Measures for sustainable computing operation

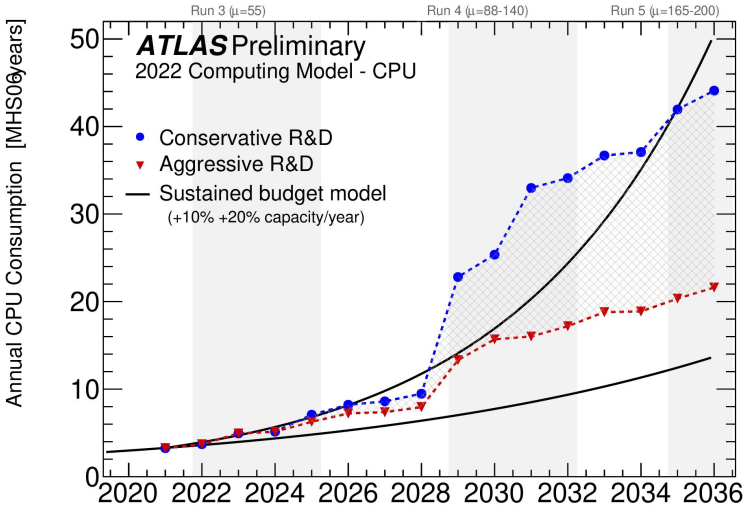
Rod Walker, LMU 31st May`23

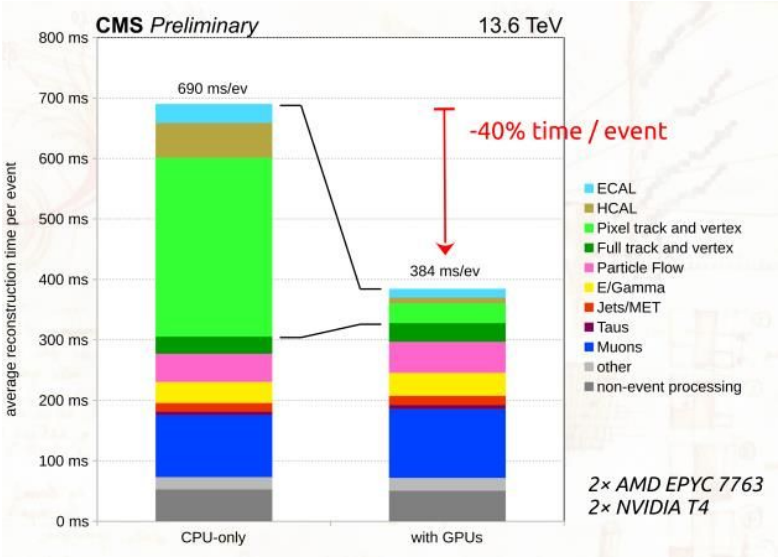
Introduction

- Working on ATLAS distributed computing, and T2 grid admin
- Sustainability activity prompted by 2022 energy crisis
- Sustainability was a hot topic at the recent CHEP
  - Nuclear and Astronomy well represented too.
- Topics:
  - Efficiency of software and datacenter
  - Electricity mix and flexible demand
  - Storage

“Reduce, Re-use, Recycle”

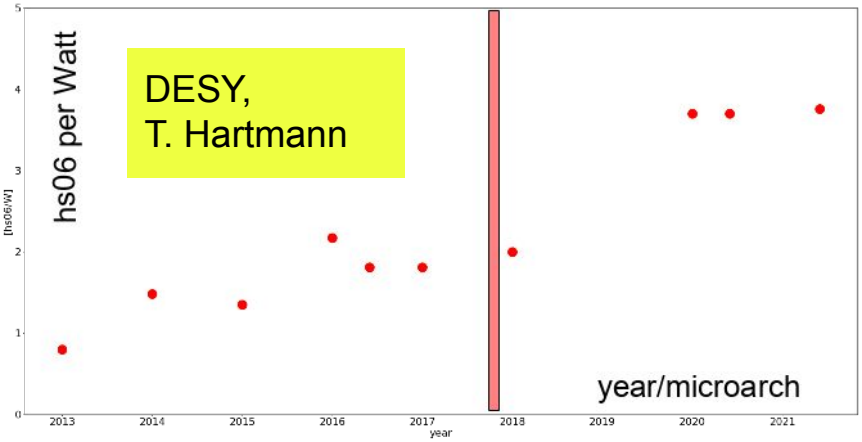
- Reduce
  - run what we need and no more
  - improve SW speed evt/HS06
    - algorithm development
    - more FastSim, with ML
  - Offloading to GPU - CMS
  - Columnar Analysis - save cpu
  - avoid duplication to reduce storage
- Re-use
  - shared LHE generator inputs
- Recycle
  - ?

