2025 EDUCADO Training School

Green computing

Oleksiy M. Kozlov
Staff scientist, Sustainable Computing Advisor

The Exelixis Lab
Heidelberg Institute for Theoretical Studies (HITS gGmbH)
exelixis-lab.org | alexey.kozlov@h-its.org

Padova – February 27, 2025

About

- Me
  - Background in Computer Science: B.Sc. 2007, M.Sc. 2013
  - 2013 – 2018 PhD, since 2018: Staff scientist
  - HPC software for evolutionary bioinformatics
  - Since 2023:. Sustainable Computing Advisor (part-time)
- My affiliation
  - Heidelberg Institute for Theoretical Studies (HITS gGmbH)
  - Privately funded, 12 PIs, ~100 researchers
  - Computational sciencies: Bioinformatics, astrophysics, statistics …

Green Computing: Challenges

- Extremely broad topic
  - From electric grid to CPU to software to users
  - Different cultures / mindsets / languages
  - “Divide-and-conquer” is problematic
- Many non-technical factors
  - Economics, politics, bureaucracy, psychology...

Agenda

- Motivation
- Energy monitoring
- Energy optimization
- Carbon-aware computing

Motivation

Green scientific computing?

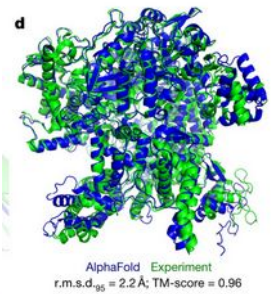
Be sure to optimize your code and all will be well!

The AI *hype* is a problem

How it started...

NEWS | 30 November 2020

'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures



aws — Carbon neutral by 2040

Microsoft — Carbon negative by 2030

Google — Carbon neutral since 2007

...and how it's going

Green | Cleaner Tech

Microsoft's AI Push Imperils Climate Goal as Carbon Emissions Jump 30%

The company's goal to be carbon negative by 2030 is harder to reach, but President Brad Smith says the good AI can do for the world will outweigh its environmental impact.

A.I.

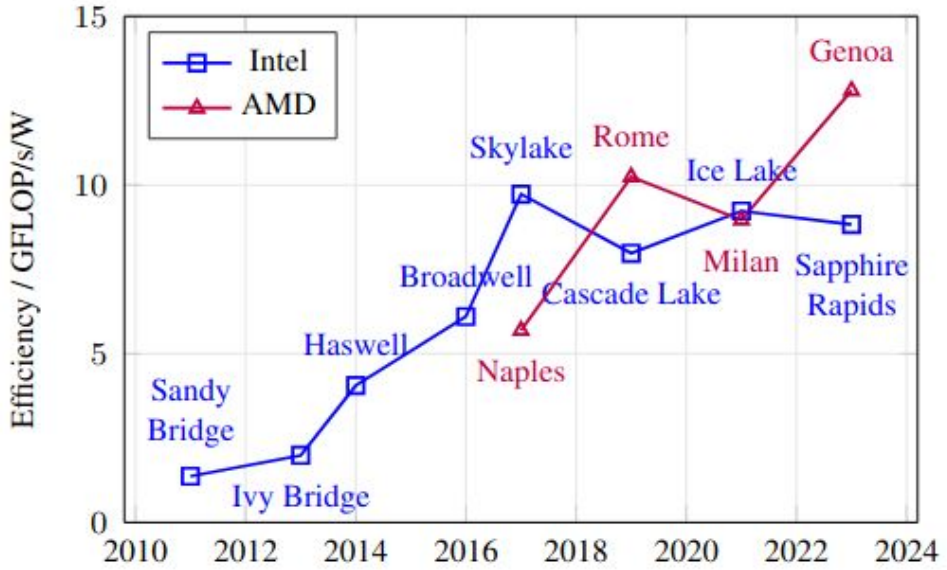
AI energy demand is ruining Google's environmental goals

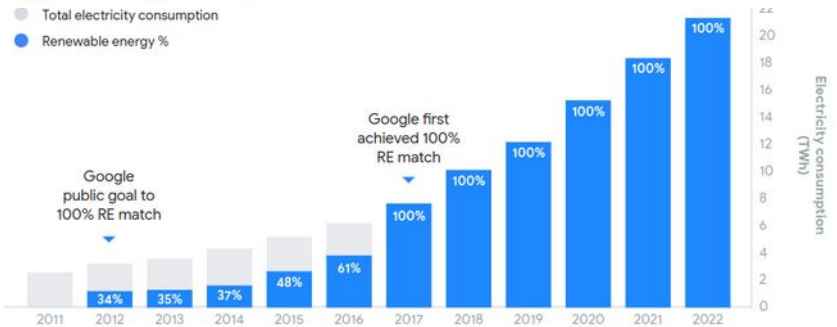
The tech giant said its carbon emissions are up 50% since 2019 because of how much energy is needed for its data centers and supply chains

HOME › NEWS › THE CRITICAL POWER CHANNEL

ExxonMobil plots natural gas power plant to exclusively power data centers

Higher efficiency = less energy use?





Suarez et al. (2024)
http://export.arxiv.org/abs/2411.16204

Google Environment report (2023)

The Jevons paradox

- W. S. Jevons „The Coal Question“ (1865):

Improved efficiency ⟵ Increased consumption rate

Lower cost

a.k.a. “rebound effect” or “induced demand”

Images: dpa, Wikipedia, techcrunch.com, https://www.busandmotorcoachnews.com ,OpenAI, DeepSeek

So what can we do?

- (Make software faster) → to keep users happy
- Measure & reduce energy consumption
- Measure & reduce CO2 emissions
- (Help to) clean up the grid + electrify everything

Energy monitoring

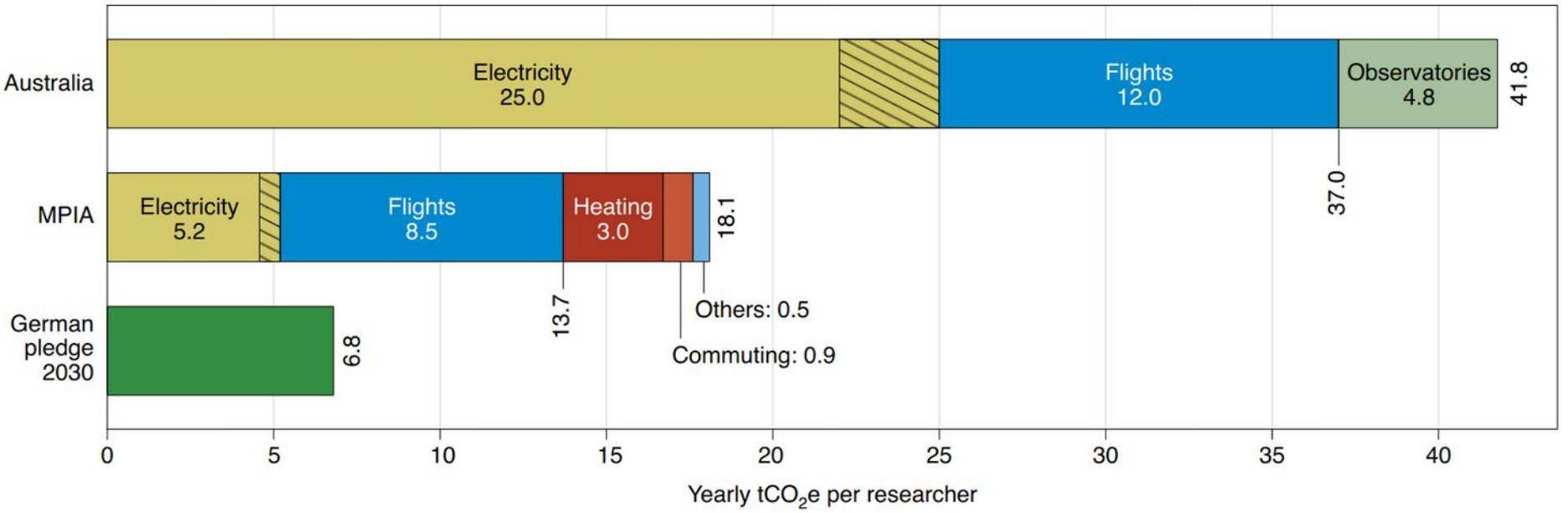
Energy monitoring: goals

- Measuring energy *accurately* is hard!
- Let’s be pragmatic
  - All measurements are wrong, but some are useful
  - Feasibility over Simplicity over Accuracy
  - Consistency important for comparability

Energy monitoring: use case #1

- General awareness
  - Impact vs. other areas, e.g. transportation
  - Averaged estimates are OK → W/core, CO2/kWh ...

Example: MPIA CO_2 footprint (2018)



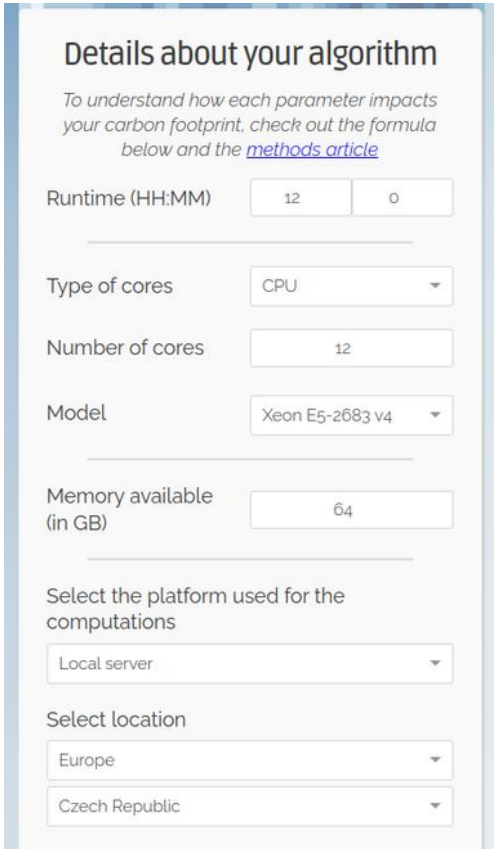
(Jahnke et al., Nat Astronomy 2020)

