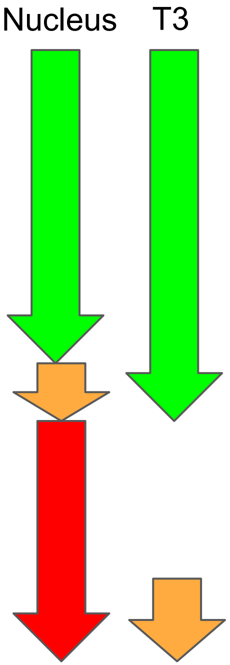
Voluntary load-shedding during peaks

Rod Walker,LMU

WLCG 9th Nov, 22

Flat reduction of energy bill: ATLAS preferred order

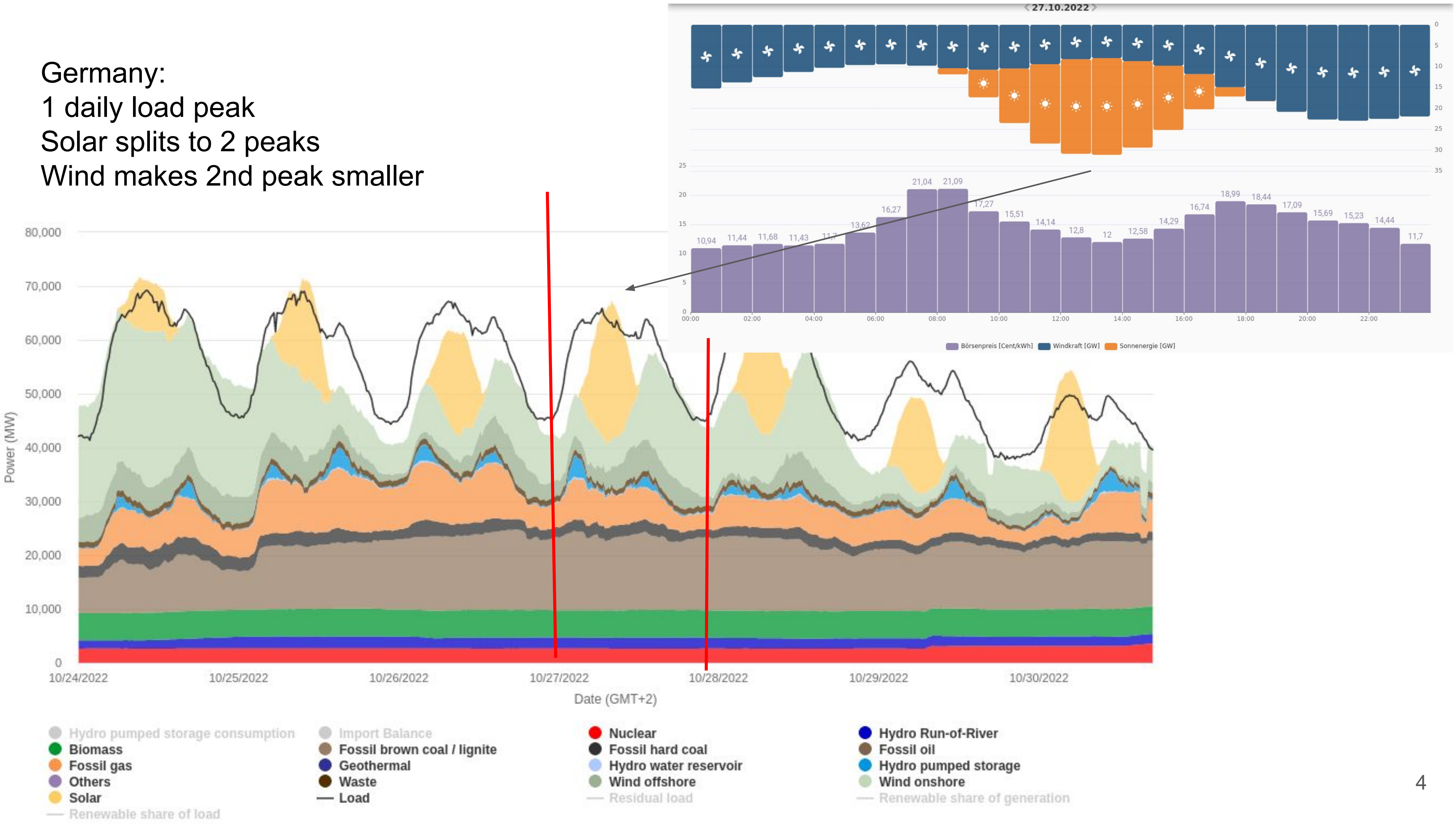


1. Turn off old hardware, during crisis or permanently
    - W/HS06 and W/TB often significantly higher for older hardware
    - O(10%) reduction in cpu or storage ok (if no ATLAS ops action needed)
        - post-crisis turn-on again, or return to pledge with 2023 hardware.
        - starting point is the pledge. Many sites way over cpu pledge(not storage).
2. Power down additional compute nodes to get to targeted saving
    - highest W/HS06
3. Compute cluster 100% powered down
4. Storage disk nodes powered down
    - keep headnode services running, and turn on pools once per week
        - DT coordination, some effort, maybe some risk