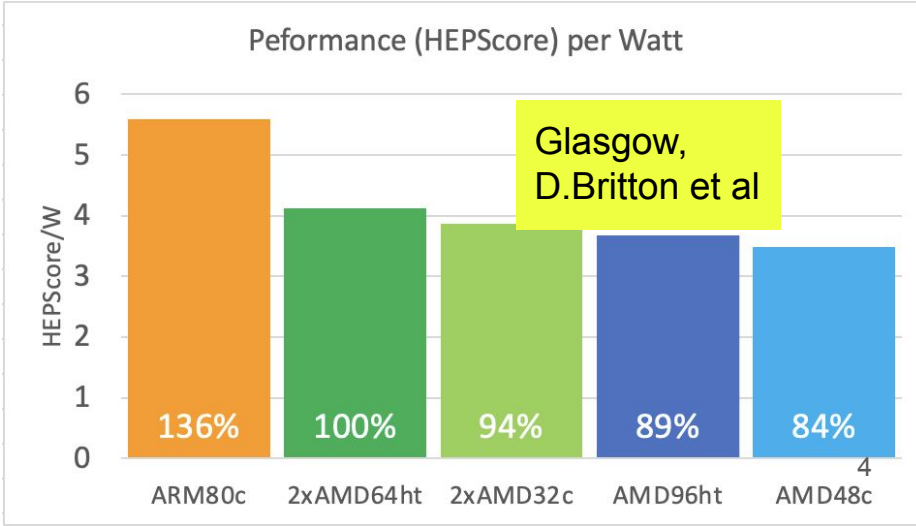




Datacenter efficiency

- If “Reduce” is maximizing evt/HS06s then here we minimize kWh/HS06s
  - retire older, less efficient hardware
  - ARM using 40% less power per HS06
- PUE: CERN Meyrin 1.45, Prevessin 1.1. 30% less energy but limit at 1.
- Waste heat: building and district heating
  - QMUL with heat pump, KIT, …





Datacenter location

- 20 years ago every HEP group had a grid compute cluster
  - often now moved to University computer center
- Were good reasons: local money, free room/electricity/manpower. Educating generation of physicists in IT & Distributed computing. Competition of ideas.
- Also big downsides
  - increased operation complexity, lost economies of scale, single idle VO
  - knowledge of no value to physicists: datacenter design, HW procurement, low-level operation
- Cloud middleware & fast networks reduce the need for in-house compute
  - full control over **interesting** aspects, inc. OS, batch system and interactive access
- Then free to choose a single, optimal location
  - growing DE renewables will give low carbon intensity **most** of the time
  - can use waste heat
  - leverage existing infrastructure and expertise
    - be on the right side of distribution network congestion

