  - Computational result is a side-product

- We’re still going to call it “power consumption” and “energy consumption” for practical purpose

Power vs. energy

- Power is instantaneous, i.e. measured at a specific point in time
  - E.g. 10 W (watts)
  - does not have a time component

- energy is power over time ($E = \int P$)
  - E.g. 10 Wh (watt-hours) or 36 kJ – kiloJoules
  - Could be measured using analogue means
  - Often just power sampling ($E \cong P_{avg} \times T$)
    - Often requires very high temporal resolution for reasonable accuracy



Power consumption of integrated circuits

- $P_{total} = P_{short-circuit} + P_{static} + P_{dynamic}$
  - $P_{short-circuit}$: Power due to short-circuit current during transistor switching
  - P_{static}: Power due to leakage current, increases with decreasing feature sizes
  - $P_{dynamic} = C \times \boldsymbol{F} \times \boldsymbol{V^2} \times \boldsymbol{\alpha} \approx \boldsymbol{F^3}$
    (C… capacitance, F… frequency, V… voltage, α… switching factor)
  - Frequency and voltage are tightly connected
    - Hence: Sometimes referred to as “cube rule”

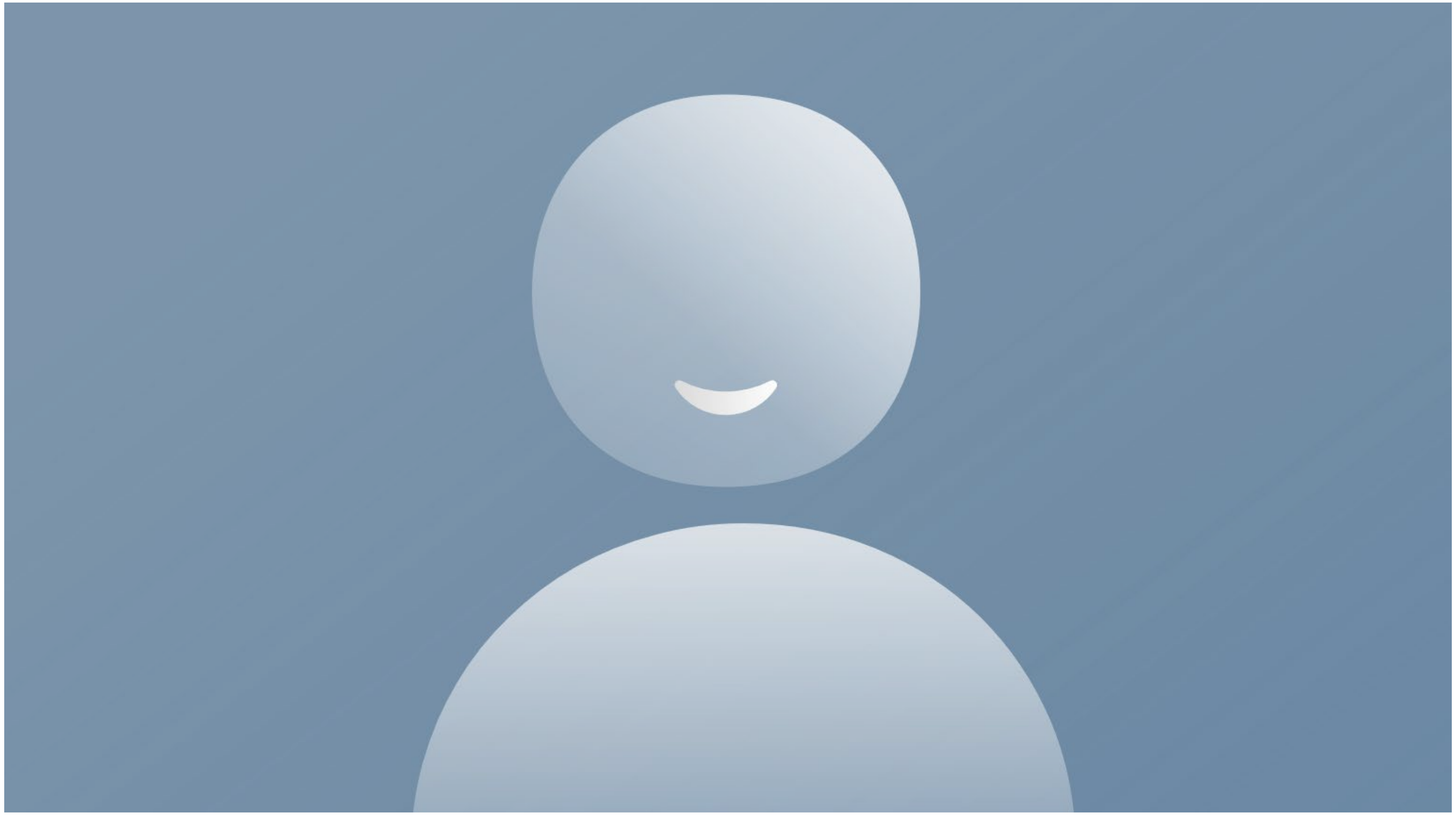
Power consumption of supercomputers

- Cores are not everything
  - Off-core entities: shared caches, memory controllers, system agents, etc.
  - Off-chip entities: RAM, mainboard, NIC, etc.

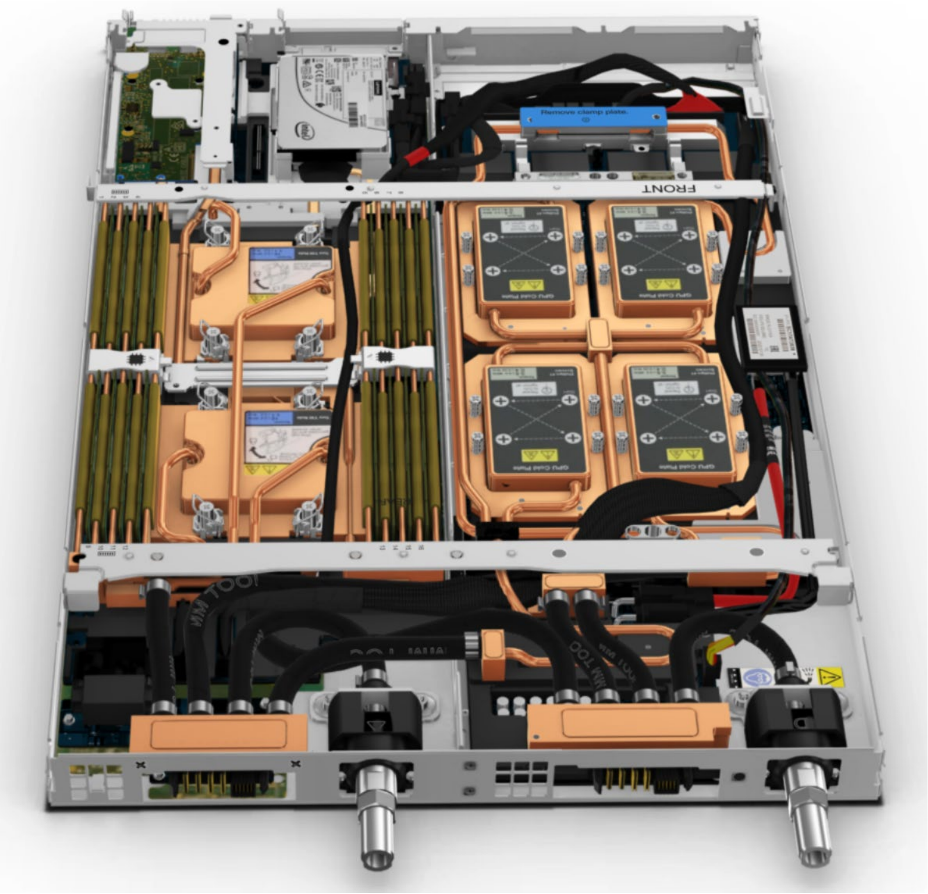
- Computing nodes are not everything
  - Network, storage, management, etc.
  - Cooling system
  - Lights, office equipment, etc.