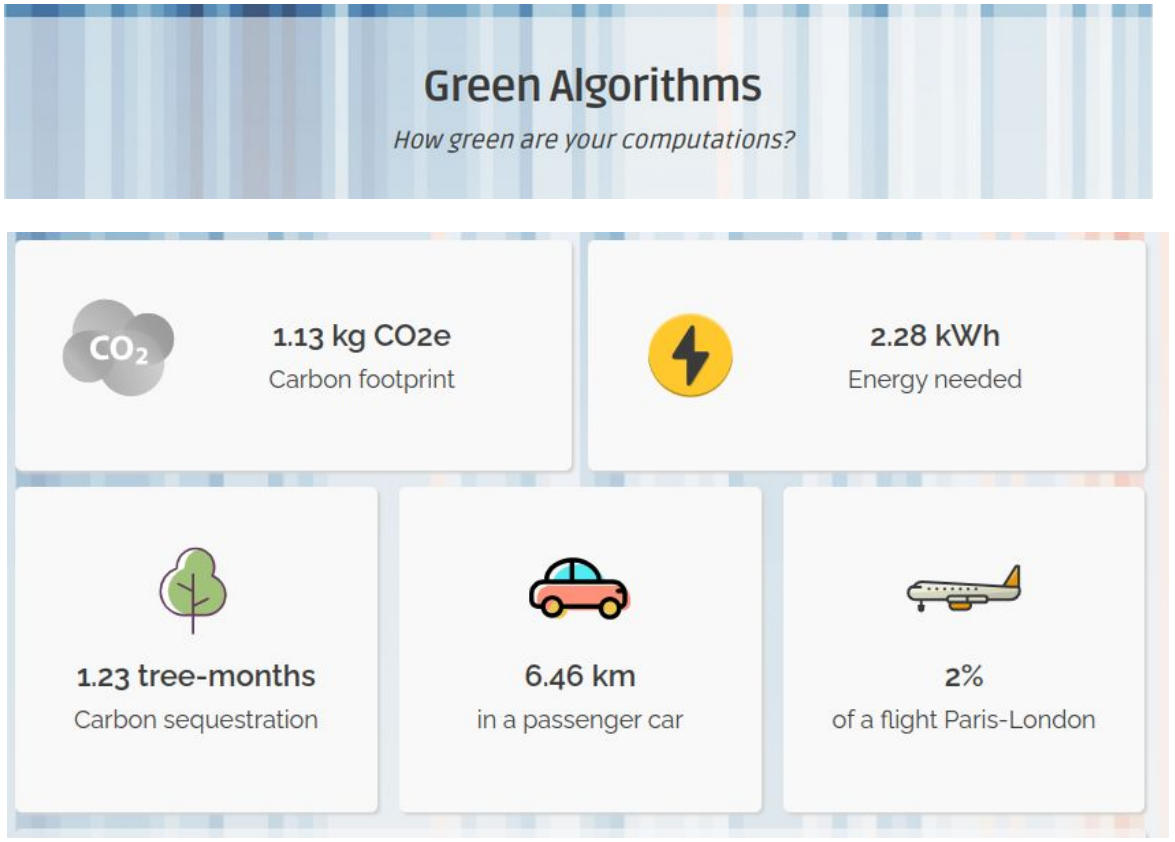
Example: Top-down estimation

- The analysis took 1,200,000 CPU-hours on the SuperMUC-NG supercomputer (LRZ, Garching, Germany)
- In 2021, LRZ energy consumption was 32,632,950 kWh [LRZ1], and in total 2,308,500,000 CPU-hours were allocated to user jobs [LRZ2]
- On average, this corresponds to roughly 14 Wh per CPU-hour, or 17,000 kWh for the full analysis
- This translates to ~7,200 kg of CO2 based on carbon intensity of the German electricity mix (0.425 kgCO2/kWh in 2021 [UBA])
- This is roughly equivalent to 17 NY->London flights (one-way) [Google]

(Stiller, 2022 Nature)

Example: Bottom-up estimation





http://calculator.green-algorithms.org/

Energy monitoring: use case #2

- Comparative analysis / benchmarking
  - Year-to-year, programs, parameters etc.
  - Actual measurements needed
  - Systematic under-/overestimation is OK

Energy measurement levels

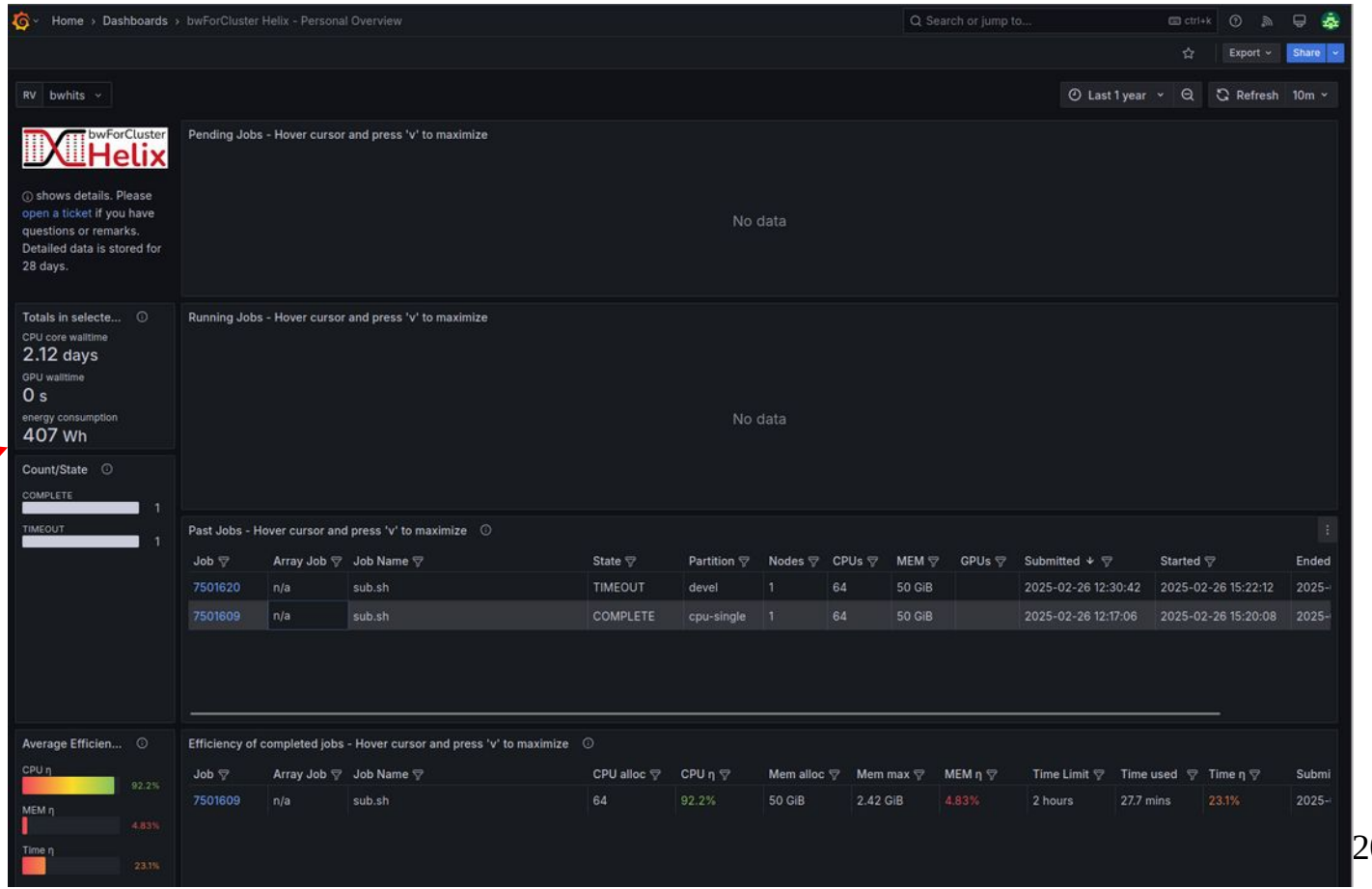
- Building / Datacenter → smart meters
- Server room, rack → smart PDUs
- Node / Server
- CPU / GPU
- Job
- User

Energy monitoring: Toolbox

	IPMI	NV-SMI	ROcM-SMI	RAPL
Platform	Servers	GPU (Nvidia)	GPU (AMD)	CPU (Intel, AMD)
Scope	Full system	GPU	GPU	CPU+DRAM
Power	✔	✔	✔	✘
Energy	✘	✘	✔	✔
Resolution	low	medium	medium	very high
Low latency	✘	depends?	???	✔
Reliability	✘	???	???	✔
Non-root access	✘	✔	✔	✔ / ✘
Power limiting	✔ / ✘	✔	✔	✔

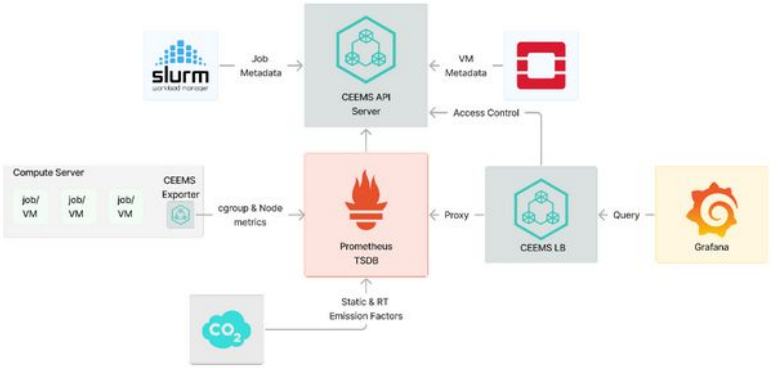
Example: HELIX job monitoring

Estimated :(



Example: CEEMS

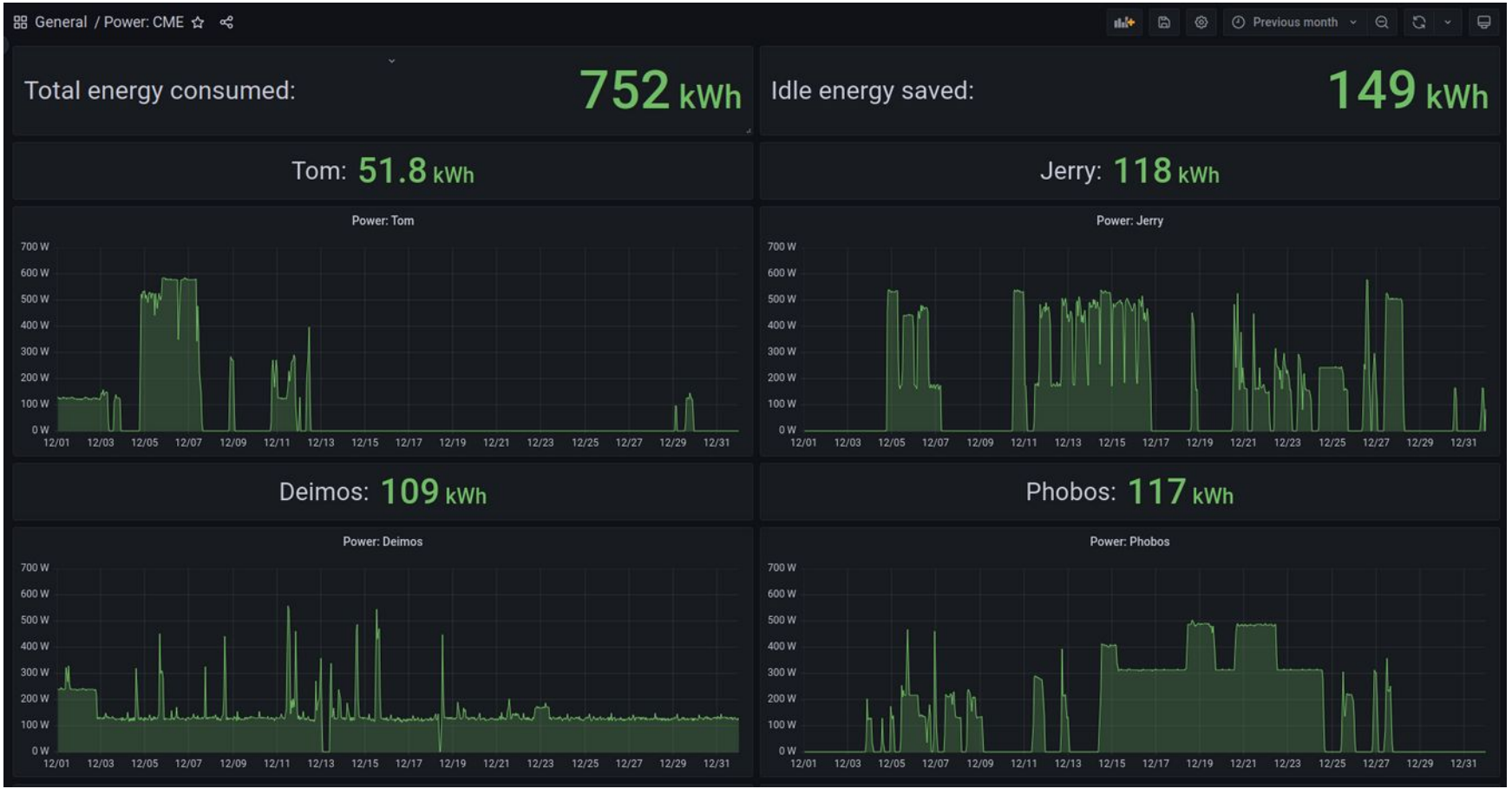
- Compute Energy & Emissions Monitoring Stack
  - Performance, Energy and CO2 monitoring
  - SLURM, Prometheus and Grafana integration
  - Requires node agent
  - https://github.com/mahendrapaipuri/ceems
- Alternatives:
  - ClusterCockpit: https://www.clustercockpit.org/
  - Llview: https://github.com/FZJ-JSC/LLview
  - Kepler (k8s): https://sustainable-computing.io