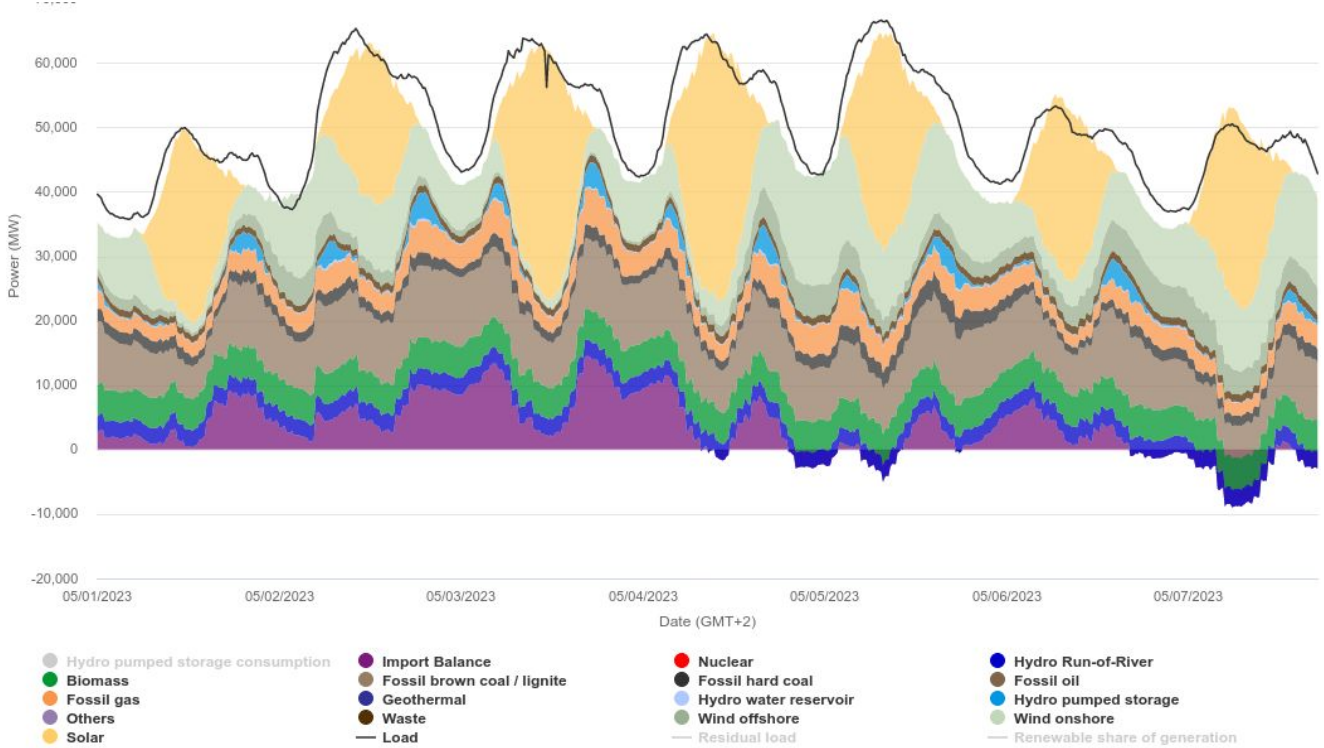
German Science Cloud

- Commercial clouds very usable but expensive
  - not aimed at HTC: storage too good, services rather then 100% cpu usage, PB egress
- Like Gauss(HPC) but for HTC & services
- Quotas for institutes, LS, experiments, including rolling share+opportunistic
  - why would BMBF split funds N ways to get less compute?
  - local funds can always buy more quota
- Avoid fragmentation of resources, with multiple points of failure
  - important for ‘new’ architectures GPU and ARM.

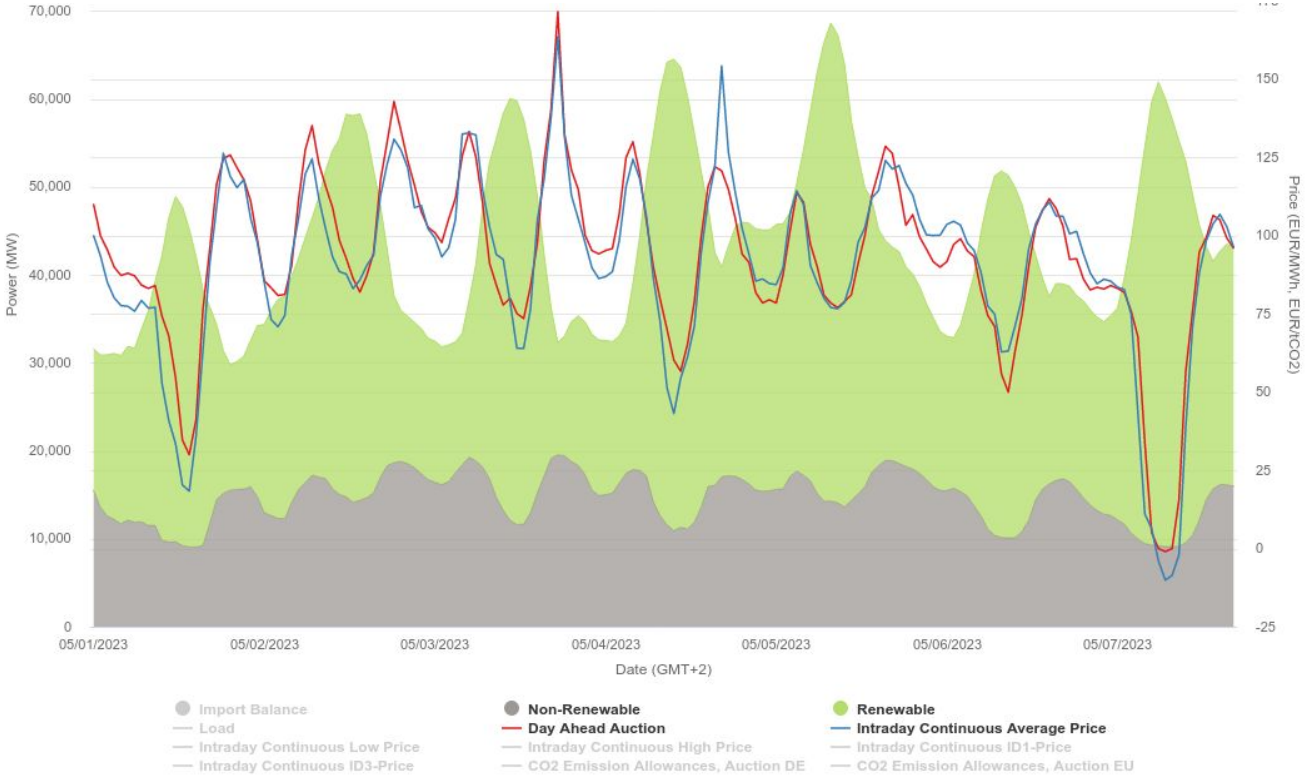
Reduced compute, efficient SW on efficient HW at an efficient Datacenter - but we still need electricity …

