Energy saving measures

DESY, 15th March 2023

Introduction

- Why am I here?
  - DESY big resource for HEP and will grow with current Uni T2->NHR plan
    - makes sense to concentrate compute in the windy North
  - already internal discussion, studies and action on energy/cost savings
  - I want to contribute not irritate.
- Saving alternatives for compute
- Variable electricity tariff
- Brainstorming

Saving alternatives for Compute

- Ignoring HW choice, datacenter optimizations, etc.
- Saving energy results in less compute work done for experiments
  - clear for cpu limited workloads, since energy proportional to cpu seconds
  - lost compute not necessarily as large as saved energy
    - e.g. non-homogeneous cluster: old hardware with less HS06/Watt
- Saved money is proportional to saved energy
  - for a fixed electricity tarif
- Alternatives
  - power down nodes
  - reduce CPU frequency

Power down nodes

- Almost 100% reduction in power
  - few Watts for wake-on-LAN(or whatever), PDU?
- Drain nodes, and/or preempt jobs
  - some wasted cpu due to idle cores, or discarded work
  - mitigation with accurate batch job walltime, short jobs for backfill, checkpointing
    - quite strong requirements for the customers(for all workloads)
  - flexible start time or long notice period
    - so not to schedule long jobs(4 days) to node planned to go down
- 100% loss of work, for nodes powered down
  - delta_work/delta_energy = 1

How can ATLAS help?

- Realistic maxwalltime per job to schedule for draining
  - already have push queues, with job specific requirements
    - mostly aCT/ARC CE but also supports HTCondorCE
    - need safety factor to allow for spread
  - have some short jobs for backfill(analy, build, merge)
    - may need short PandaQueue to avoid blockage by higher prio long jobs.

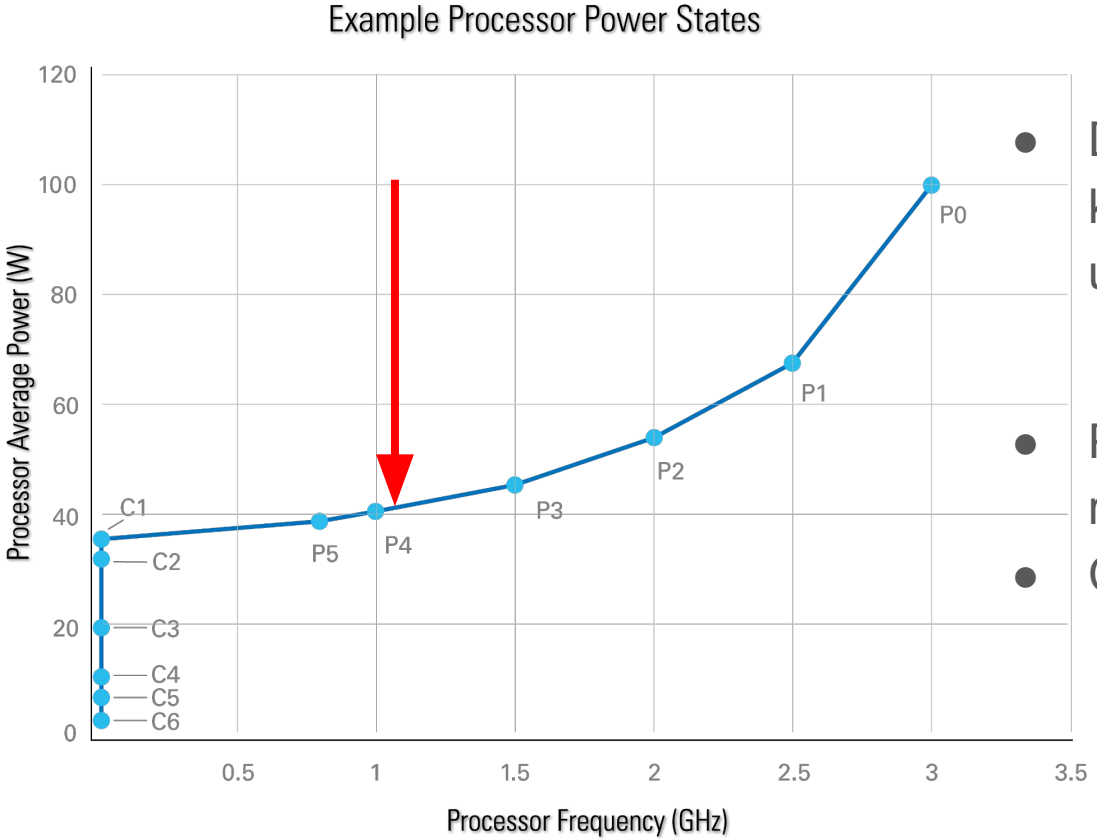
+maxMemory = 8280
+maxWallTime = 4320

- MachineFeatures WN “shutdowntime” checked every 10mins by pilot
  - will vacate node cleanly 10mins before
  - ignore timestamp older than pilot start time(still some left from December)
    - assume batch system does not start jobs on node with shutdowntime within maxwalltime(96hrs)
- Checkpointing all/most workloads is not feasible