Non-invasive energy monitoring

- Group servers
  - IPMI → telegraf → influxDB → Grafana
  - Resolution: 30 s
  - HowTo: https://github.com/amkozlov/ipmi-grafana
- HITS clusters
  - IPMI → checkmk → CSV → influxDB → Grafana
  - Resolution: 1 min

Group servers power



Cluster nodes power: aggregated

+ storage + network + cooling + PSU conversion losses

Energy monitoring in SLURM

- Plugins for RAPL, IPMI...
- Node power
- Job energy

```
[user@cascade-login ~]$ scontrol show node cascade-149
NodeName=cascade-149 Arch=x86_64 CoresPerSocket=20
   CPUAlloc=0 CPUTot=40 CPULoad=0.00
   AvailableFeatures=cpu6230,ram96,rtx2080
   ActiveFeatures=cpu6230,ram96,rtx2080
   Gres=gpu:2,cpuonly:1
   NodeAddr=cascade-149 NodeHostName=cascade-149 Version=20.11.7
   OS=Linux 4.18.0-147.8.1.el8_1.x86_64 #1 SMP Thu Apr 9 13:49:54 UTC 2020
   RealMemory=95000 AllocMem=0 FreeMem=92900 Sockets=1 Boards=1
   MemSpecLimit=2048
   State=IDLE ThreadsPerCore=2 TmpDisk=0 Weight=1 Owner=N/A MCS_label=N/A
   Partitions=debug.p
   BootTime=2021-06-09T11:20:43 SlurmdStartTime=2021-06-09T15:02:11
   CfgTRES=cpu=40,mem=95000M,billing=40,gres/gpu=2
   AllocTRES=
   CapWatts=n/a
   CurrentWatts=18 AveWatts=14
   ExtSensorsJoules=n/s ExtSensorsWatts=0 ExtSensorsTemp=n/s
   Comment=(null)
```

```
[user@cascade-login ~]$ sacct -j 880244 -o ConsumedEnergy
ConsumedEnergy
--------------
        70.11K
        70.11K
        70.10K
```

Energy in SLURM: problems

- Poor visibility, no summary stats
- Wrong RAPL measurements on modern Intel CPUs (Haswell+)
  - Fixed in SLURM v25.05
  - https://bugs.schedmd.com/show_bug.cgi?id=9956
- No support for NVIDIA GPUs (yet?)
  - But: AMD GPUs via ROCm-SMI library

Energy monitoring: RAxML-NG

- Energy usage report
  - Measured with Intel RAPL → CPU+DRAM only
  - Supported on Linux systems only
  - Can be disabled with ‘--extra energy-off’

```
Elapsed time: 42846.287 seconds

Consumed energy: 162370.469 Wh (= 812 km in an electric car, or 4059 km with an e-scooter!)
```

Energy-to-solution

Perun: Energy meter for Python

HELMHOLTZ AI | ARTIFICIAL INTELLIGENCE COOPERATION UNIT

- Python library and command line tool
- Samples performance counters
- Little to no code modifications required
- Works with MPI
- FAIR research software
- Available on the Python Package Index (PyPI)

https://pypi.org/project/perun/

```
$ perun monitor finetune_qa_accelerate.py

App name: finetune_qa_accelerate
First run: 2023-08-15T18:56:11.202060
Last run: 2023-08-17T13:29:29.969779
RUN ID: 2023-08-17T13:29:29.969779

| Host                | RUNTIME   | ENERGY     | CPU_POWER | CPU_UTIL |
|:--------------------|:----------|:-----------|:----------|:---------|
| hkn0432.localdomain | 995.967 s | 960.506 kJ | 231.819 W | 3.240 % |
| hkn0436.localdomain | 994.847 s | 960.469 kJ | 235.162 W | 3.239 % |
| All                 | 995.967 s | 1.921 MJ   | 466.981 W | 3.240 % |

The application has been run 7 times. In total, it has used 3.128 kWh,
released a total of 1.307 kgCO2e into the atmosphere, and you paid 1.02 €.
```

Questions?

Energy optimization

Idle consumption: Examples

- Intel i7-7800X, 6 cores, 64GB RAM
  - Idle: 55 W, sleep: 5 W, under load: 150W
  - 30% utilization → 42% savings (306 kWh/a)
- Xeon Platinum 8260, 48 cores, 764GB RAM
  - idle: 120 W, standby: 20 W, under load: 500 W
  - 50% utilization → 16% savings (438 kWh/a)
- AMD EPYC 9684X, 192 cores: Idle: 300 W - 400 W
- GPU server with 8 x H200: Idle: ~2000 W

Sleep-on-Idle

- Cluster nodes
  - Supported by SLURM: https://slurm.schedmd.com/power_save.html
- Desktops
  - Suspend-on-idle + Wake-on-LAN
- Rack servers
  - No suspend to RAM :(
  - PowerOff-on-idle + PowerOn-over-IPMI
  - https://github.com/amkozlov/idle-sleep

PowerOff-on-Idle

- Idle detection cron job
  - No active sessions + CPU utilization < 0.5 → idle
  - Idle since 1 hour → poweroff
  - Can be temporarily disabled:

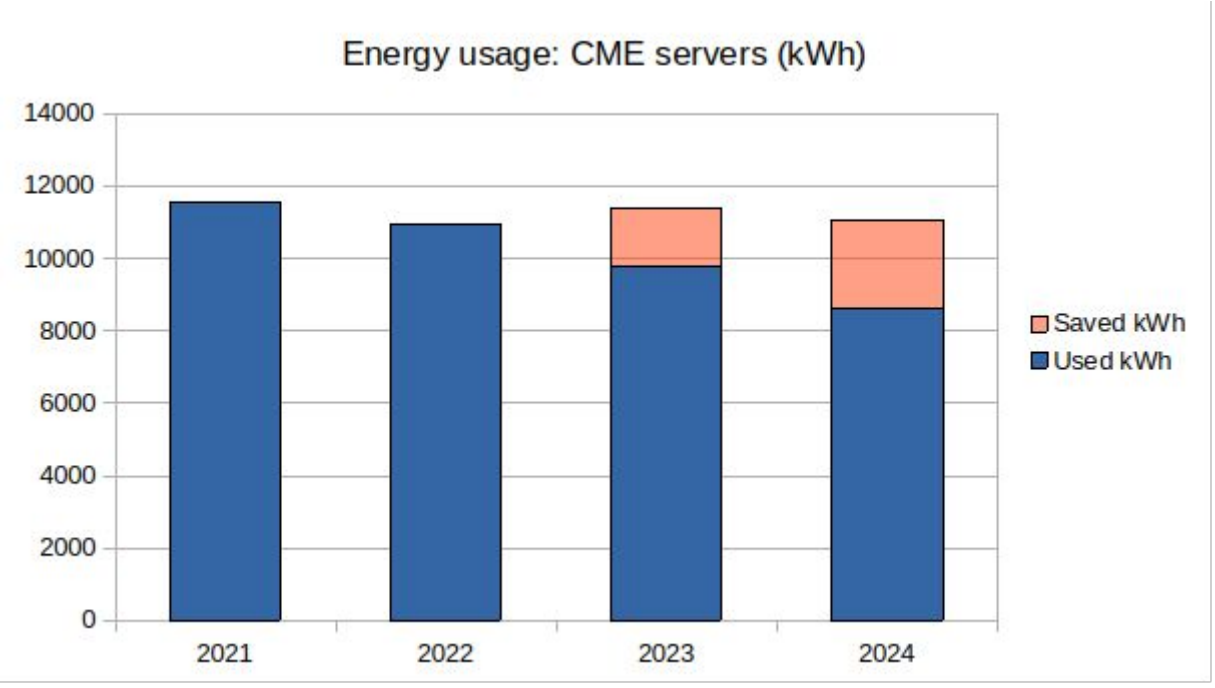
```
deimos$ ecosleep disable 12h
Server will not be powered off until: Di 16. Mai 10:13:34 CEST 2023
```

- PowerOn via SSH

```
laptop$ alias | grep ecowake
alias ecowake='ssh user@XXXX.h-its.org sudo /data/ecosleep/wakeup.sh'

laptop$ ecowake deimos
user@XXX.h-its.org's password:
Chassis Power Control: Up/On
waiting for deimos ..................................................
Server is back online!
```

Idle power savings 2024



6 group servers
(24-48 CPU cores)

Saved:
2450 kWh (22%)