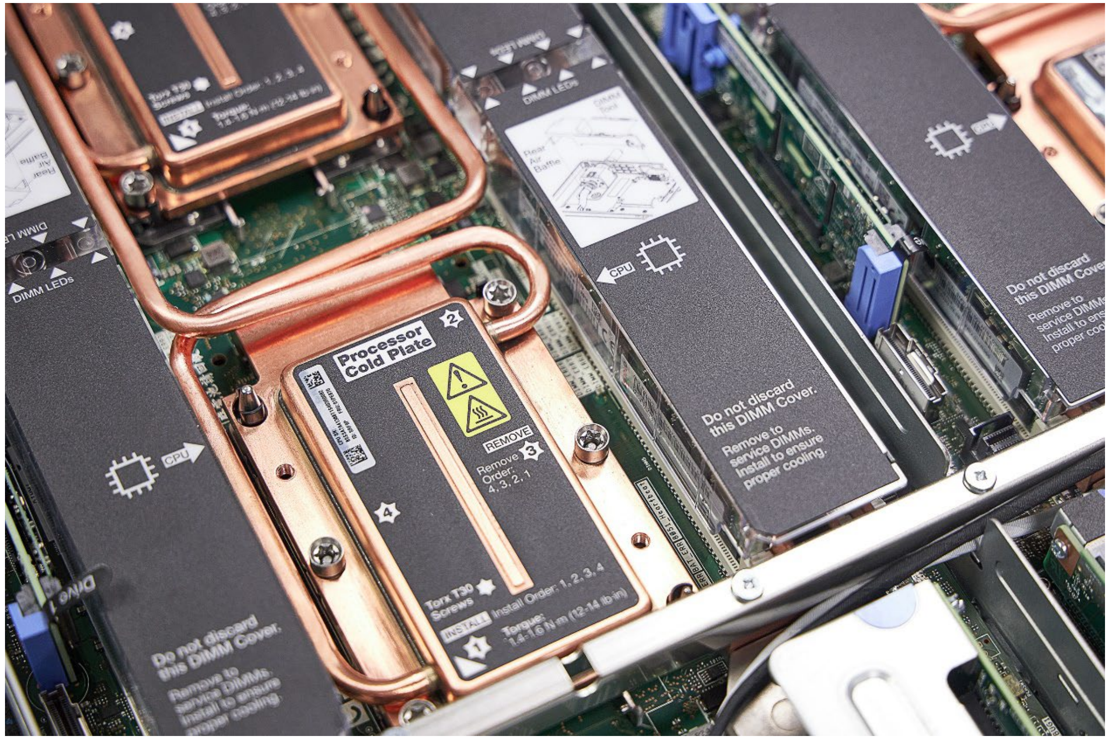
- Efficiency measured via e.g. PUE – Power Usage Effectiveness
  - Ratio of power required by supercomputer vs. power for its entire facility
  - E.g. SuperMUC-NG @ LRZ: PUE of 1.08
  - 10-20 years ago: PUEs as high as 2-3
    - Mostly caused by cooling overhead

Heat Dissipation



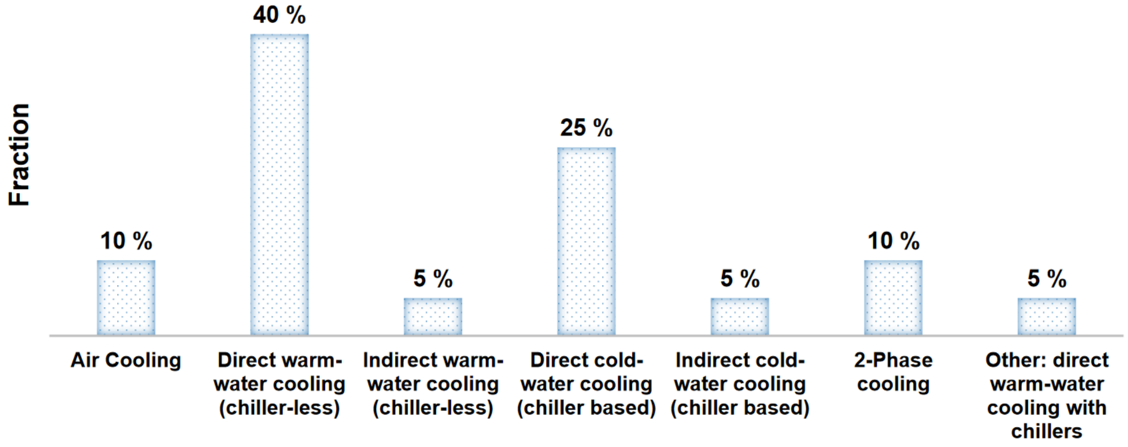
Supermuc-NG (Lenovo SD650 nodes, direct water cooling)




Processor Cold Plate
REMOVE
Remove Order: 4, 3, 2, 1
Torx T30 Screws
INSTALL
Install Order: 1, 2, 3, 4
CPU
DIMM LEDs
Do not discard this DIMM Cover.
Remove to service DIMMs. Install to ensure proper cooling.

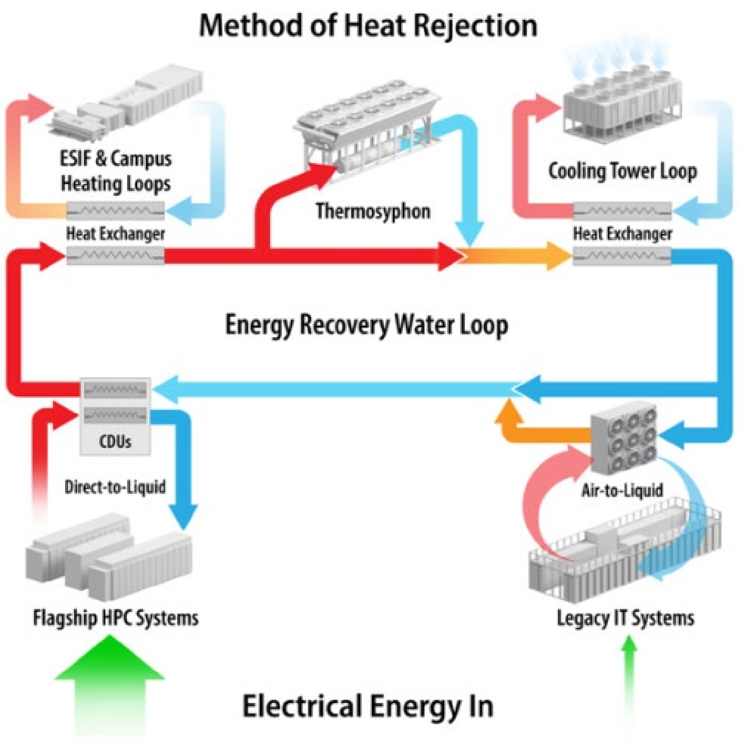
Cooling technologies

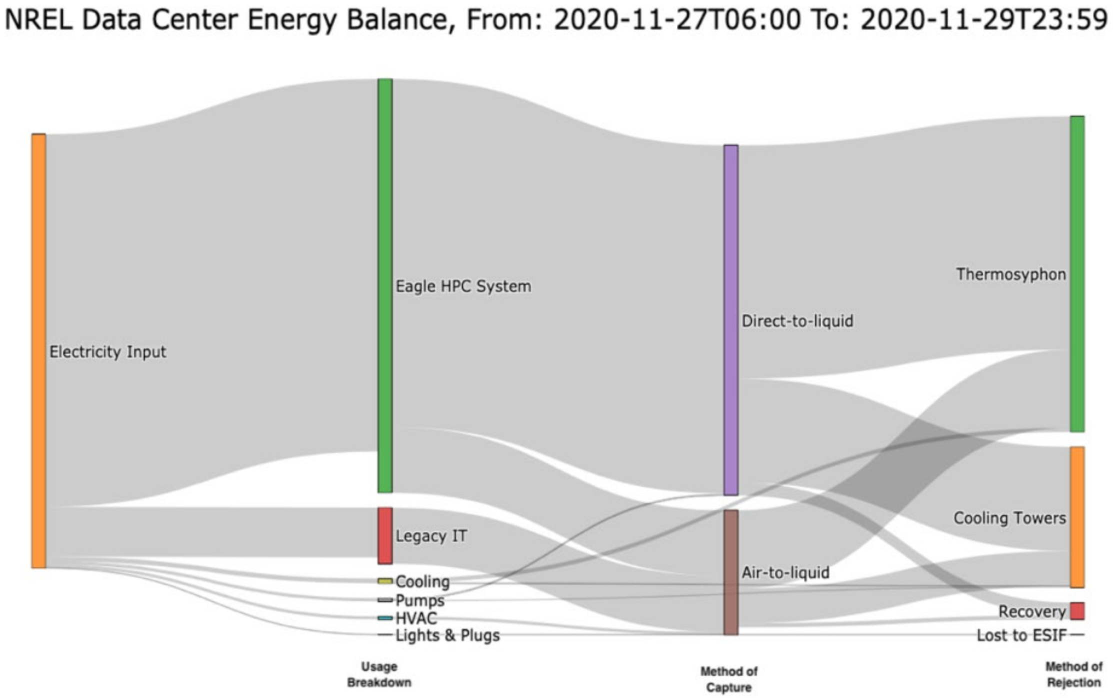
- Air cooling
  - easy to build and maintain, inefficient
- Direct water cooling
  - warm: “free air cooling”, water is not actively chilled, inlet temperature up to 50° C, difficult to build and maintain, very efficient, only for cooler climates
  - cold: water is actively chilled, difficult to build and maintain, semi-efficient, for warmer climates
- Indirect cooling
  - cool hardware with air, cool air with water
- Immersion cooling
  - VSC-3 (“the oil thing”)