German electricity demand and generation mix

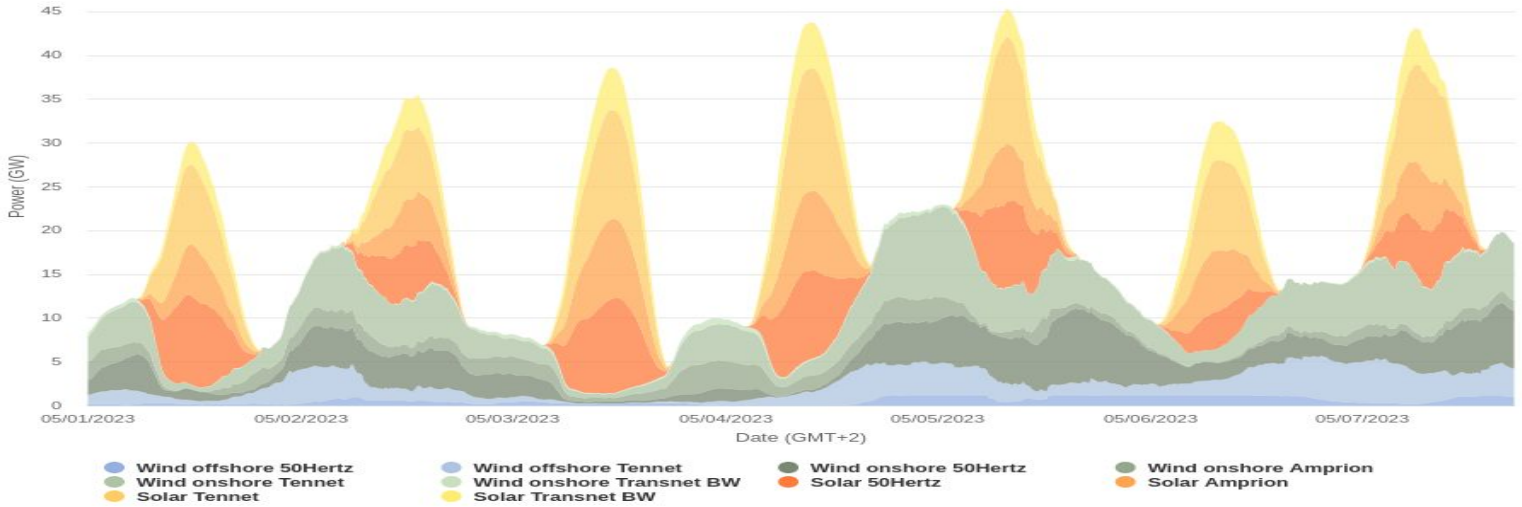
https://energy-charts.info/index.html



Follow the money …



Price and gC02/kWh vary.
Cheaper when more renewables.
It matters **when** the electricity is consumed.