Reduce CPU frequency

- Force clock speed to lowest setting to reduce power
  - Thomas's AMD 45% power reduction, but varies per CPU up to 60%.
- Can be set by BIOS/OS control, but simplest is full OS control via governor
  - instant, reversible and repeatable
  - no harm to running jobs: they just run slower
- Base power consumption for non-cpu parts of node: PDU, disk, memory, etc.
  - does not reduce with frequency so loss of work should be more than energy saving - right?
- Not clearcut.
  - work proportional to frequency, worst-case for cpu bound workloads
  - CPU power consumption ~ fV^2 where voltage is reduced with frequency
    - power falls faster than frequency, i.e. faster than work.
  - big question: does this offset the base power? Need to measure it,

Bluffer's guide to CPU power management

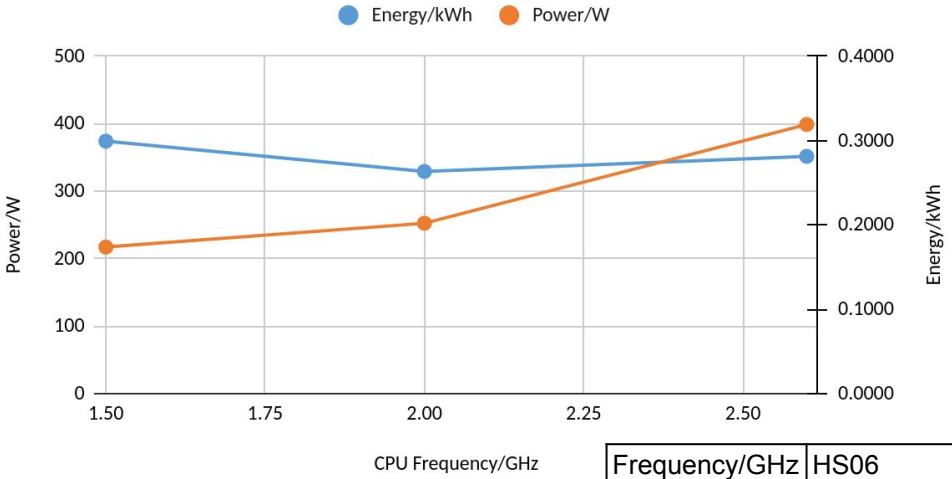


- Designed to save power while keeping performance for bursty usages, e.g. save laptop battery
  - we want to drive down power despite load - jobs still running.
- P: frequency setting, voltage reduces accordingly, **$P \sim f V^2$**
- C: shutting down parts of cpu
  - only happens when idle

Real-world measurements: HEP work vs total node power

dual-x86
E.Simili, Glasgow

Same work:
1000evt G4 sim



- HEP work per kWh not significantly less at lowest frequency
  - Glasgow 6% & Thomas 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%

T2 AMD node HEPSpec
T.Hartmann, DESY

If you buy that ….

- Work lost due to frequency reduction not larger than energy saved
  - despite the base node power, thanks to the V^2 term
  - 45% energy saved for 47% work loss
- Powering down nodes: 100% energy saved for 100% work loss
  - tiny residual power, draining inefficiency relevant for shorter pauses
- Frequency reduction on all nodes ~identical to powering down 45% nodes
  - in terms of work lost, energy saved or work/kWh
- Prague HPC reduced CPU frequency 3.3GHz to 2.1GHz
  - with nice study https://docs.it4i.cz/general/energy/

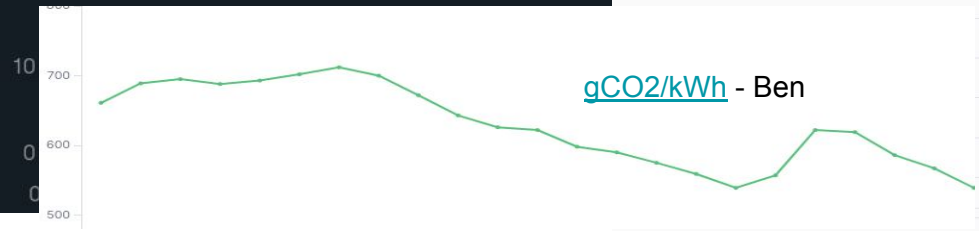
In practice

- Huge advantage is the flexibility to have frequent short reductions
  - easily automated, instant, harmless to jobs and hardware
- non-homogeneous cluster, start with worst HS06/kWh in both cases
- depending on composition and targeted saving, maybe mix of power down and frequency reduction is optimal
  - e.g. can’t save more than 45%(-60%) with frequency alone.
- sweet spot of frequency vs power might not be lowest
  - fewer voltage steps than frequency
- High-throughput users not sensitive to slow down
  - unlucky long job, mostly during reduction and already close to BS walltime limit
  - dynamically increase limit in HTCondor? Frequency term in PERIODIC_REMOVE expression.
- Reduce capacity over holidays, w/e or overnight
  - least interactive user demand but energy is cheapest and greenest, for same reason
  - sustainability(gCO2) benefit comes with variable electricity tarif