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

ESIF data center, NREL (PUE of 1.06)

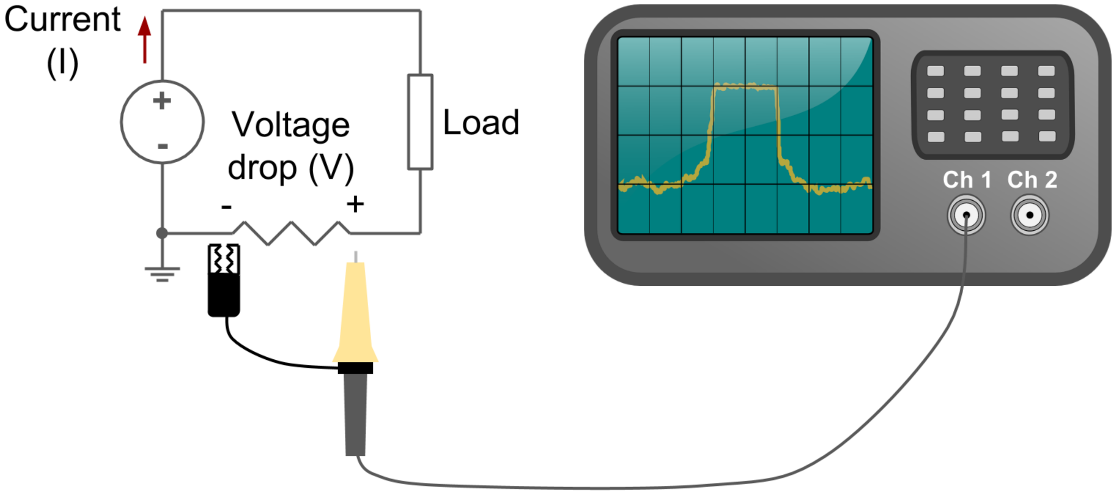




Instrumentation, measurement & modeling

Measurement methods

- Hardware instrumentation
  - in-bound: data available directly at compute resource (e.g. on- or off-core register)
  - out-of-bound: data available externally (e.g. via network interface)
    - Provided by vendor or self-made
    - Alternatively: wall socket measurements

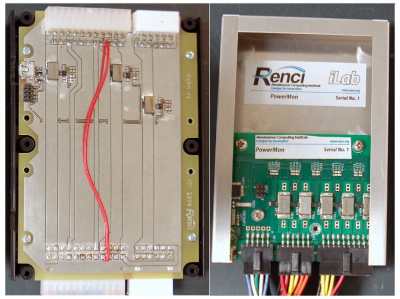


- Models and simulators
  - Wattch, SimpleScalar, Sim-PowerCMP, CACTI, McPAT, GPUWATTCH, etc.

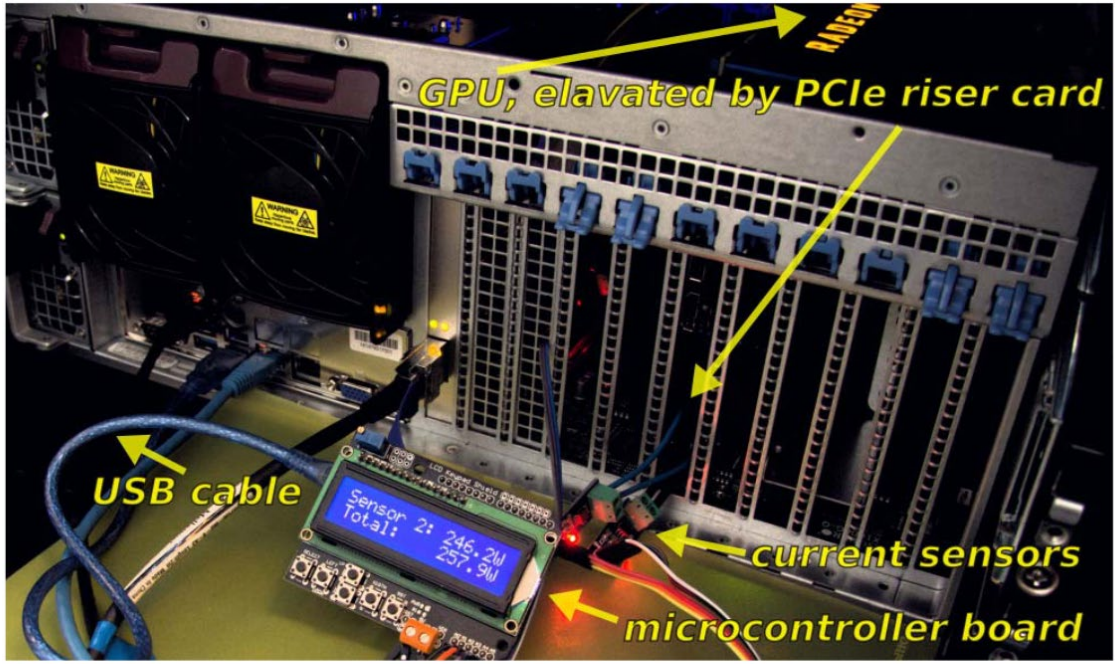
- Additional considerations: scalability & deployment of instrumentation hardware

Out-of-bound examples

Voltech PM1000+



PowerMon2



PowerSensor 2

Intel RAPL – Running Average Power Limit

- In-bound hardware method of monitoring and controlling power and energy
  - Available since Sandy Bridge (~2011)
  - Model or measurement method depends on CPU microarchitecture

- Provides low-overhead measurements via model-specific registers (MSRs)
  - Directly provide energy consumption data in “energy units”, model-dependent
    - E.g. Sandy Bridge: 15.3 µJ; Haswell: 61 µJ
  - Update once every 976 µs
  - Increase monotonically (similar to TSC register)
    - But only 32 bit! Overflows after some time, dependent on CPU stress

- Also supports controlling the hardware by power capping
  - E.g. “never exceed 43 watts”
  - Much more fine-grained than any OS/software mechanism

How to read Intel RAPL data

- Often requires root due to security implications
  - Reading MSRs requires raw register access to basically the entire CPU
  - Reading fine-grained energy/power data might enable side channel attacks

- Using the Linux kernel’s perf_event interface
  - Using `perf` and `sudo`
    - `sudo perf stat -a -e "power/energy-cores/" /bin/ls`
    - Alternative without root: requires a `/proc/sys/kernel/perf_event_paranoid` setting of less than 1

- Manually reading the MSRs (e.g. https://github.com/kentcz/rapl-tools)
  - `sudo modprobe msr && sudo chmod o+rw /dev/cpu/0/msr`
  - `sudo setcap cap_sys_rawio+ep <measurement_program>`

Additional tools

- Linux
  - Likwid: https://github.com/RRZE-HPC/likwid/wiki/Likwid-Powermeter
  - Powertop: https://github.com/fenrus75/powertop
  - Perf: `sudo perf stat -a -e "power/energy-cores/" /bin/ls`
  - PAPI: https://icl.utk.edu/papi/
  - `nvidia-smi`

- Windows?
  - Hwinfo, AIDA64, HWMonitor, PowerStrip, SpeedFan, …
  - Powercfg & windows energy estimation engine (E3), srumutil: https://devblogs.microsoft.com/sustainable-software/measuring-your-application-power-and-carbon-impact-part-1/?WT.mc_id=green-8660-cxa
  - Intel Power Gadget
  - WSL1/2?