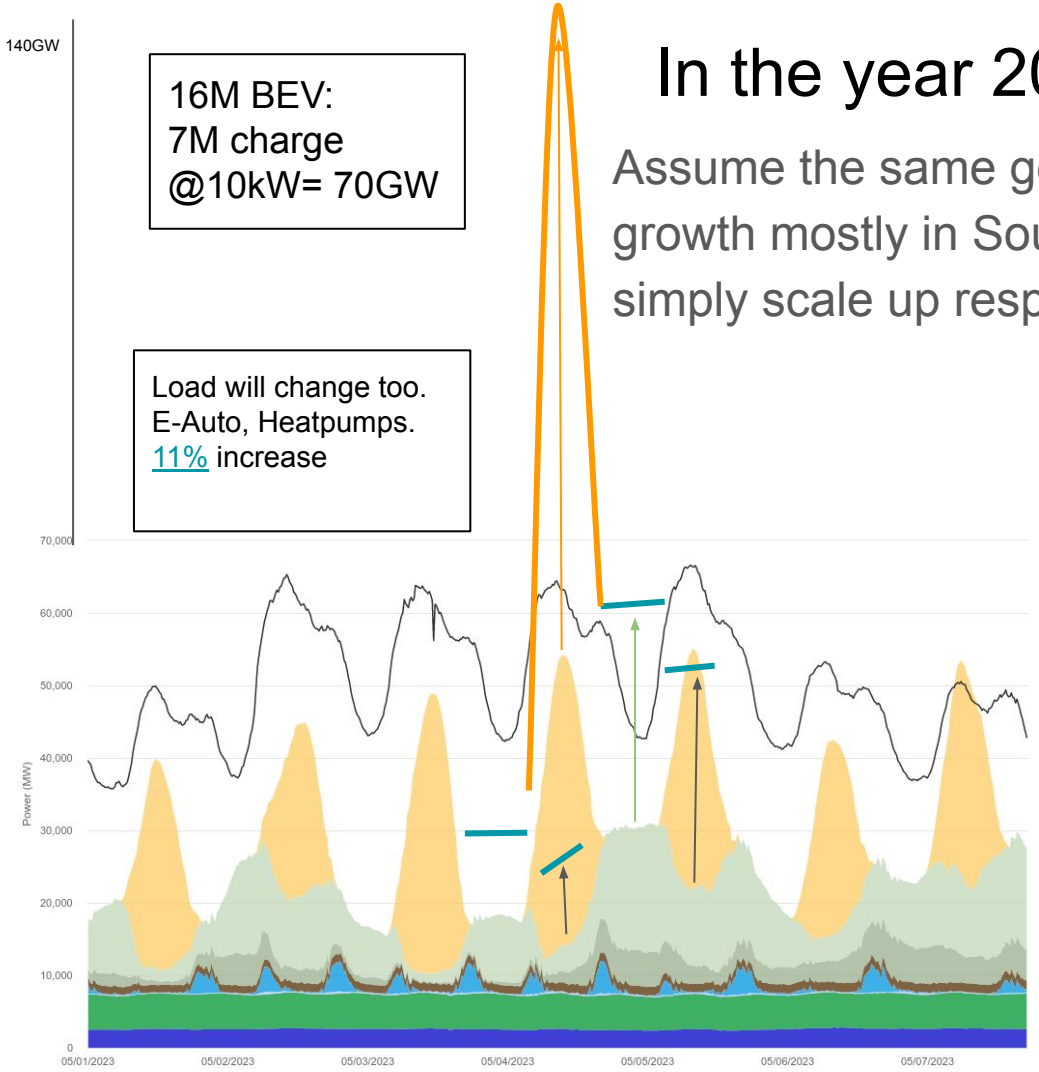
Solar & wind

In CERN Elastic Search for all WLCG countries/regions - Ben Brüers

Carbon intensity DE

In the year 2030, if we are still here ….

Assume the same geographical distribution(onshore growth mostly in South?) and weather(!), then these simply scale up respective contributions.



Capacity	2022 (GW)	2030 (GW)	Factor
Offshore Wind	7.8	30	4
Onshore Wind	56	115	2
Solar	66	215	3

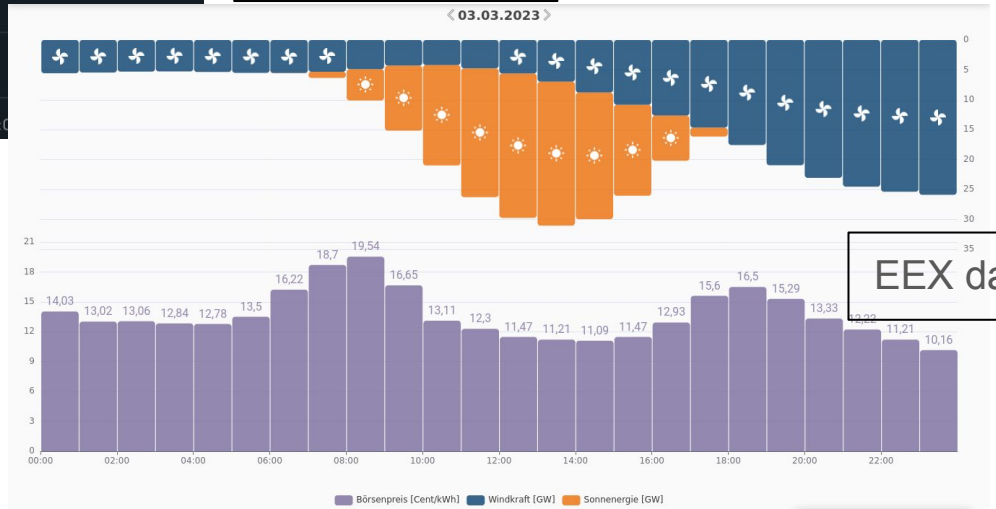
https://www.bmwk.de/Redaktion/DE/Dossier/erneuerbare-energien.html

Variable electricity tariff

Tibber, Awattar consumer tariffs but also available for business

All companies must offer one by 2025

Fixed price/kWh



EEX day ahead price

Save money by reducing power at peaks.
Save gCO2 when pricing politics catches up: higher carbon price/tonne

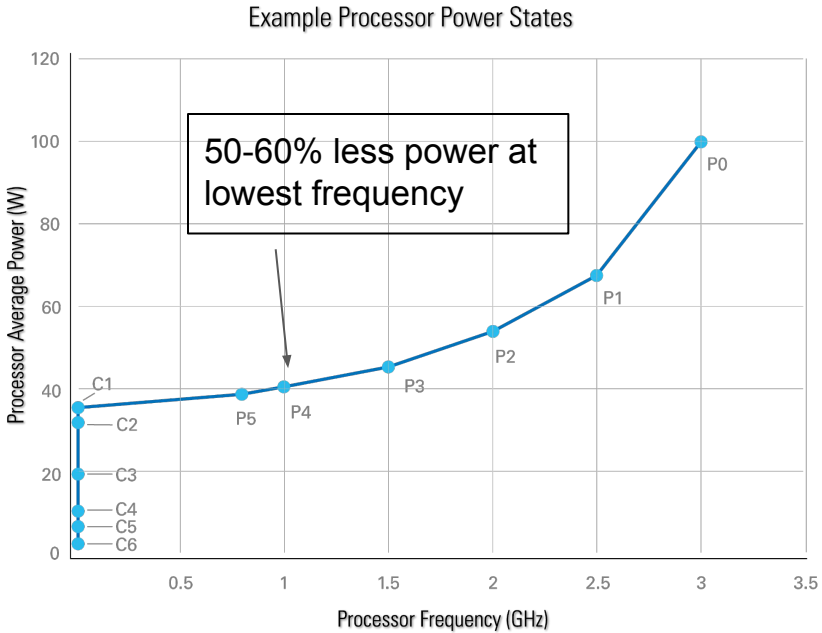
Can Datacenter modulate power consumption?

