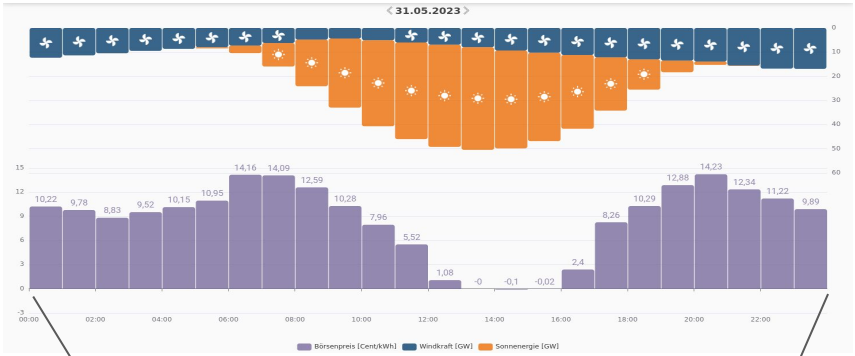
Power Modulation of a Compute Cluster

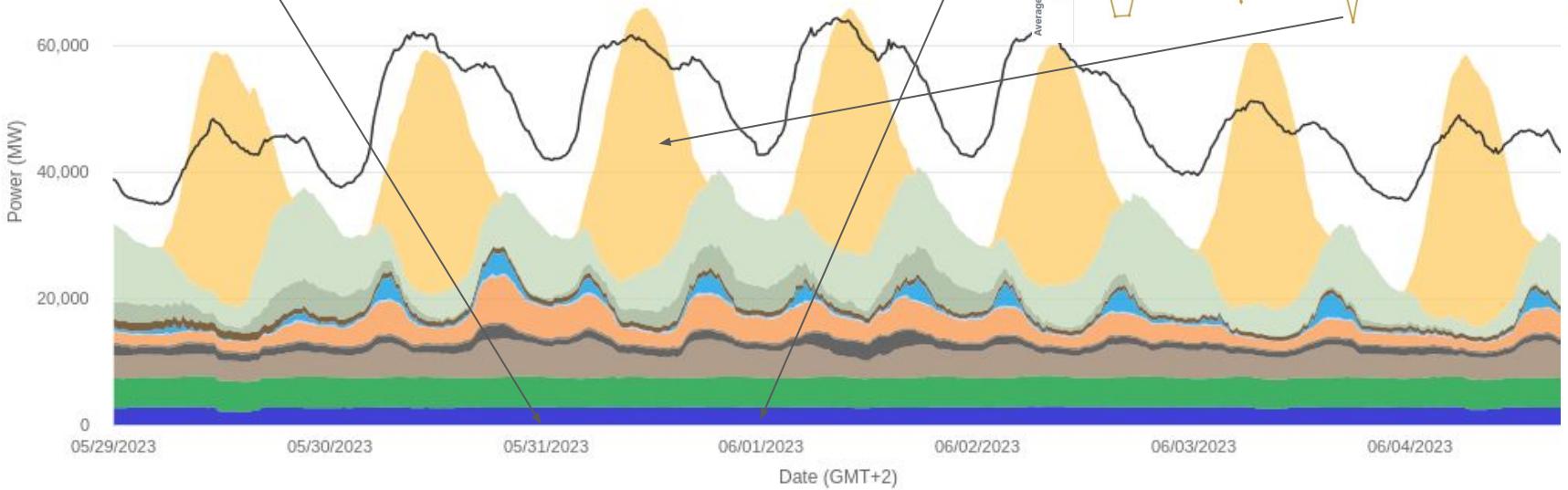
Rod Walker, LMU 3rd Aug `23

Topics

- Motivation for modulating the power consumption
  - In Germany
  - In Cuba
- Methods to modulate power
  - Power-off, CPU frequency throttling & Battery
- Embedded carbon
- Goals and possible plan

DE generation Week 22





Hydro pumped storage consumption
Hydro Run-of-River
Fossil hard coal
Geothermal
Others
Wind onshore
Import Balance
Biomass
Fossil oil
Hydro water reservoir
Waste
Solar
Nuclear
Fossil brown coal / lignite
Fossil gas
Hydro pumped storage
Wind offshore
Load