- Mac
  - Intel Power Gadget

Other vendors

- A lot of “RAPL-compatible” systems from other vendors
  - Doesn’t really mean “compatible”, RAPL is used as a deonym here
  - Just means there is a user-accessible interface for getting power or energy data

- In-bound: AMD CPUs (RAPL/APM), Apple M1/M2 (?), etc.
- Out-of-bound: Nvidia (?), AMD GPUs (?), IBM (Amester), Sun (?), etc.

- Third-party: PowerMon, PowerPack, PowerInsight, etc.
  - Always out-of-bound

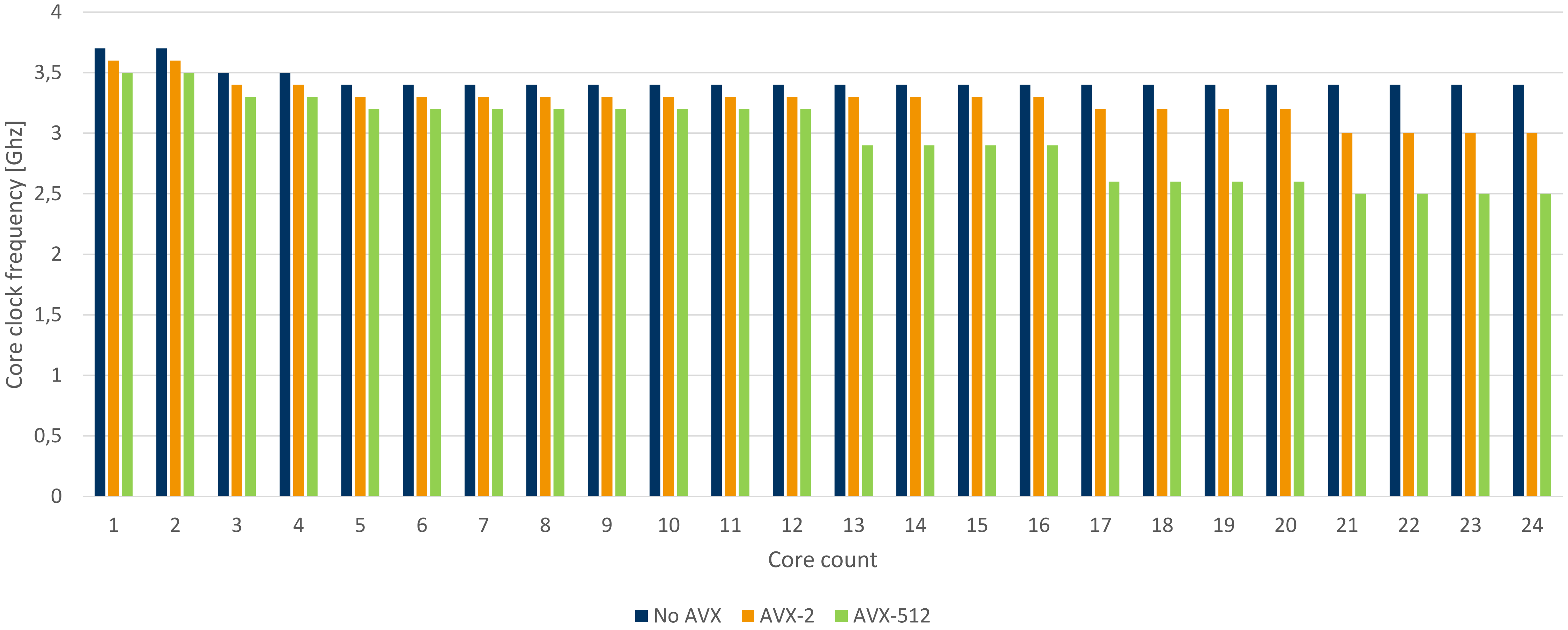
DVFS – Dynamic Voltage and Frequency Scaling

- Hardware can operate at multiple clock frequency points
  - Usually requires to scale voltage along with frequency
  - Originally one frequency for all components, these days individual frequencies for cores, last-level cache, etc.

- Originally controlled in software by selecting next DVFS state – 30ms latency
  - OS directly requests a so-called P-state by writing into a register
  - PCU halts CPU e.g. every few milliseconds, reads the requested & current P-state, acts accordingly

- Modern CPUs have much more autonomy in this
  - OS requests a certain range of frequencies, minimum QoS or maximum performance
  - CPU controls the actual setting in hardware (e.g. Intel SpeedShift, ~1 ms latency)
  - Required by AVX, Turbo and alike to work properly

AVX Turbo Frequencies (Intel Xeon 8174)



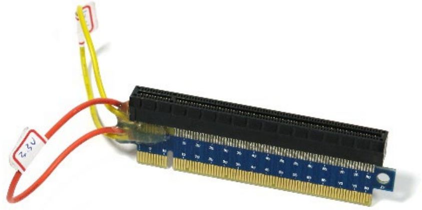
Green500 measurement methodology

- 33 pages of definitions: measurement devices, topology, workload requirements, averaging, etc.
  - https://www.top500.org/static/media/uploads/methodology-2.0rc1.pdf

- 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!)