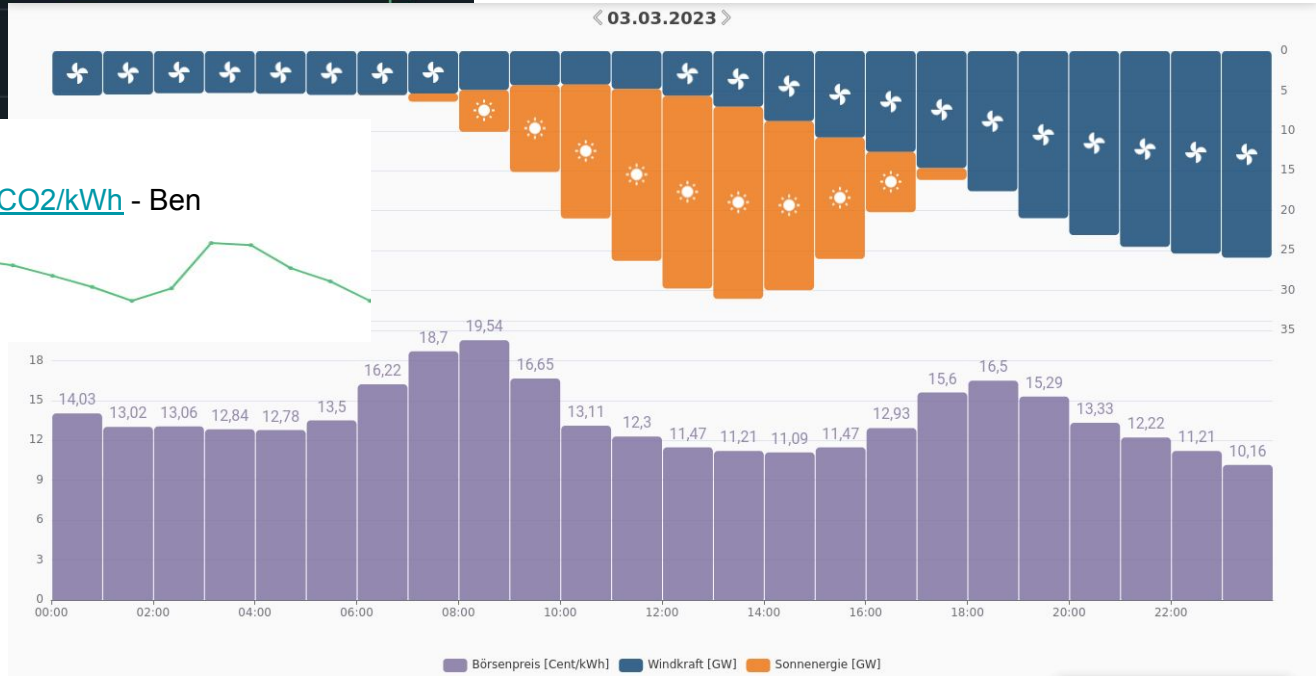
Variable electricity tariff

EEX day ahead price, plus base

Tibber, Awattar but also business tarifs



gCO2/kWh - Ben

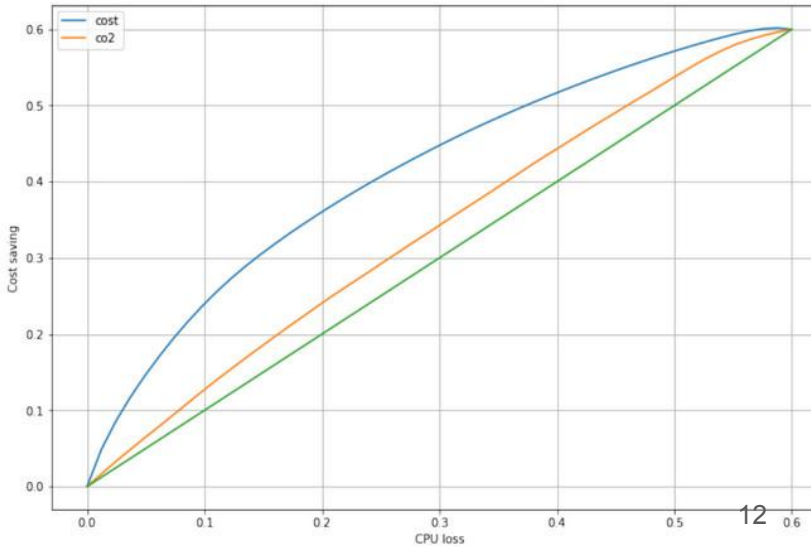


Save more money than work lost by reducing power at peaks.