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

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

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

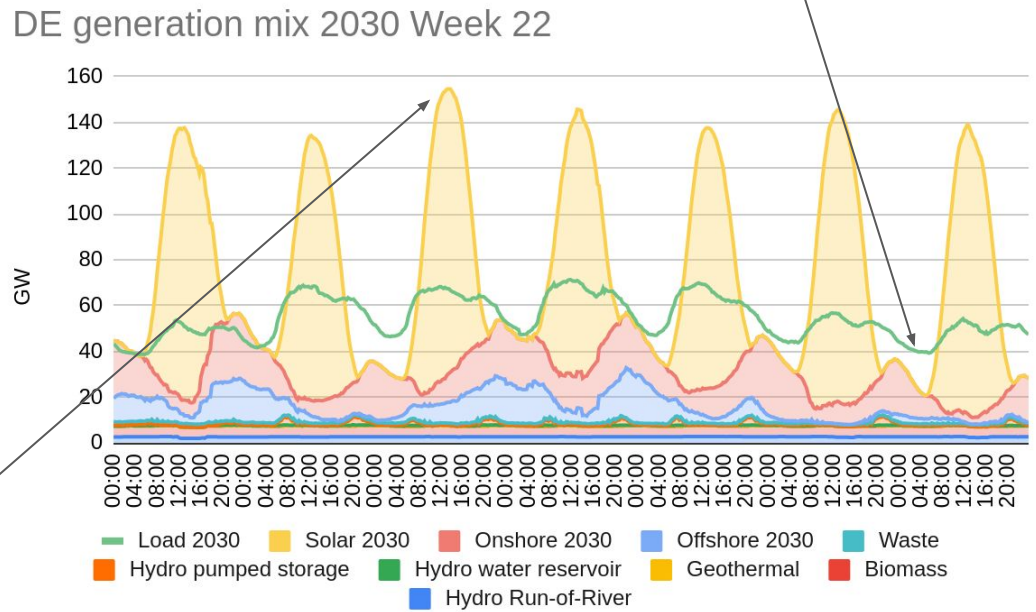
https://www.bmwk.de/Redaktion/DE/Dossier/erneuerbare-energien.html

Assume the same geographical distribution and weather(!), then these simply **scale up** the respective contributions.

Load will change too.
E-Auto, Heatpumps.
11% increase

16M BEV:
7M charge
@10kW= 70GW

Still periods needing gas generation, unless we can reduce the load.



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 preemption
  - 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(PDU,RAM,..) 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

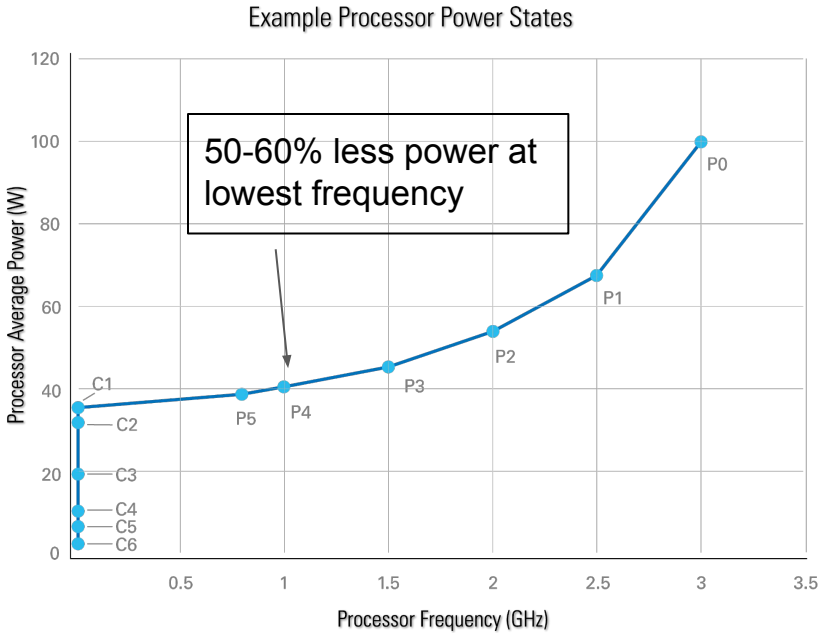
Preemption: Types of checkpointing

- Have ATLAS EventService, but only for sim
  - Output stored externally so preempt loses only in-flight events.
- Snapshot application memory to file , store state of file descriptors, network connections, etc
  - dmtcp integrated into apptainer, but not tested for pilot/athena
  - need to keep workdir, and restart on same WN - lots of dev, also at sites
- Signal job to end cleanly. Pilot stores partial output files in rucio
  - Athena needs to handle signal, drop in-flight events, close files
  - Panda pulls partial files on next attempt
  - TRF checks what is done and continues
    - would also work without storing out - keep workdir and restart job.
- No general checkpointing available now.