Consumption:
-21% vs. 2022

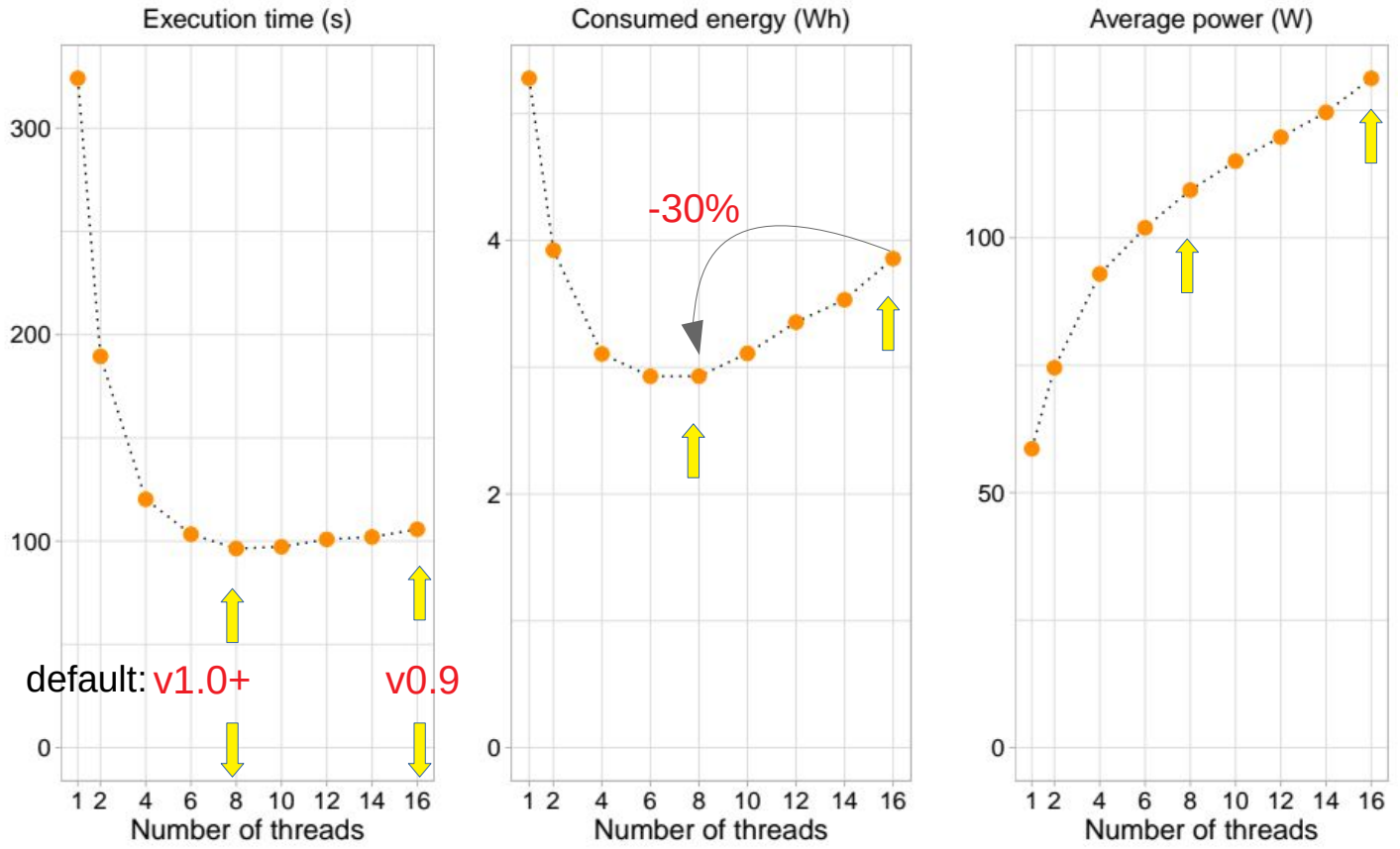
Scripts & HowTo: https://github.com/amkozlov/idle-sleep

Problems / Improvements

- screen/tmux sessions lost
  - tmux-resurrect might help
- boot delay 1-2 min.
- (force NFS remount after boot)
- No HW failures due to frequent on/off cycles!

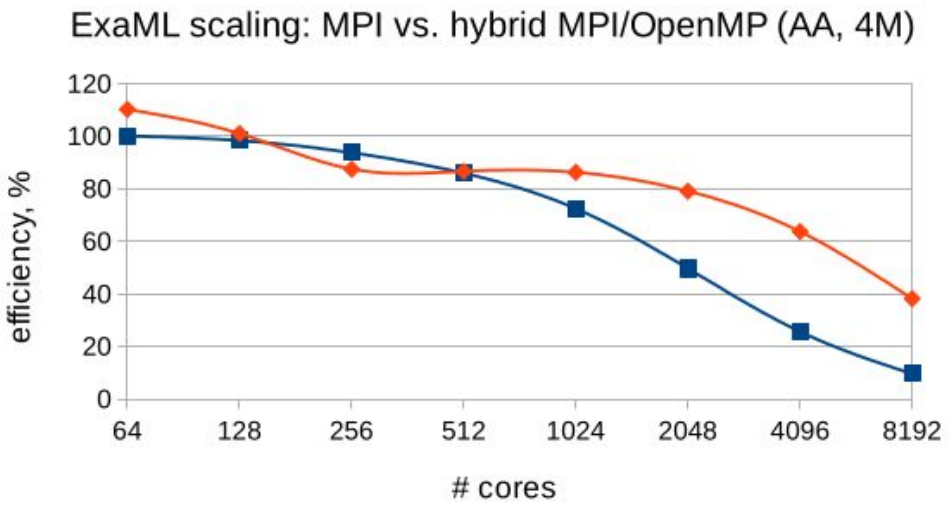
Parallelization tuning

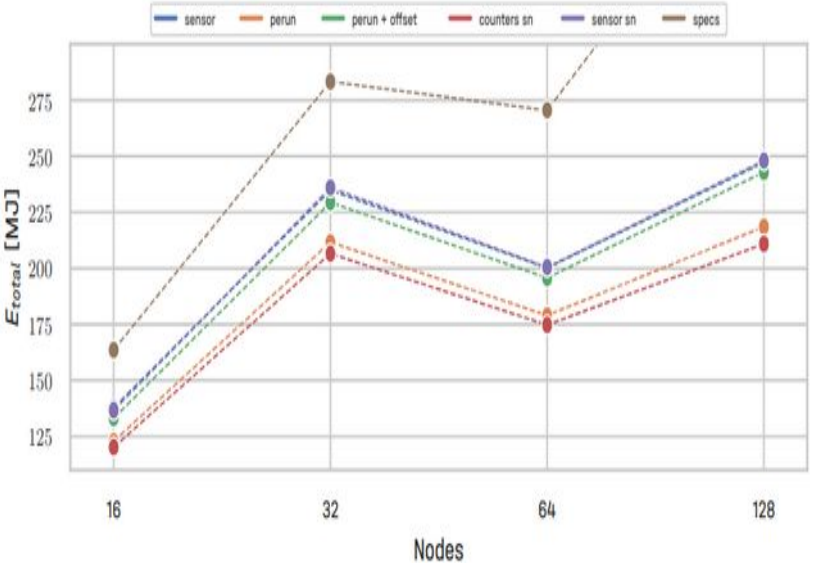
Automatic parallelization tuning (intra-node)



Amdahl’s law

- Diminishing returns from parallelization
- Fewer nodes / GPUs *usually* more efficient





Power scaling

Power scaling on CPU and GPU

- Widely available: Intel/AMD/NVIDIA
- Power and/or frequency limits

- Typical range: 50% - 100% TDP
- Easy-to-use, transparent to workload

NVIDIA GeForce RTX 2080 SUPER

```
$ nvidia-smi -q -d POWER,CLOCK
  Power Management                  : Supported
  Power Draw                        : 4.24 W
  Power Limit                       : 250.00 W
  Default Power Limit               : 250.00 W
  Enforced Power Limit              : 250.00 W
  Min Power Limit                   : 125.00 W
  Max Power Limit                   : 250.00 W
```

```
$ sudo nvidia-smi -pl 200
Power limit for GPU 00000000:17:00.0 was set
to 200.00 W from 300.00 W.
```

Intel Xeon Platinum 8260

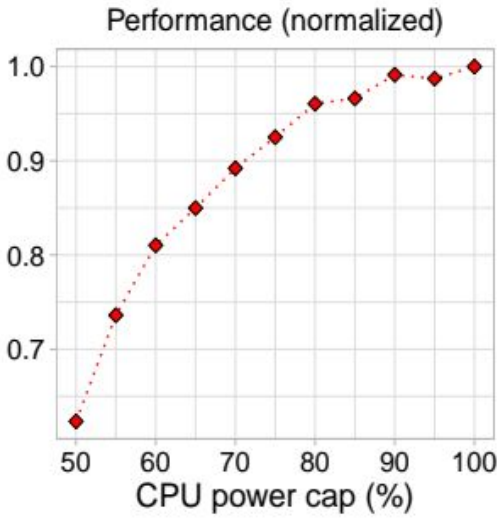
```
$ sudo cpupower frequency-info
  hardware limits: 1000 MHz - 3.90 GHz
  available cpufreq governors: performance powersave
  current policy: frequency should be within 1000 MHz and
3.90 GHz.
```

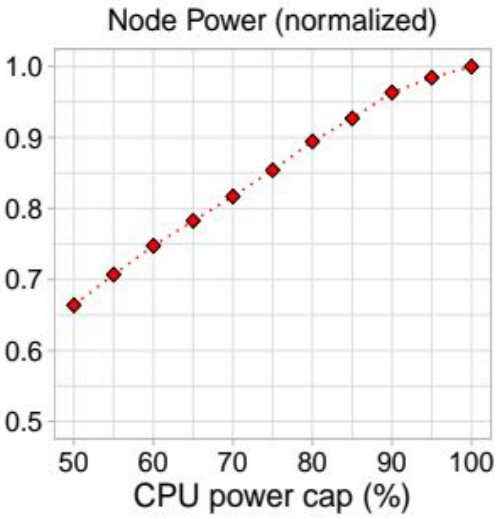
AMD EPYC 7452

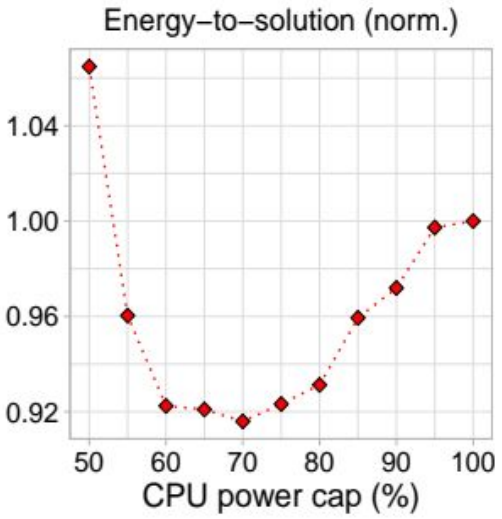
```
$ sudo cpupower frequency-info
  hardware limits: 1.50 GHz - 2.35 GHz
  available frequency steps:  2.35 GHz, 2.00 GHz, 1.50 GHz
```

```
$ sudo cpupower frequency-set -u 2000000
Setting cpu: 0
...
```

Energy efficiency “sweet spot”