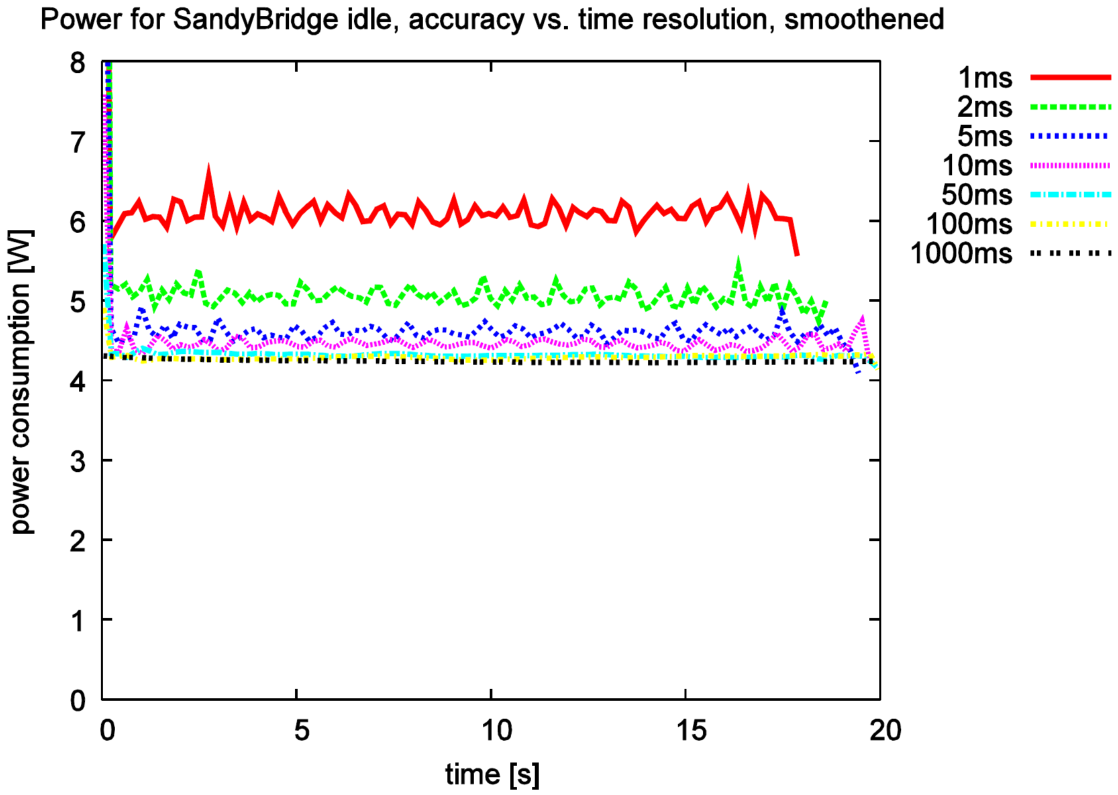
Issues

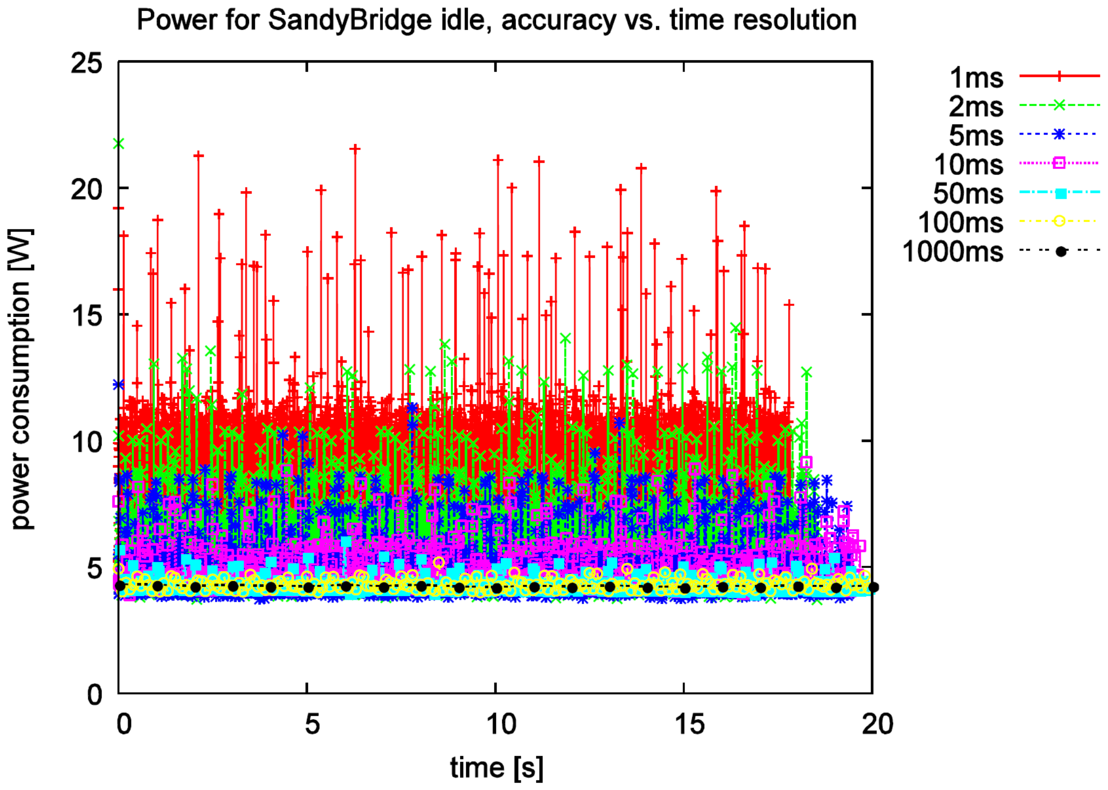
- Spatial resolution / topology
  - E.g. Intel RAPL offers separate readings for entire package, cores and off-core entities (e.g. RAM controller or iGPU)
    - What about RAM? Mainboard? GPUs? Storage? Network?
  - GPUs can draw up to 75 watts via PCIe, rest via ATX power connectors,
    - If no in-bound measurement available, you need PCIe riser cards



- Temporal resolution and accuracy
  - High-frequency CPU loads are hardly visible at the wall socket
  - Resolution != accuracy
    - E.g. RAPL can have 15.3 µJ resolution but a mJ accuracy or worse
- External conditions
  - Temperature
  - Measurement perturbation (often caused by high-frequency in-bound measurements)

In-bound measurement perturbation on a i7-2600k (RAPL)

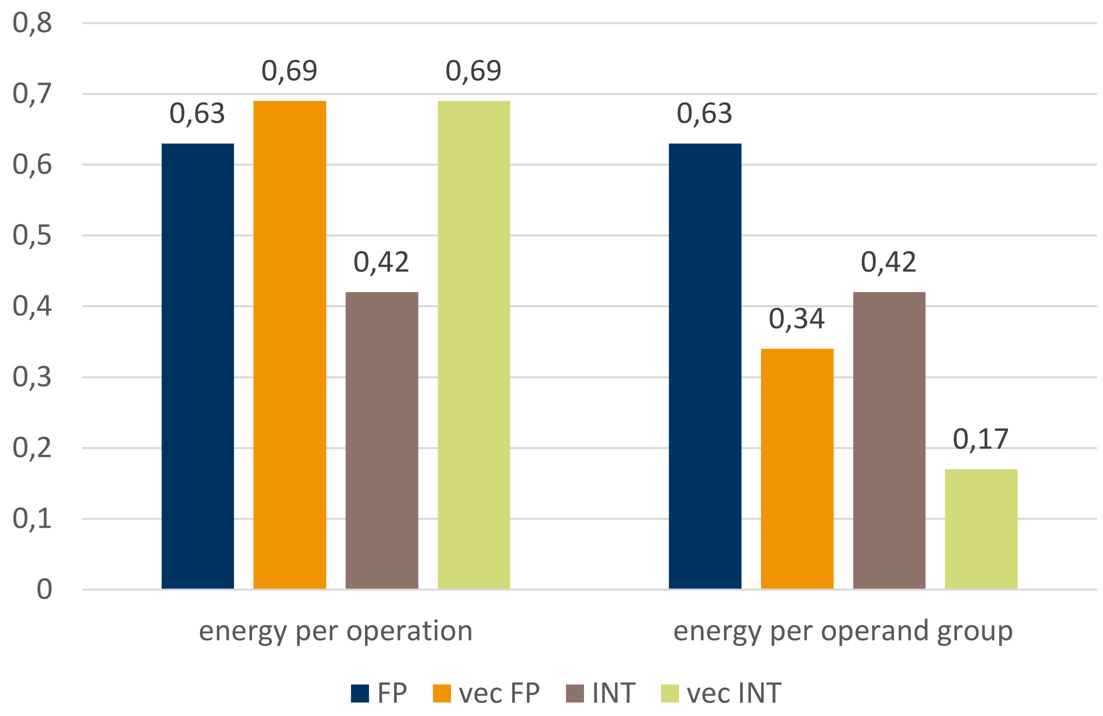




Modeling Example

- Lots of models available
  - Some as simple as linear combination of all performance counters
    - $E_{dynamic} = \sum \alpha_i \, c_i$
  - All the way up to Support Vector Machines (SVEs), Deep Learning, etc.