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

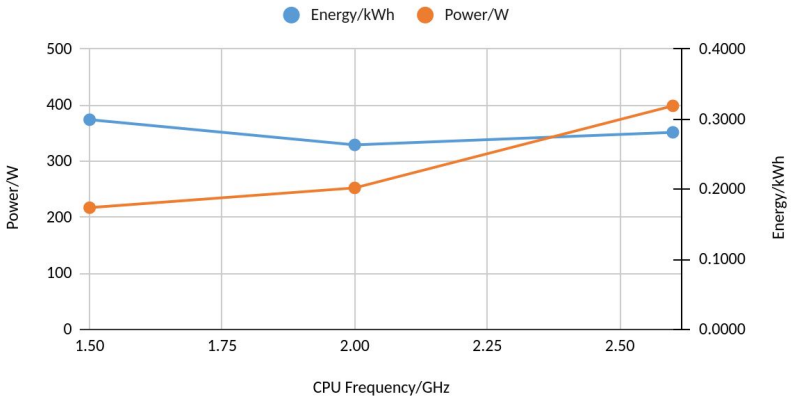
Useful work ~ frequency, but power falls faster than frequency

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

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

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%

Does not apply to ARM

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 low carbon(and cheap)
    - DE - variable tariff can extend life of old hardware
- Relocate old hardware
  - 1kWp solar produces ~1000kWh/a in DE but ~1600kWh/a in Cuba

Cuba: Goal and requirements

- Close to 100% self-sufficient in energy
  - but use mains power when necessary
- Can achieve availability level TBD, e.g.
  - Full compute power 12hrs/day
  - >40% compute power 24hrs/day
  - minimal battery capacity to reach this level
- Resilient to mains power outage - target N hrs
- Smooth mains lumpy power?
  - important function of UPS

Possible plan

- Donate old(5yr) T2 HW - export!!!
  - More HepScore/Watt. Homogeneous. RAM/core~>2GB.
  - Or buy ARM based. Ok for all workloads?
- Maximum number of solar panels(space/cost) and inverter
  - Grid-tied will give higher availabilities for less battery
    - can supply solar power to other in-house consumers
      - also with redundant power from battery
- Battery
  - LiFePO4 with 6000 cycles, ~ 20 years lifetime
- Algorithm and procedure to reduce power consumption when necessary
  - night, clouds, power cut, (price)
  - combination of preempt/drain-power-off, reduce CPU frequency

Summary

- Consumers unable to modulate power will have higher footprint and pay more
  - price fluctuations will increase to enable network stabilization and avoid bottlenecks
- Reduce CPU frequency on all nodes is identical to turning off 50% nodes
  - 50% less power, 50% less work - but way easier to do
- Second-life hardware on 100% renewables can offset work lost modulating
- Battery necessary when grid electricity unstable and to maximise solar usage
- Battery also useful to modulate power-usage to reduce cost/gCO2
  - reduces the need to power-down or modulate CPU frequency
    - algorithm&procedure still common need for DE and Cuba
  - cost-neutral with 10ct/kWh electricity price difference. 3-6% of compute server cost.
    - standard datacenter component, like rack or cooling

Back-up

Battery capacity and cost per server

80 core server, 5000€, 400W

4.8kWh battery, 1250€

2hrs peak on battery -> 0.8kWh

Charger/inverter: 3kW, 900€

7 servers: 2.8kW 1250+900=2150 c.f. 7*5000= 35000€, i.e. 6%

2 cycles/weekday: 6000/(2*52*5) > 10yr battery lifetime

AC vs DC coupled

- DC more efficient
  - 1 fewer AC-DC conversion to charge battery
- Open questions
  - Both ok for blackout protection?
  - charge batteries from mains?