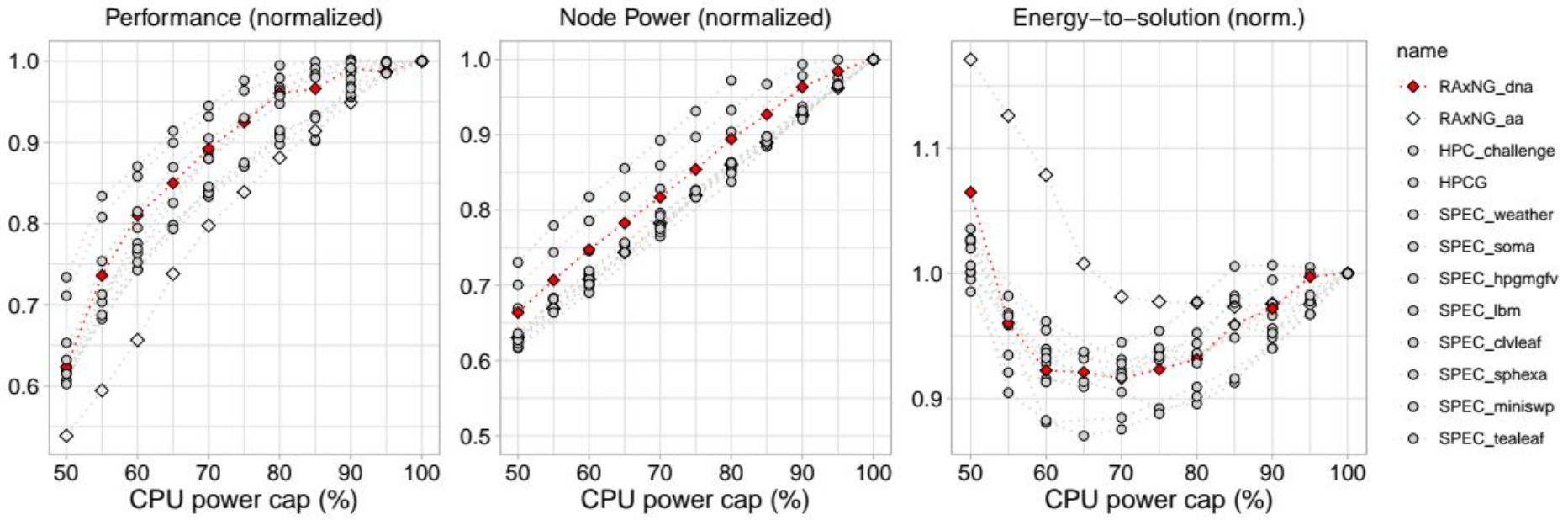




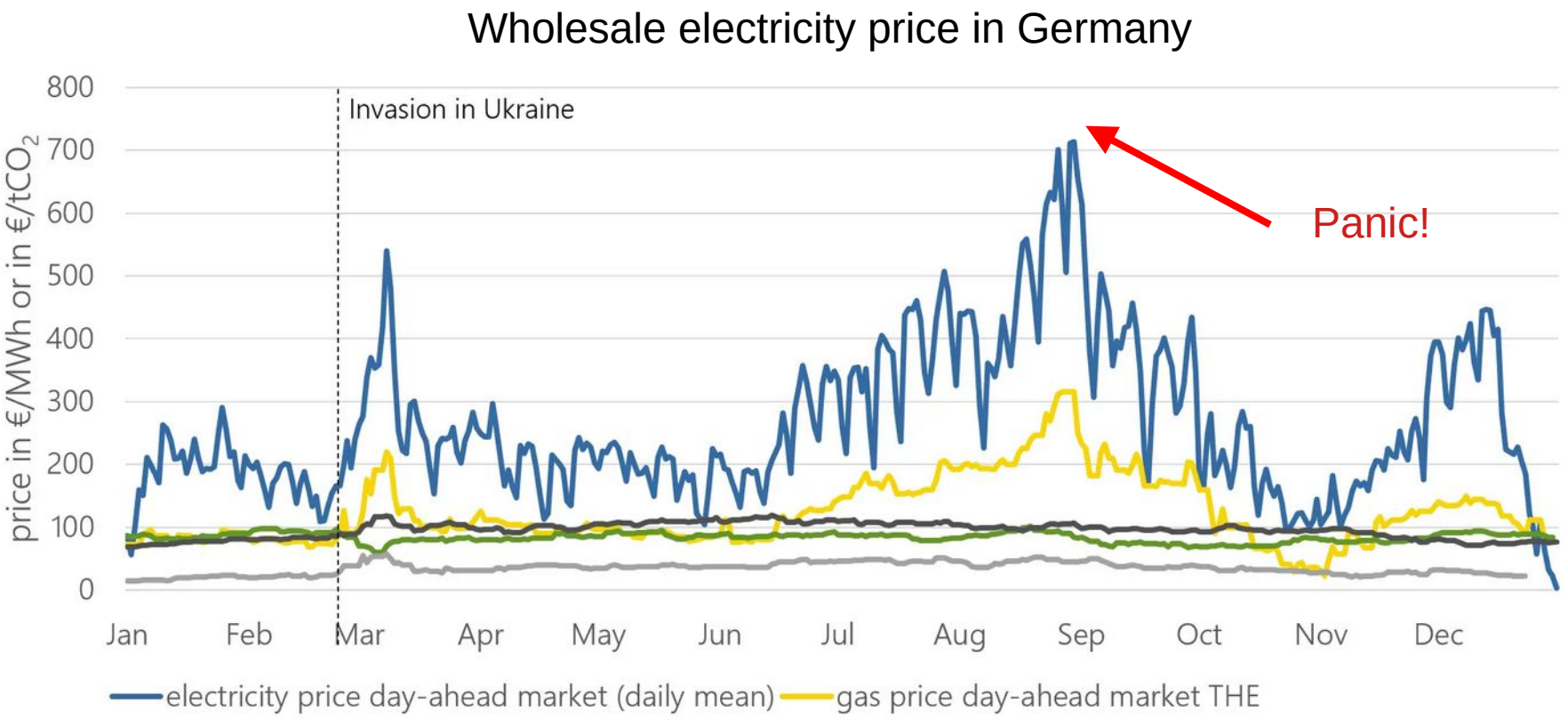


(RAxML-NG 1.1, 2x Intel Xeon Platinum 8260, 48T)

Workload variation



2022 energy crisis



Source: https://www.ffe.de/en/publications/german-electricity-prices-on-the-epex-spot-exchange-in-2022/

Energy crisis: reactions

Voir en français

CERN to implement additional energy-saving measures for 2022–2023

30 SEPTEMBER, 2022

https://home.web.cern.ch/news/news/cern/cern-implement-additional-energy-saving-measures-2022-2023

Federal Cabinet

Additional energy saving measures approved

The Cabinet has approved additional energy-saving measures that will help secure the energy supply in the short and medium term. Among other things, less office space is to be heated, while buildings, monuments and advertising installations are no longer to be illuminated at certain times.

Wednesday, 24 August 2022 3 min reading time

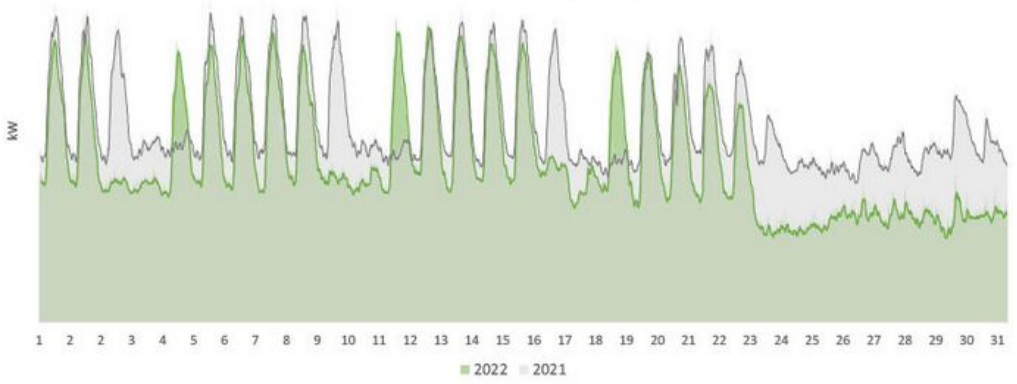
Brendan Rouse

31 March 2023

Category: General

Energy savings at EMBL – a 2022 update

Data centres account for 36% of all the electricity consumed by EMBL, so savings here have a material impact. Various measures have been taken across EMBL to improve IT service's energy efficiency, including limiting CPU speeds, moving data to more efficient data centres, and reducing the demand for cooling. All these measures have resulted in energy savings of 670,000 kWh (11%).



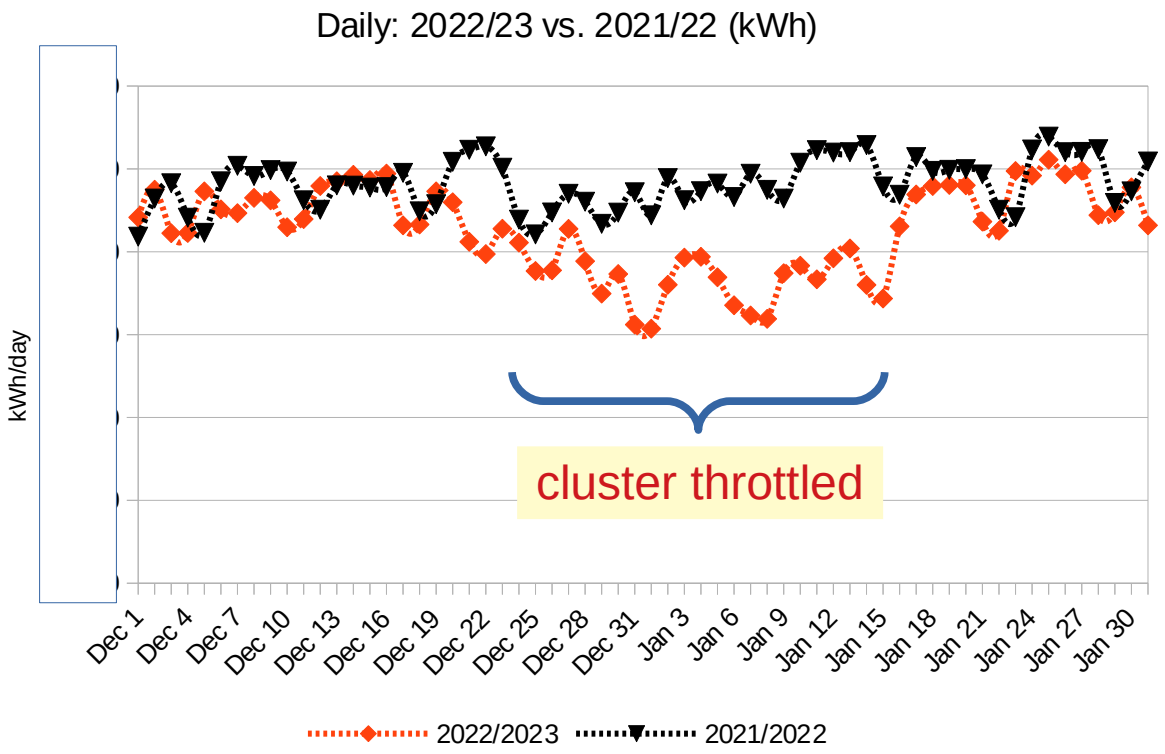
Electricity use in EMBL Heidelberg for the month of December in 2022 and 2021, showing an obvious reduction in 2022.

https://www.embl.org/about/info/sustainability/energy-savings-at-embl-a-2022-update/

“Christmas experiment”

- 23.12.2022 → Apply power throttling
  - Haswell cluster (URZ): CPU 2000 MHz
  - Cascade cluster (HITS): CPU 90 W / GPU 175 W
- 16.01.2023 → Back to normal power
- 16.01. – 29.01. → Baseline data collection