- Mostly HTC where a few hours or days delay is irrelevant.
- Obvious saving is from turning off compute nodes
  - ok for longer predictable pauses, e.g. infamous Dunkelflaute
    - but lengthy draining of long jobs, without checkpointing
  - twice per day is unfeasible, also due to HW strain of power cycling?
- Can we reduce power without draining jobs
  - freeze processes to let CPU sleep
    - large drop in node power, but base usage still there for no work done
  - reduce CPU frequency to minimum, or sweet-spot
    - smaller drop in node power, ~50%, but still doing work
    - does it pay off?
  - switch to battery
    - traditional UPS probably expensive, but solar battery systems prices tumble
      - battery/inverter costs with 6000 cycles gives ~ 10ct/kWh stored and returned

CPU frequency modulation

Free, fast, repeatable, harmless to workloads
Set CPU governor to PowerSave

$P \sim f V^2$



dynamic voltage and frequency scaling (DVFS)

voltage reduces with frequency

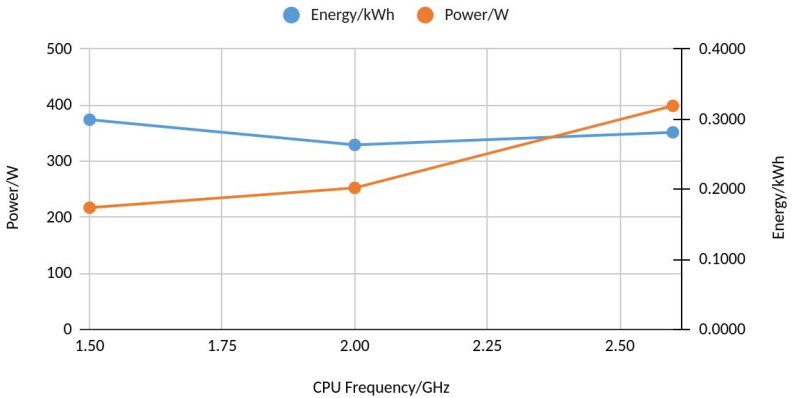
Useful work ~ frequency, but power falls faster than frequency

Could offset base/non-CPU node power consumption

Real-world measurements: HEP work vs total node power

1) Same work at different frequencies: 1000evt GEANT4 simulation

dual-x86
E.Simili, Glasgow



- **HEP work per kWh not significantly less at lowest frequency**
  - Glasgow 6% & DESY 2%
- Middle frequency best for both!
  - fewer voltage steps?
  - highest frequency at lowest V

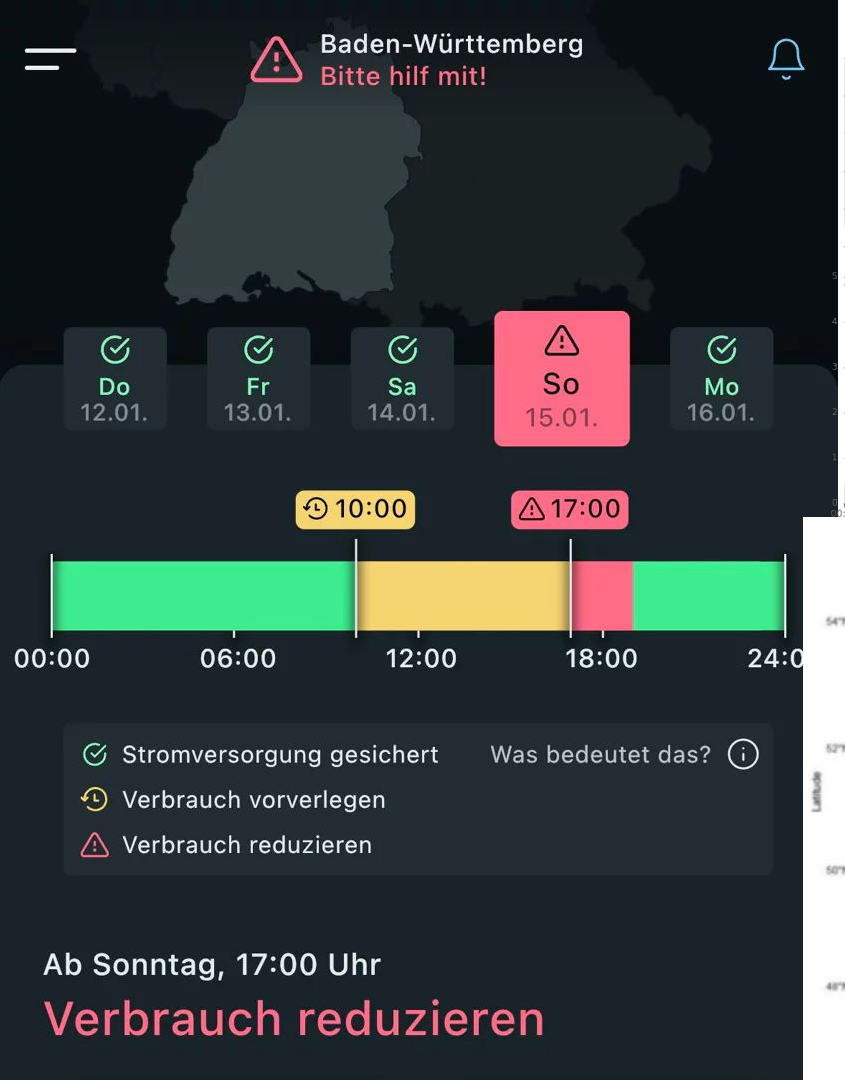
Frequency/GHz	HS06	Power/W	HS06/GHz	HS06/W	Ratio to high
1.5	1085	286	723	3.79	98%
2.15	1424	330	662	4.32	111%
2.85	2032	524	713	3.88	100%