Energy consumption in nJ for IBM POWER7 FMA operations



Consequences and applications

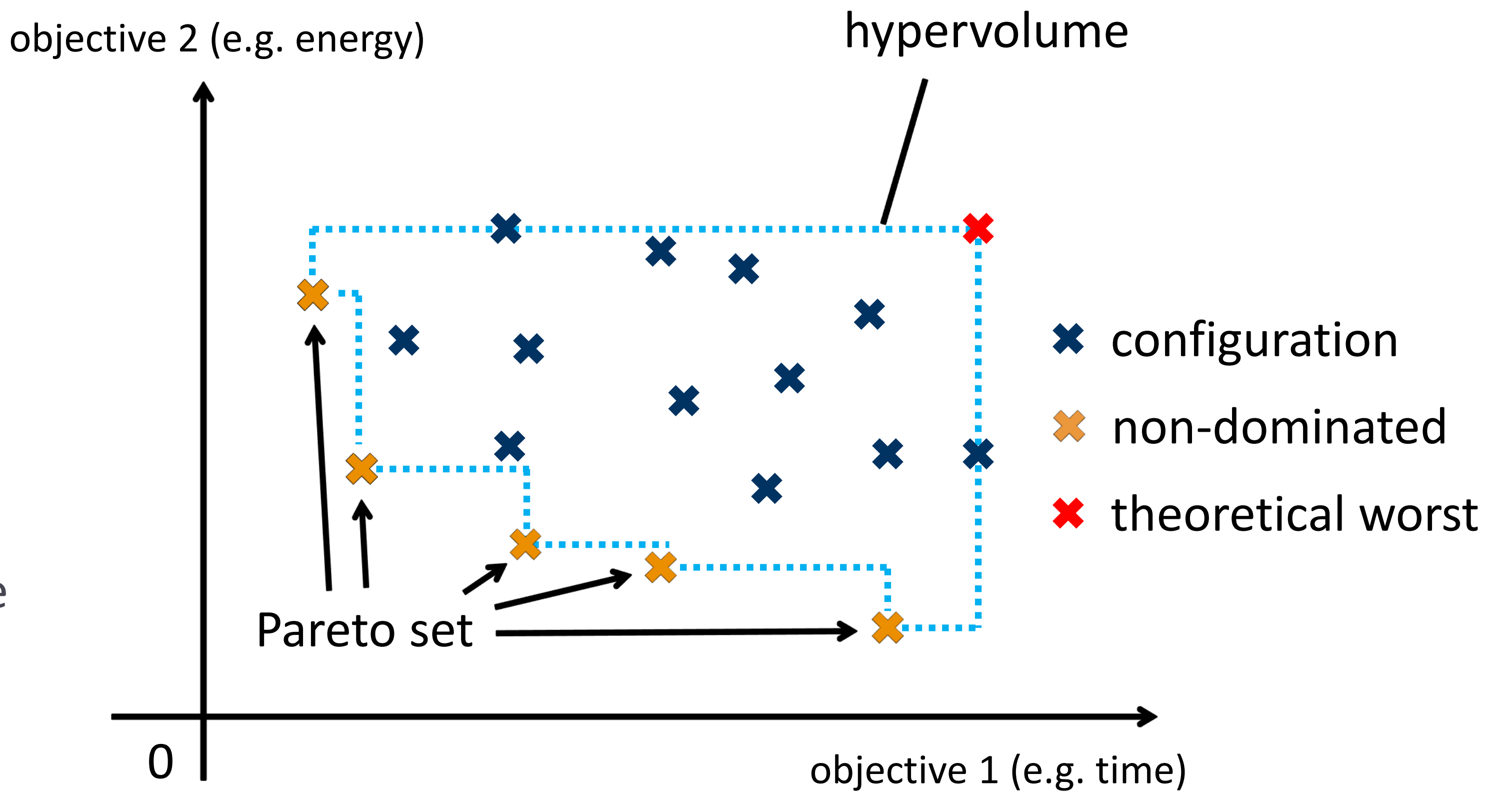
Multi-objective optimization

- We now have at least two objectives which are (partially) in conflict
  - Faster program execution (requires more power and energy)
  - Lower power or energy consumption (requires longer execution times)

- Two possible optimization approaches
  - Weight function to combine all objectives into a single one
    - E.g. EDP – Energy Delay Product, also: ED^2P, ED^3P, PDP, …
  - Alternatively use true multi-objective optimization, keeps flexibility
    - e.g. Pareto optimality

Multi-objective optimization and Pareto optimality

- Definition of Pareto set:
  - For each point in the set, no other point “dominates” it (=is better in all objectives)
  - Entails that it’s impossible to improve one objective without worsening another objective
- Superior to weighted approach
  - Provide all configurations in the Pareto set to the user, choose dynamically according to current preferences

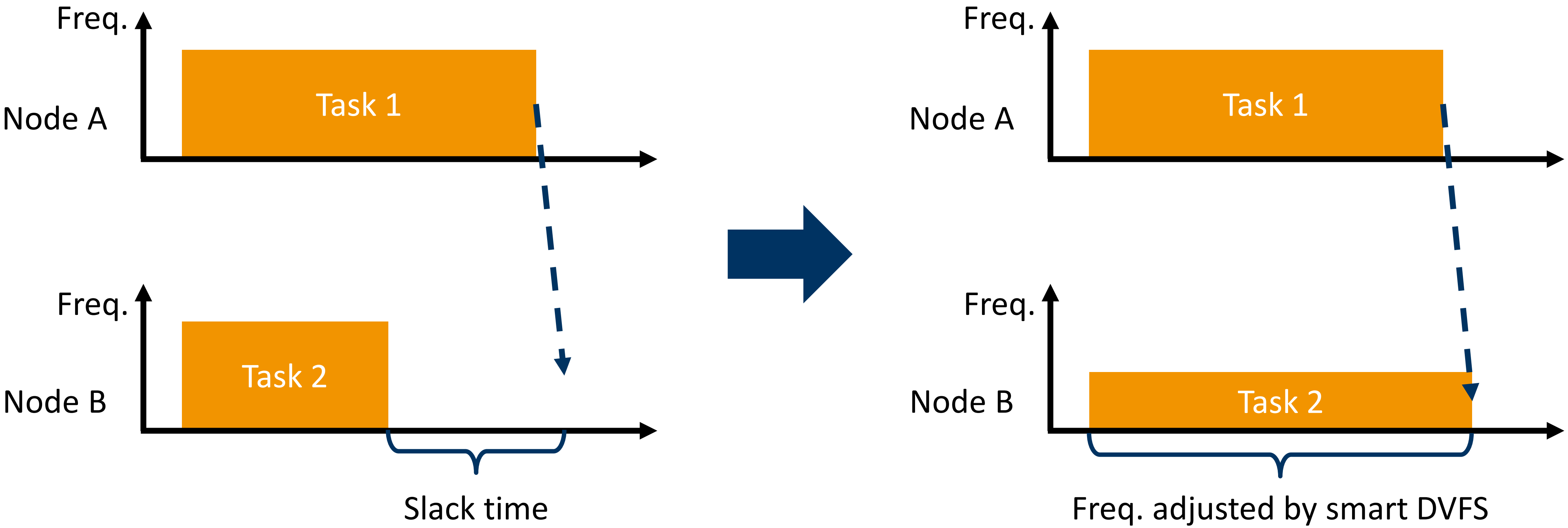


MPI slack time optimization

- Recognize slack time in parallel applications
  - Wait states
  - Periods of extended memcopy operations or I/O
  - Even computation if not on the critical path
  - etc.

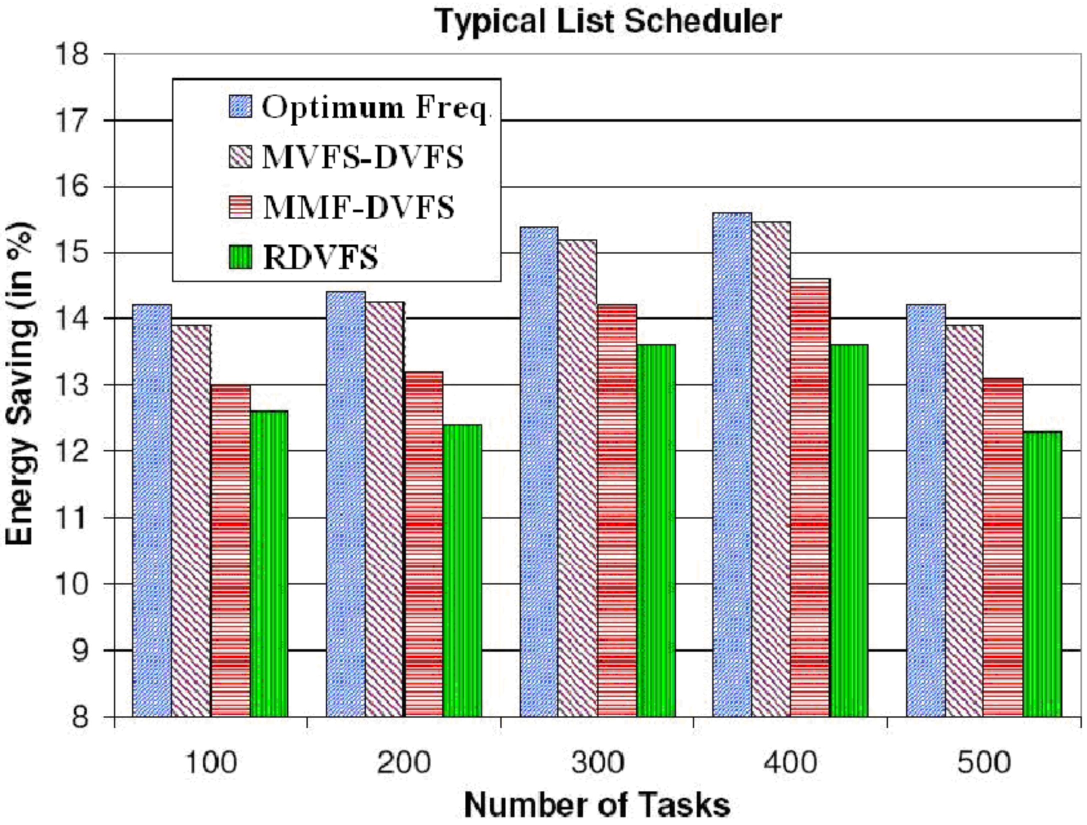
- Use DVFS to reduce energy footprint with minimal impact on wall time
  - Lots of work on that from 5-15 years ago