Energy consumption: HITS campus



Energy consumption: Clusters

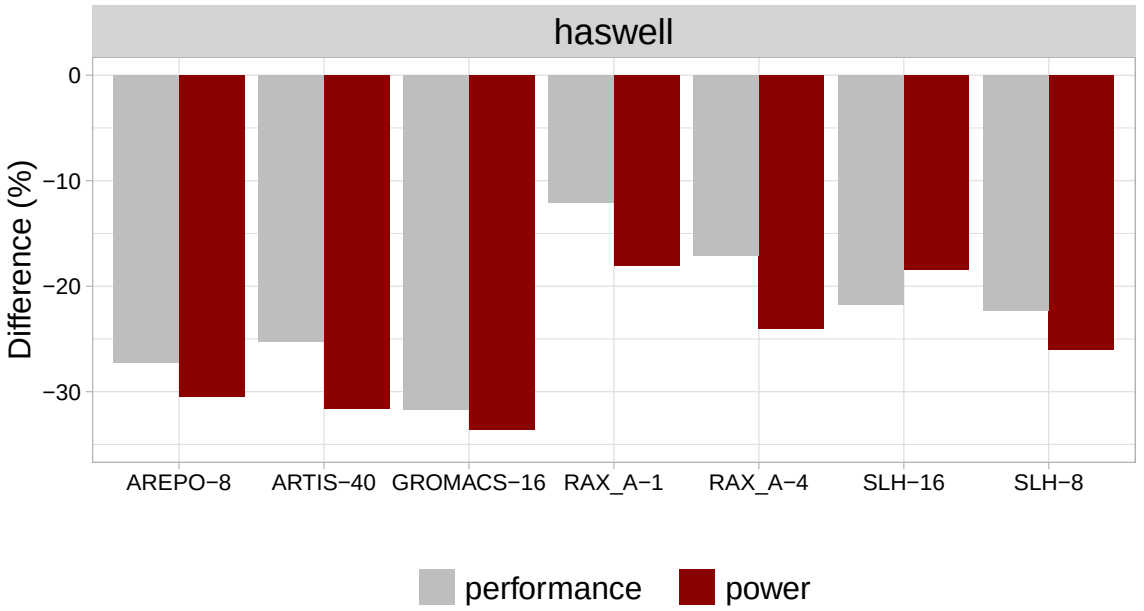
	Avg. node power (excl. idle)			Energy saved	Performance reduction (min … max)
	Regular 16.01 – 29.01	Reduced 23.12 – 15.01*	Diff.		
Haswell / URZ	150 W	100 W	-33 %	5110 kWh	-12 % ... -32 %
Cascade / HITS	294 W	220 W	-25 %	4791 kWh	-10 % ... -25 %

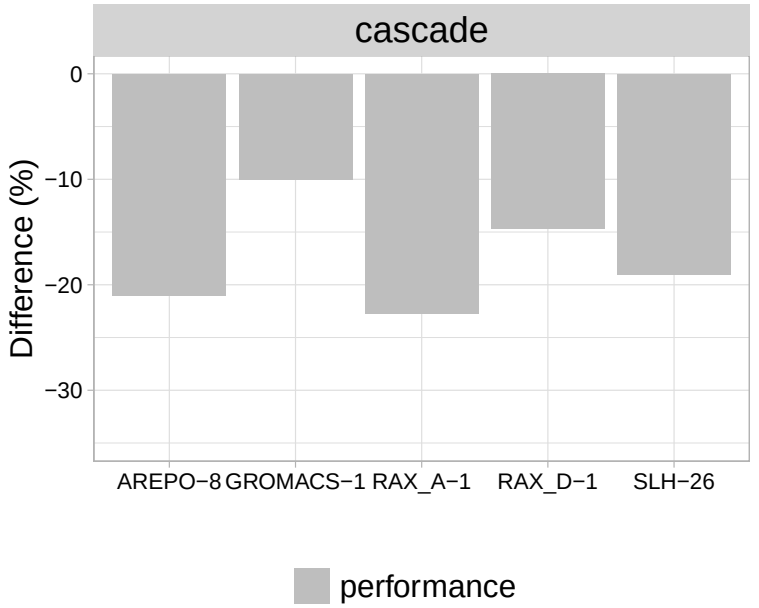
* cascade: 07.01-11.01 excluded due to storage failure

Estimated total savings (3 weeks):

 10,000 kWh =  5-7 years =  50,000 km

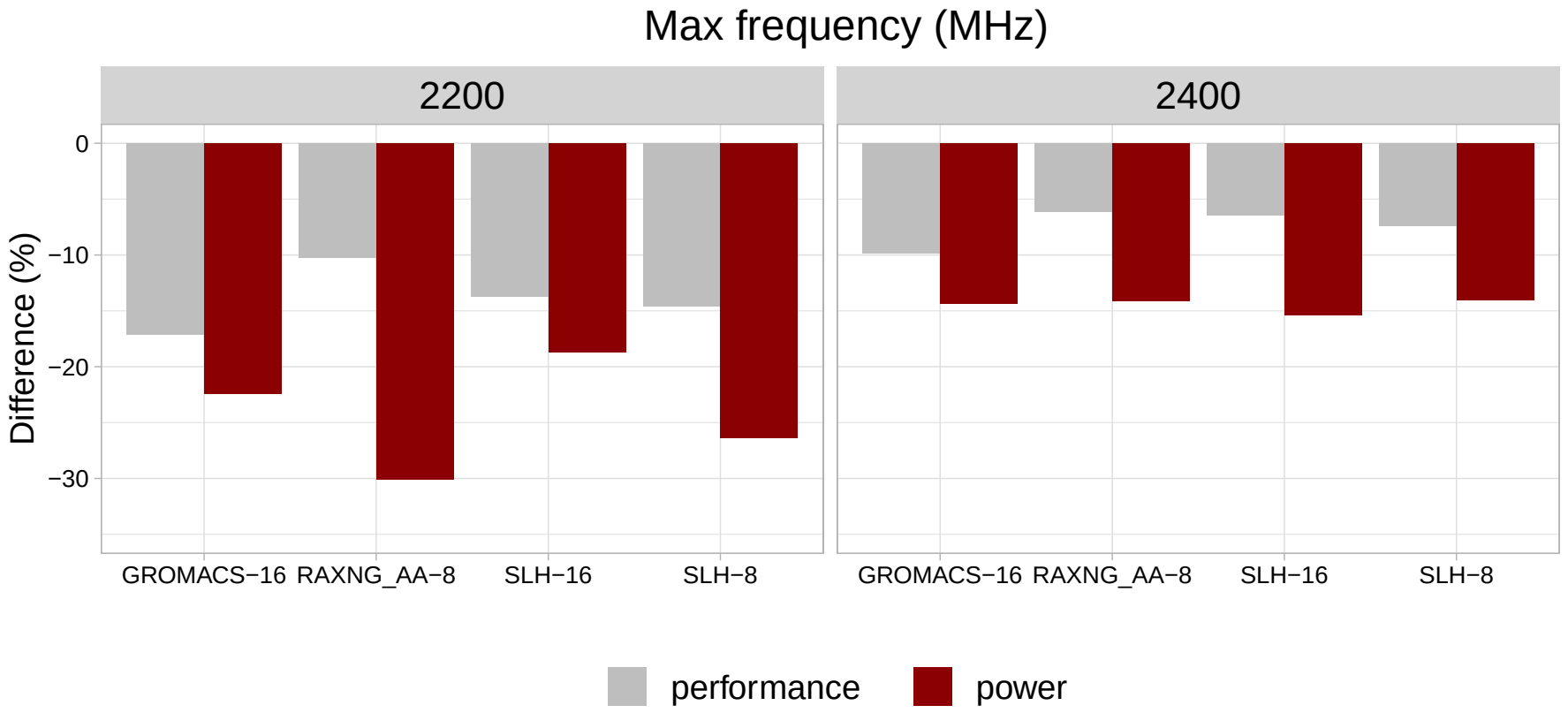
Performance vs. power reduction





- performance loss is sublinear w.r.t. power
- BUT: for many workloads, “free lunch” is small: 2000 MHz below efficiency sweet spot?

Searching for the sweet spot



EAR: Energy Aware Runtime

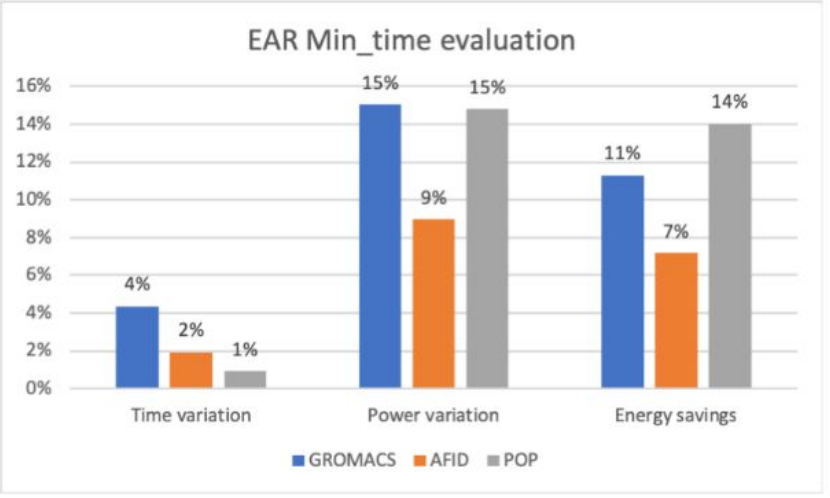
- Energy monitoring & optimization framework
  - Policy-based CPU frequency tuning
  - Open-source: https://gitlab.bsc.es/ear_team/ear
  - SLURM integration

EAR Energy Aware Runtime

Lenovo

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

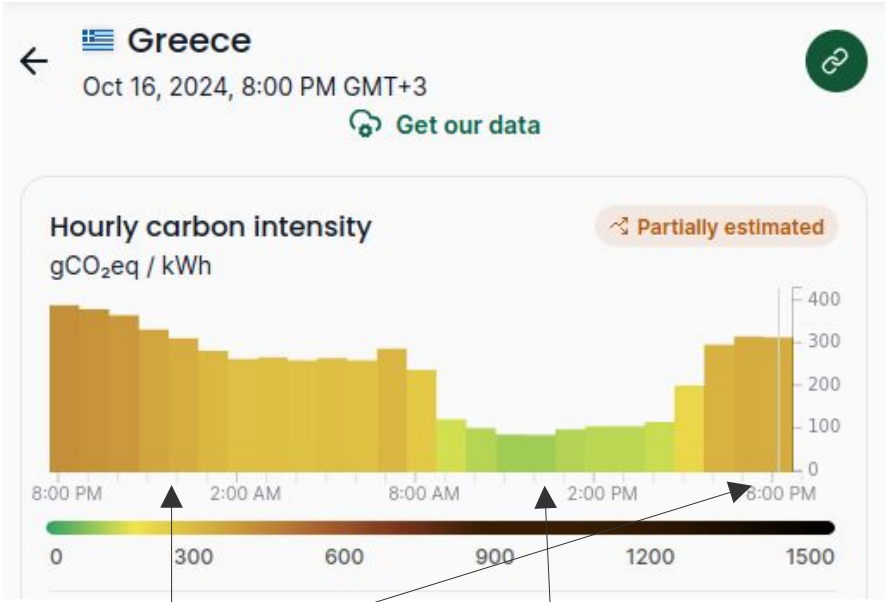
```
#!/bin/bash
#SBATCH -J jobname
#SBATCH --time=24:00:00
#SBATCH --partition=test
#SBATCH --ear=on
#SBATCH --ear-policy=min_time
#SBATCH --ear-cpufreq=2400000
```

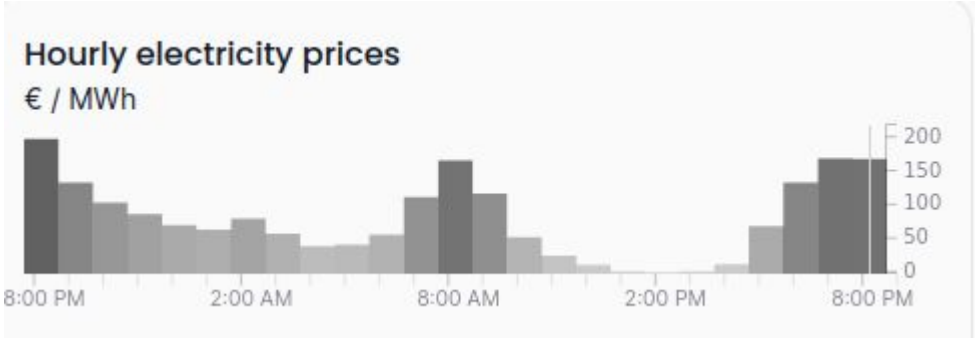


Questions?