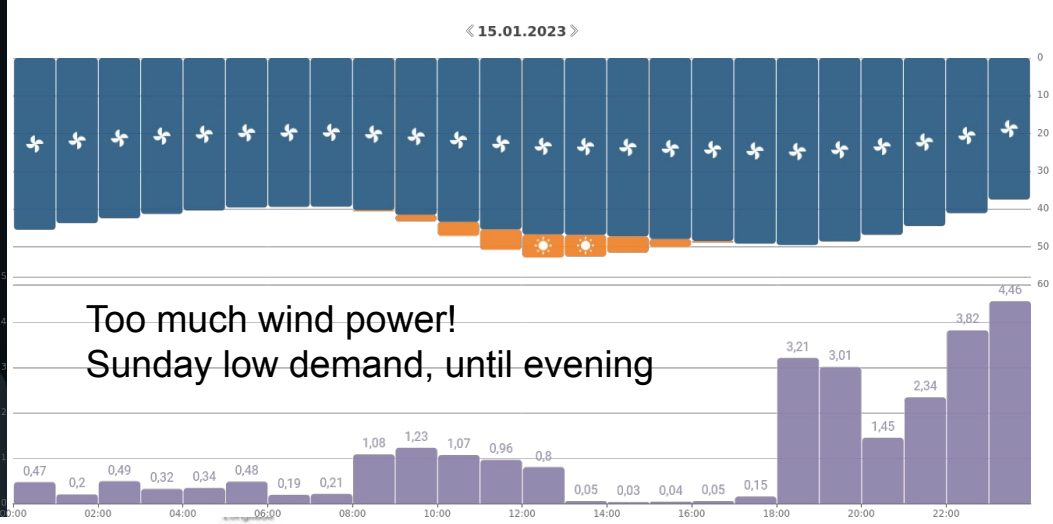
2) AMD node HEPSpec vs f
T.Hartmann, DESY

Embedded Carbon and Second-life

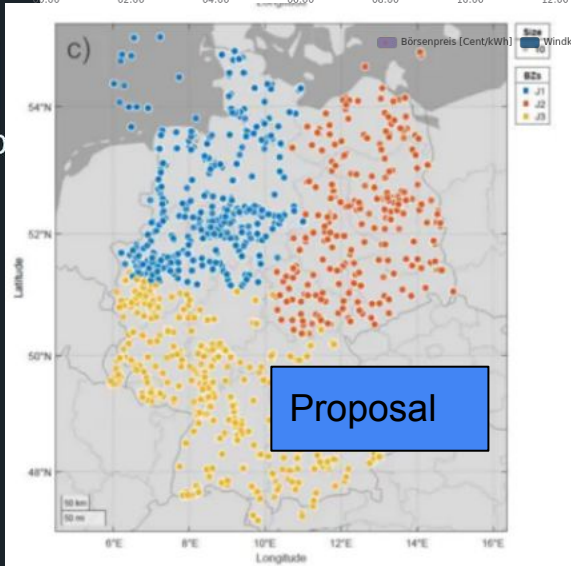
- Some estimates of only 50% carbon footprint due to electricity in server operation.
  - rest is ‘embedded’ carbon due to manufacture
- As gCO2/kWh electricity reduces, the embedded part grows proportionally
  - means we should use HW as long as possible
- But older HW less efficient in HS06/W
  - does not matter when electricity is cheap and low carbon
  - variable tariff can extend life of old hardware
  - don`t need to run new HW 24*7 to reach pledge

	kHS06	Hours	Avg Hrs/day	HS06 Mhrs	HS06/W	MWh
Pledge	1000	8760	24	8760	4	2190
New HW F_hi	1000	7300	20	7300	4	1825
New HW F_Lo	500	1460	4	730	4	183
New HW Hi+Lo	917	8760	24	8030	4	2008
Old HW	800	912.5	2.5	730	2	365
Old+New				8760		2373