5. Storage 100% down

Why only the peaks?

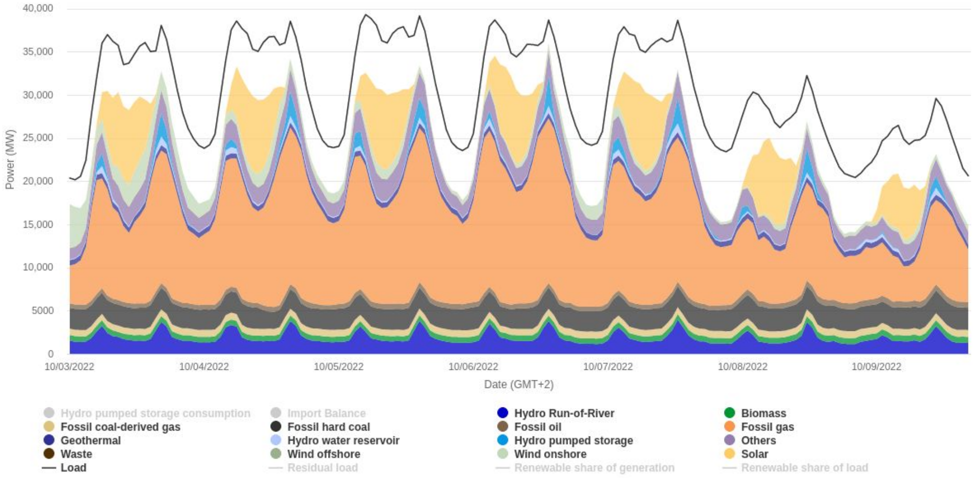
- Short-term price increase caused by replacing Russian gas, French nuclear
    - gas used for electricity generation particularly in consumption peaks
    - EU wants voluntary 10% flat reduction, but mandatory 5% in peaks
        - “..identify the 10% of hours with the highest expected **price** and reduce demand during those peak hours.” amounts to 3-4hrs per weekday.
    - address underlying physical problem, leading to the financial one
- Long-term ‘normal times’ prices likely to vary more with time of day
    - daily load peak conflated with intermittent renewables, network congestion, storage state
        - leads to peaks in fossil fuel usage, and price (ideally the same peaks)
    - shapeable loads will be vital for grid stability and tarif priced accordingly