Carbon-aware computing

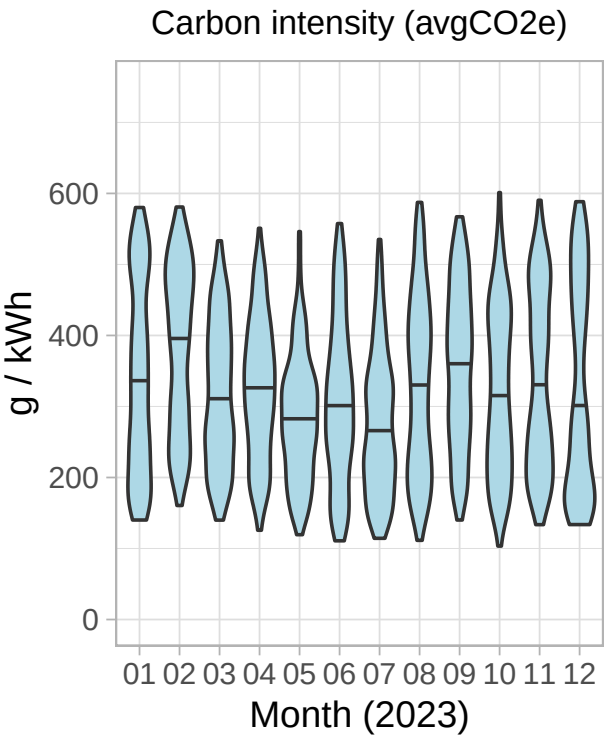
Real-time energy mix

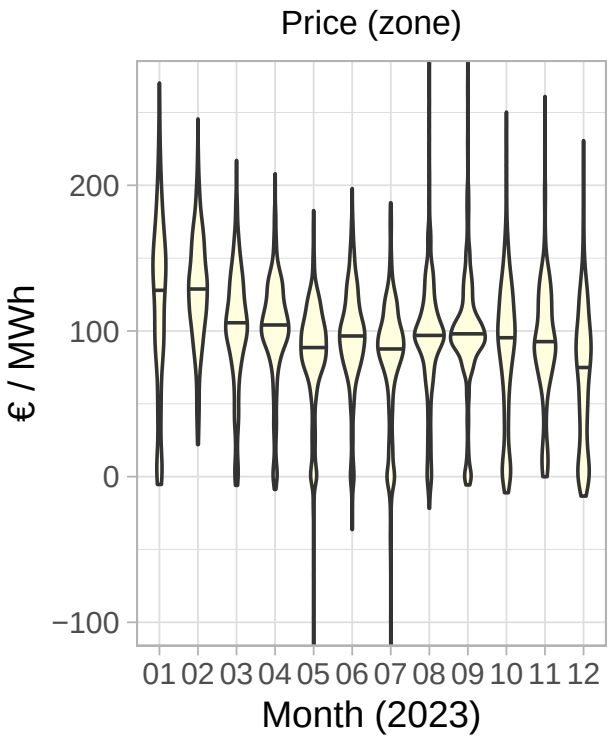


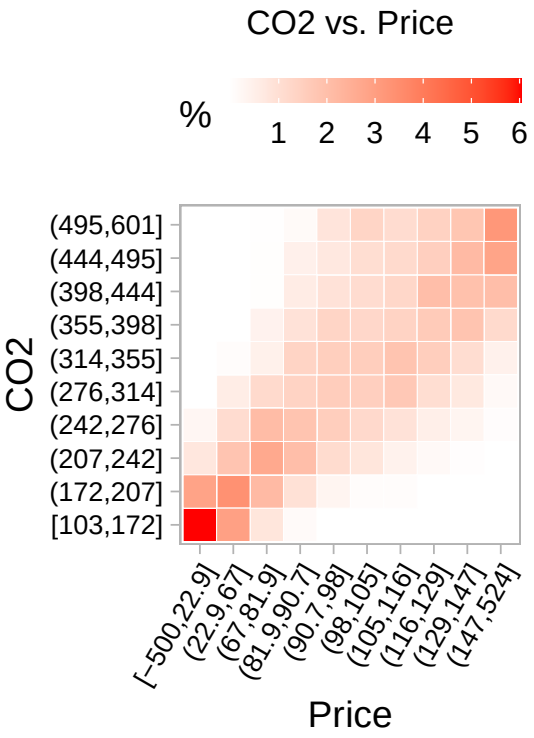


https://app.electricitymaps.com

CO2 and price profiles: Germany



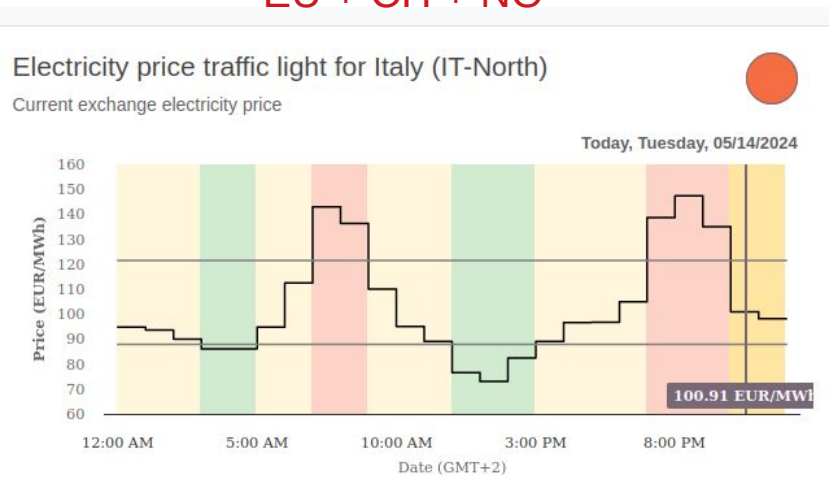




Carbon “traffic light”

- Discrete signal, e.g. green / yellow / red
- Ideally, reflects *local* marginal CI

EU + CH + NO



https://energy-charts.info

UK

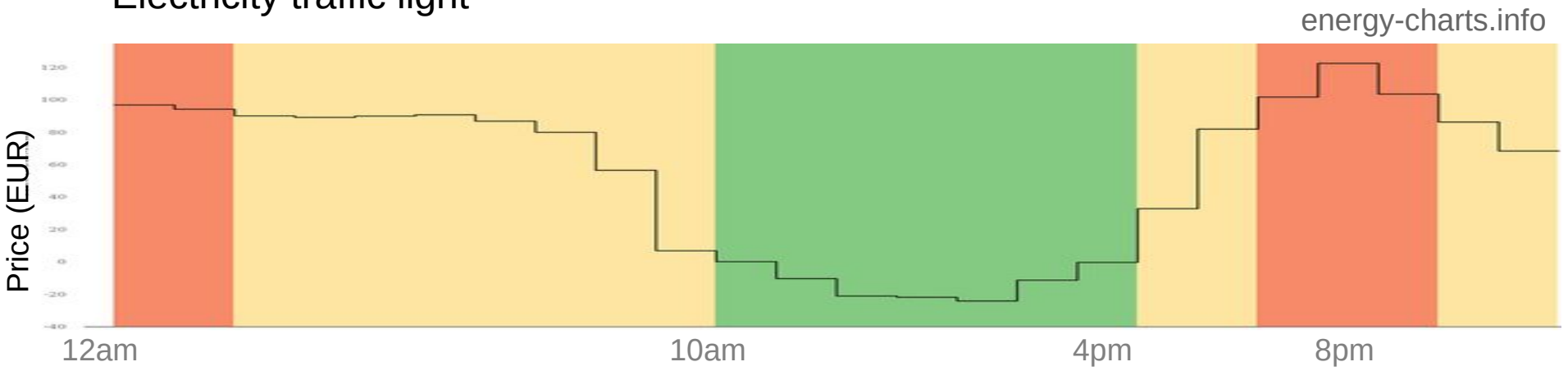
https://carbonintensity.org.uk/

Germany, BW

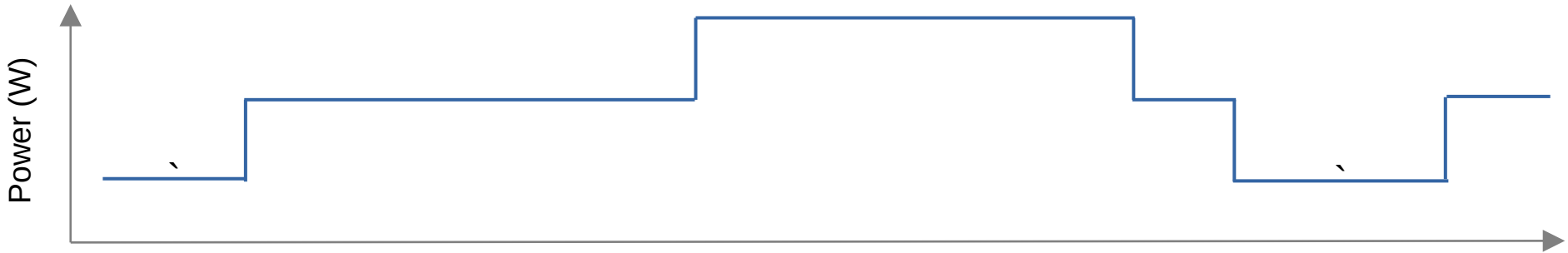
https://www.stromgedacht.de/

Carbon-aware power scaling

"Electricity traffic light"