Slack time optimization example



Slack time optimization results

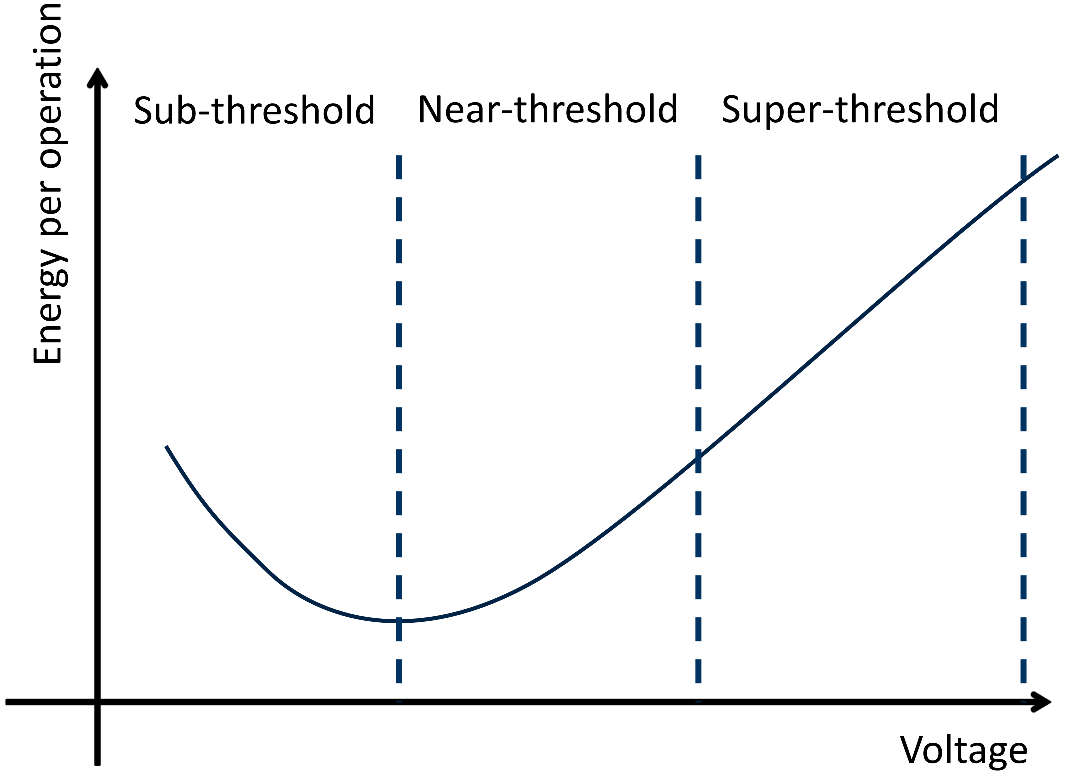
- Rizvandi et al. „Some Observations on Optimal Frequency Selection in DVFS–based Energy Consumption Minimization”
  - Simulations with 3000 randomly generated task graphs
  - Energy savings 10-20%



https://arxiv.org/ftp/arxiv/papers/1201/1201.1695.pdf

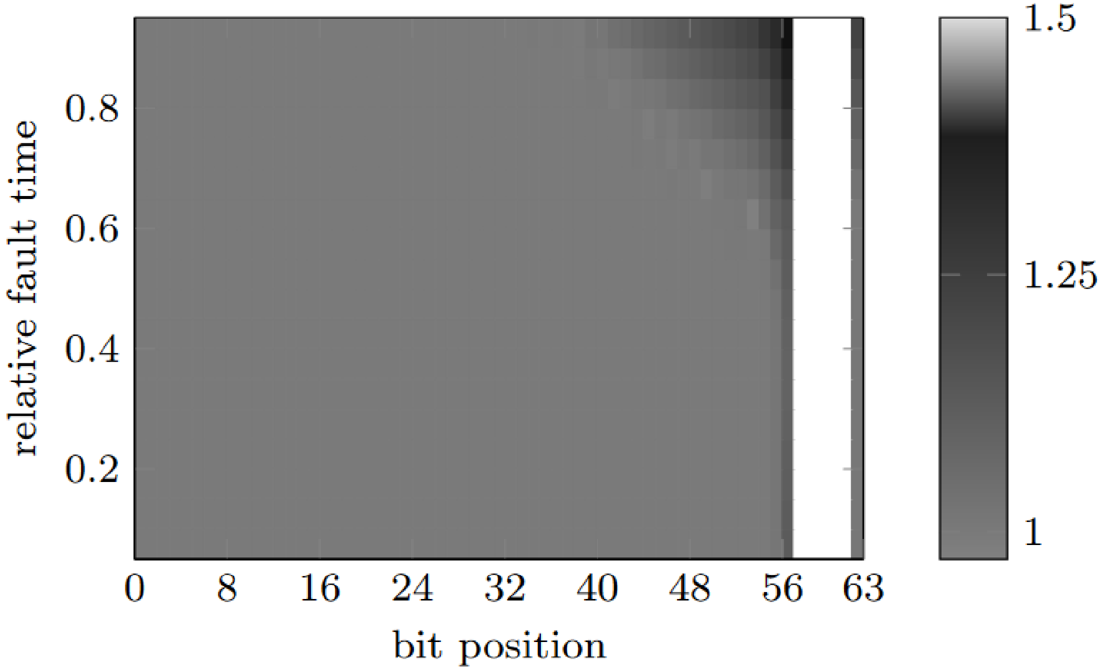
Near-Threshold Voltage

- Idea: reduce voltage below safe operating levels
  - Near-threshold voltage computing (NTV/NTC)
  - Can still operate transistors, but at large clock frequency reductions
    - Slowdown of 5x-10x
    - Might produce computational errors



Approximate Computing

- Idea: use NTV and mitigate slowdown with parallelism
  - Single, reliable core at super-threshold voltage
  - Several unreliable ones at near-threshold voltage (under the same power envelope!)
  - Switch cores depending on state of computation
- Investigate effects of bit flips in floating point data for converging algorithms (e.g. jacobi)



Approximate Computing cont’d

- Idea: reduce voltage below safe operating levels
  - Near-threshold voltage computing (NTV/NTC)
  - Replace single, reliable core with several unreliable ones under the same power envelope
    - Speedup through parallelism results in energy savings

- Investigate effects of bit flips in floating point data for converging algorithms (e.g. jacobi)