What's in it for us?

G.Duckeck: Using 2021 DE spot price history, scheduled frequency pauses
60% reduction of energy and work
Plot the cost saving versus lost work.
Plotted gCO2 estimate too - scheduled on cost.

- Can reduce cost by 25% when losing only 10% of the work
  - N.B. total electricity price includes more than spot
- Ideally energy price fully correlated with carbon intensity/gas generation
  - saving money is an easier sell
- DESY-HH saved 35MWh over Xmas
  - 10000 Euros for 17days(~5% of year)
    - drained and turned **off** older nodes
  - same saving with frequency reduction at peaks
    - 1-2hrs/day over year, with flat price
  - variable, 5% cpu loss=15% cost saving



Renegotiate energy contract - variable tariff

- No need to go “all in”
  - “Beim Spotmarktmodell können Sie flexibel Ihre Energiebeschaffung festlegen. Beispielsweise kann ein Teil des Energieverbrauchs als Fest- oder Tranchenprodukt am Terminmarkt eingekauft werden, wohingegen der andere Teil als Tranchenkauf am Spotmarkt beschafft wird.” - Stadtwerk Magdeburg
  - buy most energy at fixed price, and part that can be modulated at spot price.
  - variable cheaper even without any load-shedding
    - about who carries the risk for price increase, and that is not for free
    - when it goes crazy, the government steps in anyway, e.g. 40ct cap.
- Maybe DESY contracted for N years, but worth asking
  - SWM not averse to publicity for currently high profile themes
    - **Digitalisierung der Energiewende – Smart Meter**
- Conclusion: get variable tariff and schedule frequency reductions

Brainstorming section

- Opportunistic NAF
- Storage?
- Second life, satellite datacenter
- Funding

Opportunistic NAF

- We use spare cycles on several opportunistic resources
  - mostly quite awkward HPC with limited workloads, e.g. whole nodes only, G4
    - some with preemption
  - seems odd to have NAF nodes idle or off, rather than use opportunistically
    - perfect for any workload, direct access to RSE
- Low interactive usage corresponds to low energy price
  - weekends, overnight, holidays have lower demand and hence price
  - can reduce T2 usage during high price periods, to get same ATLAS work output
- Concern was always the latencies introduced for users on NAF
  - powering on a node is certainly faster than draining production jobs
  - choose target #cores freed per hour, use job mix and possibly preemption to achieve this.
  - combined NAF/T2 gives more flexibility. Next free slot on T2 can run NAF job.

What about Storage?

- Around 40% of T2 energy used for storage
- HH see ~10GB/s read/write, local+remote
  - 3PB RAID capable of 3 times this
- 2PB is 90% spun down
  - factor 10 less energy and latency similar to tape
  - no robot, no winding, variable #'drives'
- Complete datasets on single disk
  - schedule BringOnline just like for tape
- Needs careful planning
  - disk failure probably whole dataset gone
    - can reproduce data
- 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

Second-life

- embedded CO2 significant part of life-cycle, c.f. energy to run
  - where power is ‘free’ collect and run retired compute nodes
  - more peak capacity gives flexibility to load-shed and maintain throughput(pledge)
- Extreme load-shedding perhaps running only few hrs per day.
- No machine room required
  - dry, ventilated, clean with power and network
  - Lancium use ‘hen hut’ with filter walls & no cooling.
- Probably no physical benefit to being in a barn in Schleswig-Holstein
  - assuming no electricity network congestion North of HH
  - financial benefit, <10ct/kWh due to Netzentgelt etc, could be smoothed over with funding
    - pending more helpful tariff structure, i.e. dynamic NNE/Abgabe/Steuer
    - Bundesnetzagentur 2015 position(p66) did not age well
      - need smart meters, would work too well, cannot be done with fax machines!

Funding

- BMBF sustainability goals
  - focus on green technology, rather than CO2 reduction but some possibilities for us
- Manpower
  - development: scheduling SW, spin-down dcache pseudo-tape and QoS
  - operation, e.g. load-shedding, spin-down storage
- Hardware
  - second-life peripheries , spin-down storage prototype
- Energy tariff compensation
  - fully variable tariff, without fixed base price, clearly optimum to encourage load-shedding
    - fixed parts redistributed but add up to same, for H0 or flat profile.
  - schedule as if we had one, and funding compensates the difference
    - study with partner(SWM, Fraunhofer, …)