Use less energy in red, more in green phases → EcoFreq

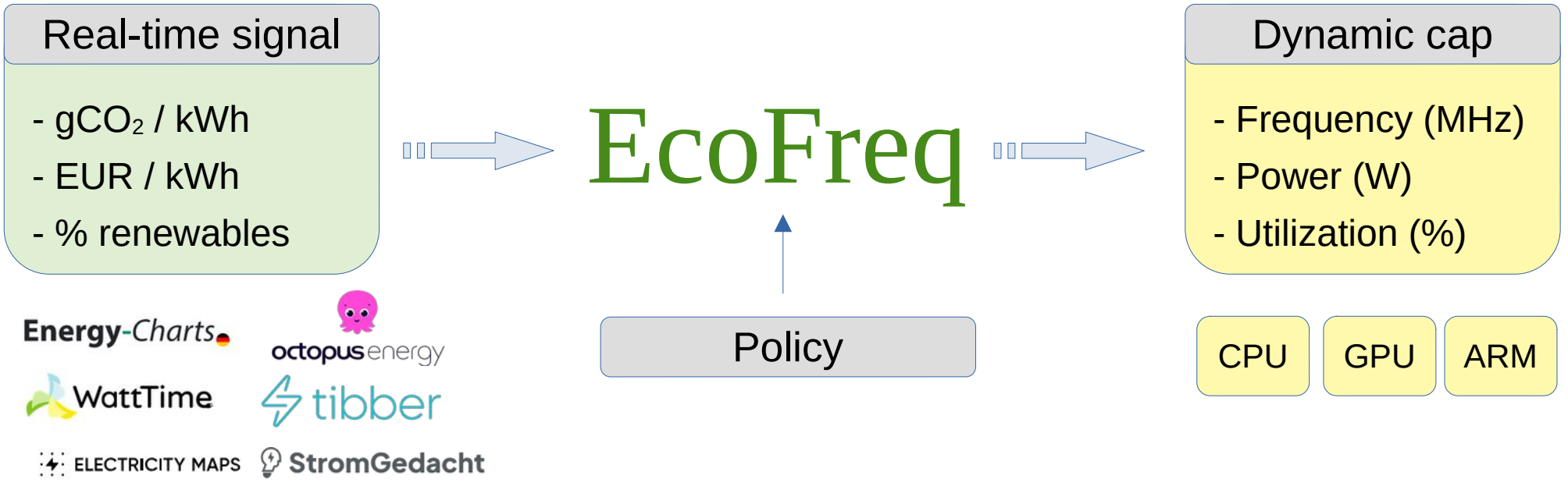


Power scaling methods

Hardware support		CPU			GPU	
Method	Granularity	x86 Intel	x86 AMD	ARM	NVIDIA	AMD
DVFS	core / chip	+	+	+	+	+
Power cap	chip	+	+/-		+	+
Utilization cap	process	+	+	+		

- No modification to workload / submission scripts

Proof-of-concept: EcoFreq



- ✅ Zero workload modification
- ✅ Low latency (seconds)
- ✅ Broad hardware support

Scaling policy

- Based on relative carbon intensity

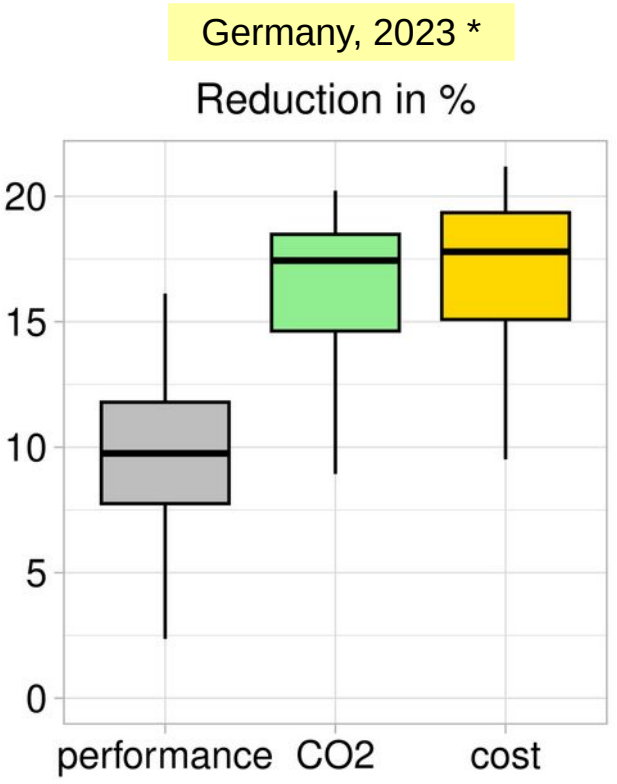
$$P_{lim}(CI) = \begin{cases} 100\%\ \text{TDP} & \text{if } 0\% < CI <= 33\% \\ 80\%\ \text{TDP} & \text{if } 33\% < CI <= 66\% \\ 60\%\ \text{TDP} & \text{if } 66\% < CI <= 100\% \end{cases}$$

- Based on discrete “traffic light” signal

$$P_{lim}(CI) = \begin{cases} 100\%\ \text{TDP} & \text{if } CI \text{ in } [\text{'low'}, \text{'very low'}] \\ 80\%\ \text{TDP} & \text{if } CI \text{ in } [\text{'moderate'}] \\ 60\%\ \text{TDP} & \text{if } CI \text{ in } [\text{'high'}, \text{'very high'}] \end{cases}$$

EcoFreq: evaluation

- Over-proportional savings
  - 15-18% lower CO_2 & electricity cost
  - @ 10% throughput loss
  - Adjustable via scaling policy
- Presented at ISC'24
  - Paper: https://doi.org/10.23919/ISC.2024.10528928
  - Code: https://github.com/amkozlov/eco-freq



(Kozlov & Stamatakis, ISC 2024)

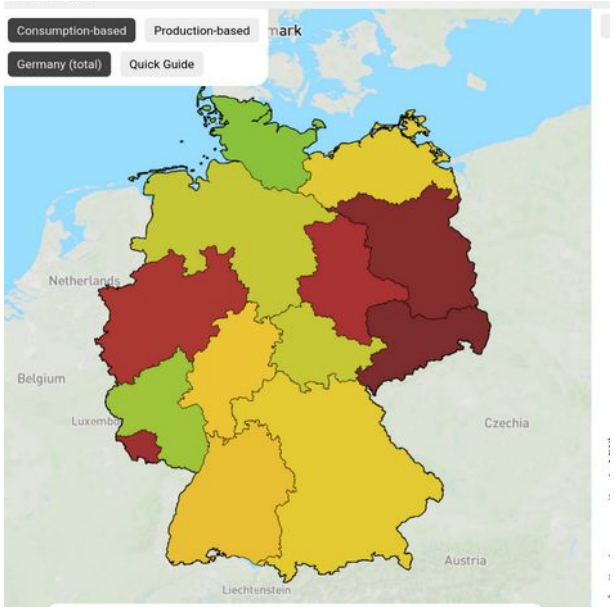
* electricity day-ahead market data; 3-step scaling policy (100%/80%/60%); 14 HPC workloads, Xeon Cascade Lake

Infrastructure & other

Datacenter location

- Major differences even within the same country

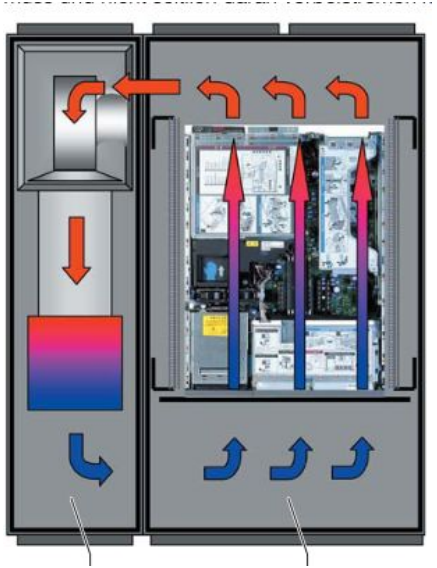
https://carbonintensity.org.uk/



https://co2map.de/

Cooling & heat reuse @ HITS

- Hybrid air / cold water cooling (~15°C)
- Side rack fans (LCP) + room A/C units
- Water outflow T < 20°C
  - Winter: too low for direct heating (>40°C)
  - Summer: too low for free cooling



- Alternative: direct water cooling

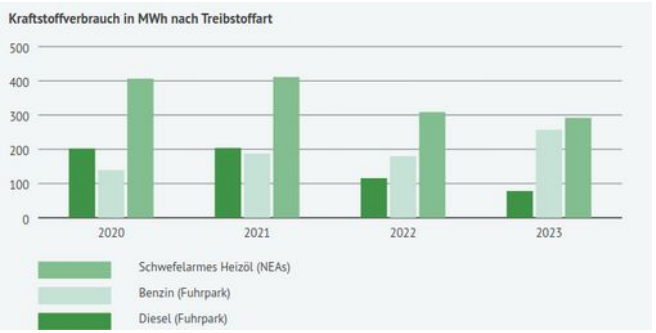
UPS

- Uninterrupted power supply: losses up to 10%
  - esp. older models at low utilization
  - right-sizing, eco-mode, upgrade to LFP

Efficiency AC-AC up to (at cosphi 1.0) (depending on Module power)	%	*Load* : *100 %* *75 %* *50%* *25%* DPA 30-50 : 94.5% 94.5% 94% 93% DPA 10-25 : 94% 94% 93.5% 91%
Efficiency with Linear Load at cosφ =0.8ind Efficiency Non-linear Load (EN 62040-1-1:2003)		Typically up to 1 % higher of above values Typically up to 1 % lower of above values
Eco-Mode efficiency at 100% load	%	98 %

Backup generators

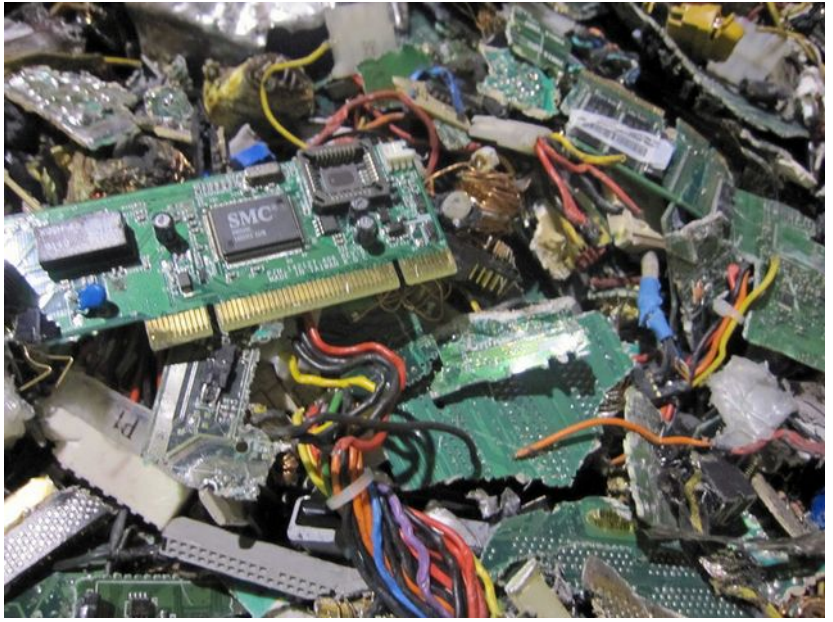
- Datacenter’s “dirty secret”
  - Burn diesel / heating oil → pollution
  - Require regular test runs → emissions
- Why?
  - Cheap & users want 100% uptime
- Alternatives
  - Batteries, fuel cells, <100% uptime



Umwelterklärung 2024 – TelemaxX

Embodied carbon / E-waste

- Solar panels: >30 years
- Wind turbines: >20 years
- Batteries (Li): >10 years
- HPC hardware: 3-7 years
  - Second life?
  - Recycling?



https://technofaq.org

Further reading

- Green Software for Practitioners (online course)
  https://training.linuxfoundation.org/training/green-software-for-practitioners-lfc131/
- Resource-aware research on Universe and Matter: call-to-action in digital transformation
  Bruers et al. (2024)

  https://link.springer.com/article/10.1140/epjs/s11734-024-01436-4
- Energy-aware operation of HPC systems in Germany
  Suarez et al. (2024)
  http://export.arxiv.org/pdf/2411.16204

Take-home messages

- Familiarize yourself with energy / electricity
  - measured in kWh (for all practical purposes)
- Look for absolute consumption / emissions
  - Not hypothetical “savings”
- Don’t wait for management or IT
  - Find motivated colleagues, build horizontal links