Too much wind power!
Sunday low demand, until evening

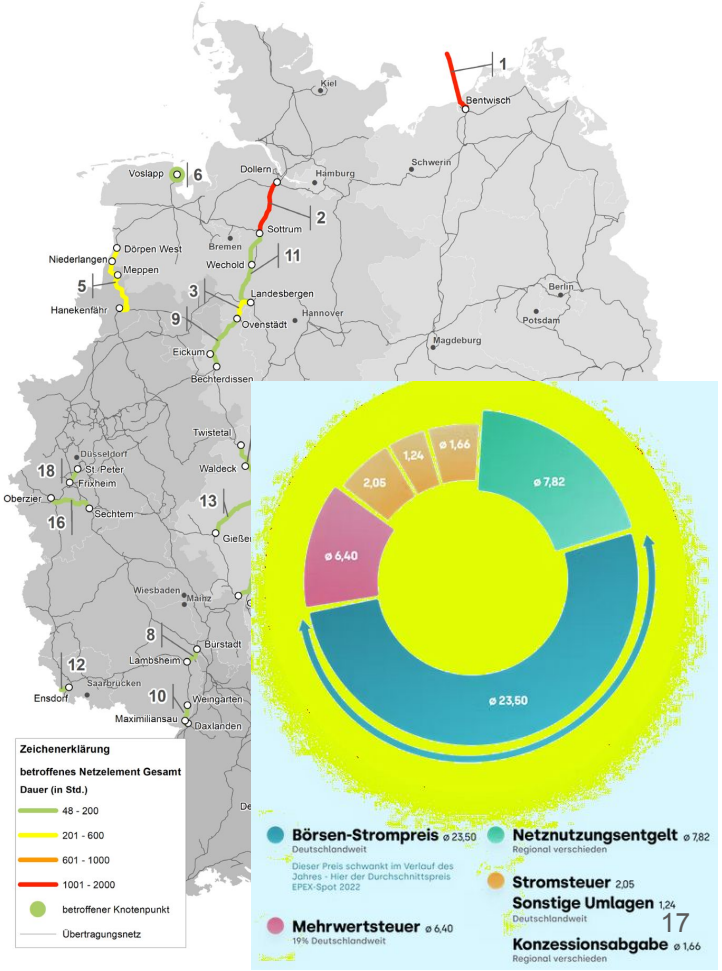


Fossil power stations reduce output due to low price(for all DE).

Northern wind energy cannot reach South due to missing power lines.

More variable: Stronger price signal

- Vary Netznutzungsentgelt to avoid congestion
  - several studies and groups suggesting this
  - Verbraucherzentral, FfE, Agora,
- Stromsteuer, Abgaben too
  - make it overall revenue neutral, then who cares?
- Fine tune for network congestion at all levels, e.g. medium-low voltage transformer
  - street full of EVs should not all charge at once
    - VNB wants to control it. Price signal better.
- 5.8TWh abgeregelt in 2022, cost 820M€
- Sometimes the price/kWh will be really zero



What about Storage?

- Around 40% of T2 energy used for storage
- HEP uses tape for more than just archive
  - Data Carousel controls pre-stage, optimized write/read
  - other communities? Joint dev area.
- T2 has ~10GB/s read/write, local+remote
  - 3PB RAID capable of 3 times this
- If 2PB is 90% spun down
  - factor 10 less energy. Latency similar to tape
  - no robot, no winding, variable #'drives'
- Complete datasets on single disk
  - schedule BringOnline just like for tape
  - dev on QoS in e.g. dCache
- Big benefits in cost and energy

Standard T2 disk

RAID 6, 12*10TB disk(€200), 100TB usable.
Server €10000 Euro
Bandwidth: 10 * 1GB/s
1PB 10*(10000+12*200) = 124k€
Power: 10*(12*10+200)*8000/1000 = 25,600kWh/a

JBOD with spin-down

Server €5000. 100*10TB disk(€200), 10 active
Bandwidth: 1GB/s
1PB: 5000+100*200 = 25k€
Power: (10*10+200)*8000/1000 = 2400kWh/a

TAPE

€8/TB, €5000/drive, Server(€10000)
Bandwidth: 1GB/s with 3 drives @ 300MB/s
1PB: 8000 = 8k€, some drives and servers effectively dedicated. Robot.
Power: 40W/drive (200+3*40)*8000 = 2560kWh/a

Input from Computing Operations

- SW development for efficiency, new architectures, checkpointing, storage
- Establishing Science Cloud is optimal use of taxpayers money for computing
  - economies of scale: manpower, purchasing, operations, …
  - important to maintain ownership, control, flexibility, e.g. IaaS, steering committee
- Still need electricity, and location is important
  - luckily good locations exist in DE
- Leverage HTC flexibility to reduce carbon footprint and energy cost
  - CPU frequency reduction viable since no reduction in work/kWh
  - run old, less efficient hardware(2nd life) only when energy low carbon and cheap
- Vary whole electricity price
  - lobby work from others, but maybe a pilot project