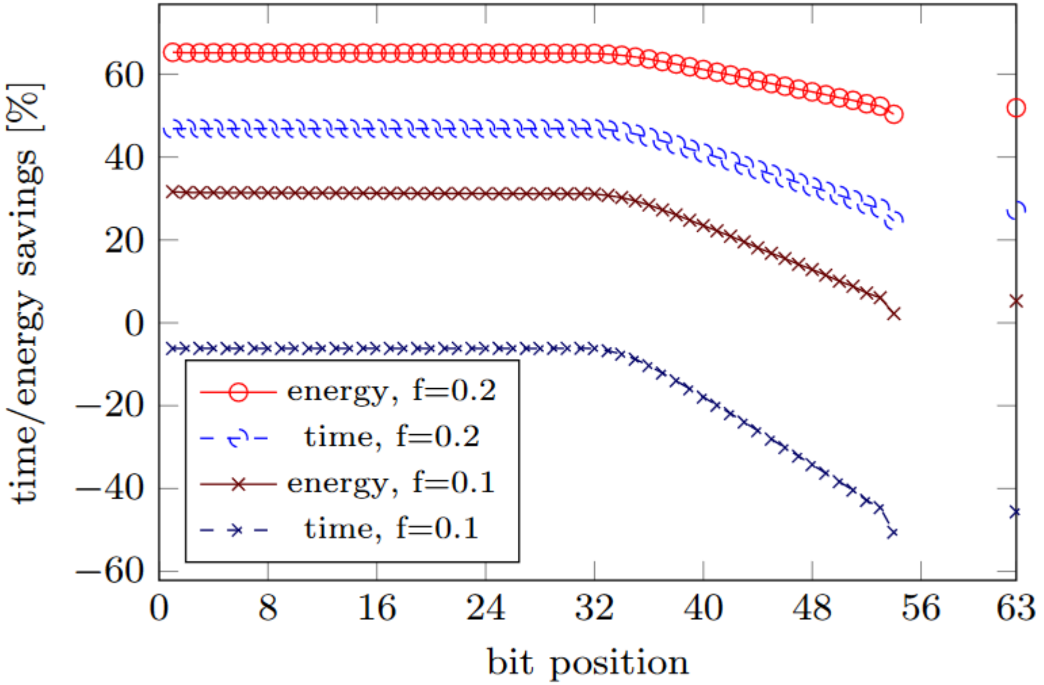


Fig. 4: Relative energy and time savings of an unreliable, parallel run of Jacobi on 16 cores compared to a reliable, sequential one. The missing data at bits 55–62 denotes divergence.

Additional Consequences

- Job Scheduling: Profile applications, compute roofline model, set optimal DVFS setting for consecutive runs
  - E.g. EAR – Energy Aware Runtime (SLURM tool @ SuperMUC, LRZ, Germany)

- Influences load balancing & scheduling decisions
  - Data movement is expensive
  - Move tasks to location of data instead of data to location of tasks

- Stacked memory, processing-in-memory, etc.

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)

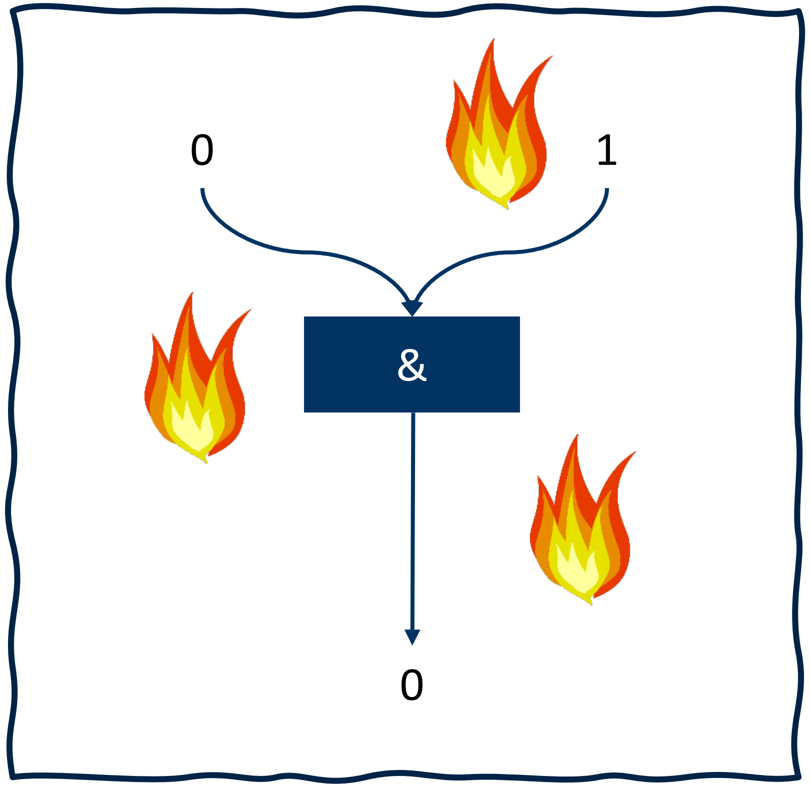
- The metrics are often flawed
  - e.g. PUE cannot be used to compare HPC sites in different climate zones
- There are diverging interests
  - Operator: minimize power/energy, maximize workload throughput
  - User: minimize wall time
  - Taxpayer/politicians: minimize costs

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
- 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

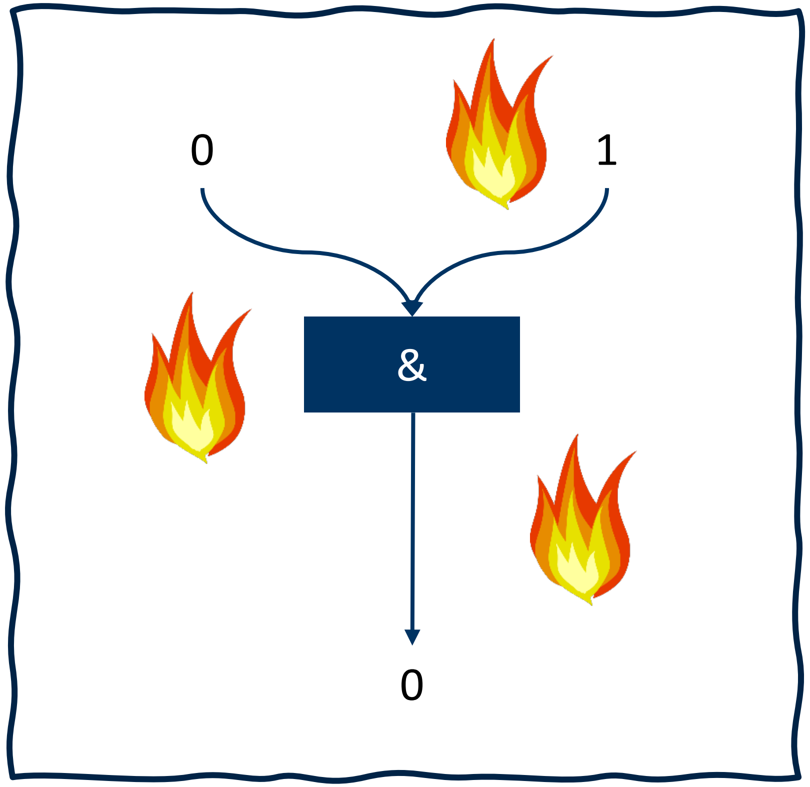
Reversible computing and Landauer principle: the future?

- 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? cont’d

- 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



Summary

- Energy is a hot topic in HPC

- Instrumentation and measurements got a lot easier over the past decade
  - Availability not an issue these days
  - Accuracy and granularity however is

- Discussed several research perspectives on the topic
  - Multi-objective optimization, slack time optimization, approximate computing, …

Image Sources

- Voltech PM1000: https://www.voltech.com/media/8d8595acca6d01f/pm1000-user-manual-v14.pdf
- Shunt transistor: https://articles.saleae.com/oscilloscopes/how-to-measure-current-with-an-oscilloscope
- PowerMon: https://ieeexplore.ieee.org/abstract/document/5453824?casa_token=Ax4_mcUUF0IAAAAA:YFY5X2H6aCU2pMDs5gwvMqbvA28huJePfLkRDveibf6d1TKkKmvqXfgCVtyVz1nZCp_z8-mIT9U
- PowerSensor 2: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8366941&casa_token=T70QZiDS9F4AAAAA:O02oYwOTJXaLIRj8as2ZAGQSmYUDeigrjd5Mt-sAIGfegoz0NlH25rfXL1gsEM8mmM6WYtr6DdE
- PCIe riser: https://www.igorslab.de/en/power-recording-graphics-card-power-supply-interaction-measurement/3/
- Supermuc-NGn nodes: https://www.lenovo.com/us/en/p/servers-storage/servers/high-density/thinksystem-sd650-n-v2/77xx7dsd672?orgRef=https%253A%252F%252Fwww.google.com%252F
- Intel Architecture Day Slide: https://download.intel.com/newsroom/2021/client-computing/intel-architecture-day-2021-presentation.pdf
- Alder Lake Die Shots: https://www.reddit.com/r/intel/comments/qhbbow/10nm_esf_intel_7_alder_lake_die_shot/