Germany:
1 daily load peak
Solar splits to 2 peaks
Wind makes 2nd peak smaller



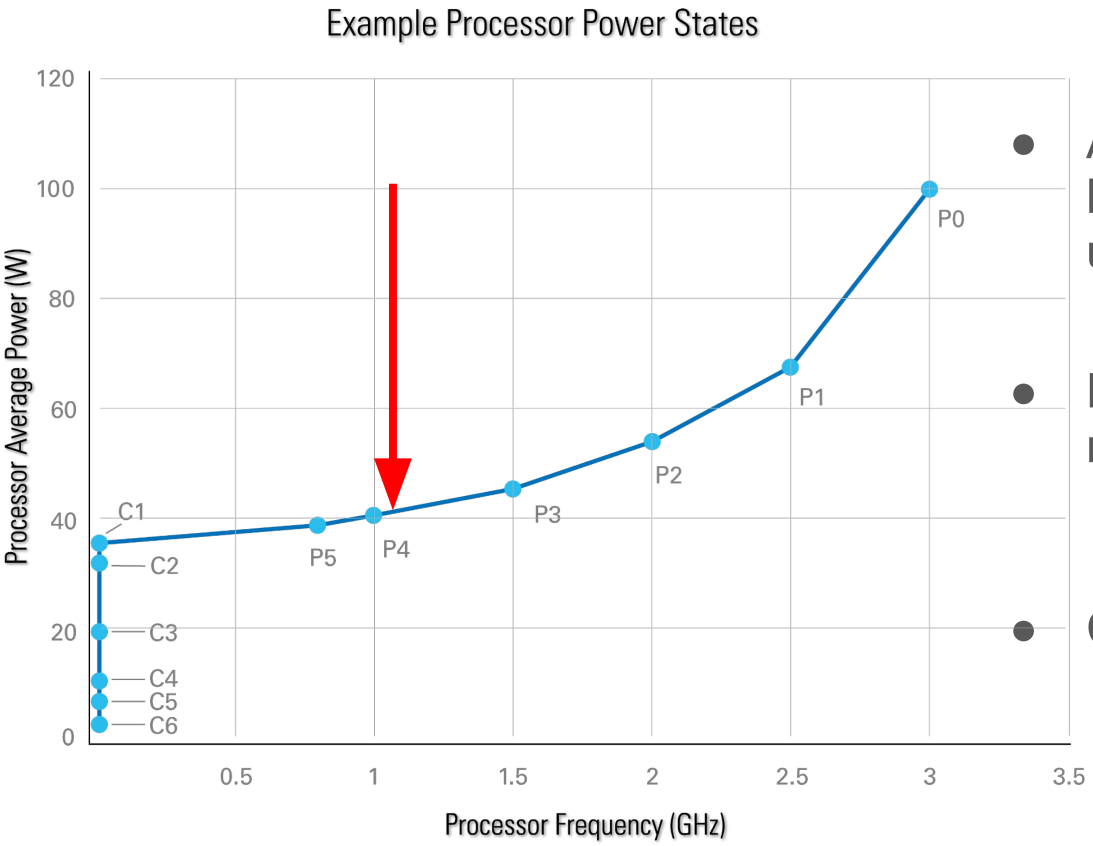
Power reduction at peaks

- Typically twice per weekday for 2 hrs
  - can’t drain nodes of jobs with lengths up to 4 days
  - can’t preempt jobs at this frequency, due no checkpointing and so cpu waste
  - probably not wise to power cycle nodes/power supplies/disks at this rate
- Repeatable cycle to save power with no bad effect on running jobs or pilots?



Italy from https://energy-charts.info
Clear peaks, but gas base too.
Complex - rely on experts and the market price having it all in.

Bluffer's guide to CPU power management



- All 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, $P \sim f V^2$
  - set by bios, OS governor or manually
  - OS total control, or combined with bios
    - pcc_cpufreq module
- C: shutting down parts of cpu
  - only happens when idle
    - @core granularity
    - SIGSTOP processes?

Hardware configuration(still bluffing)

- Operating system power management can only operate within constraints of BIOS firmware config
    - set min,max freq and who controls it (firmware/OS)
- LRZ-LMU SLES15 nodes load module pcc_cpufreq
    - OS gives hints to firmware, but firmware changes frequency
        - found the frequency did not reduce with governor change, and not completely with suspended processes.
        - Apparently should not be used for >4 cpus(cores), patched from 4.19 (sles 4.12)
    - need to blacklist this module or reconfigure BIOS
- OS control of cpu frequency should be possible for all sites
    - but might need reboot.

Power saving actions

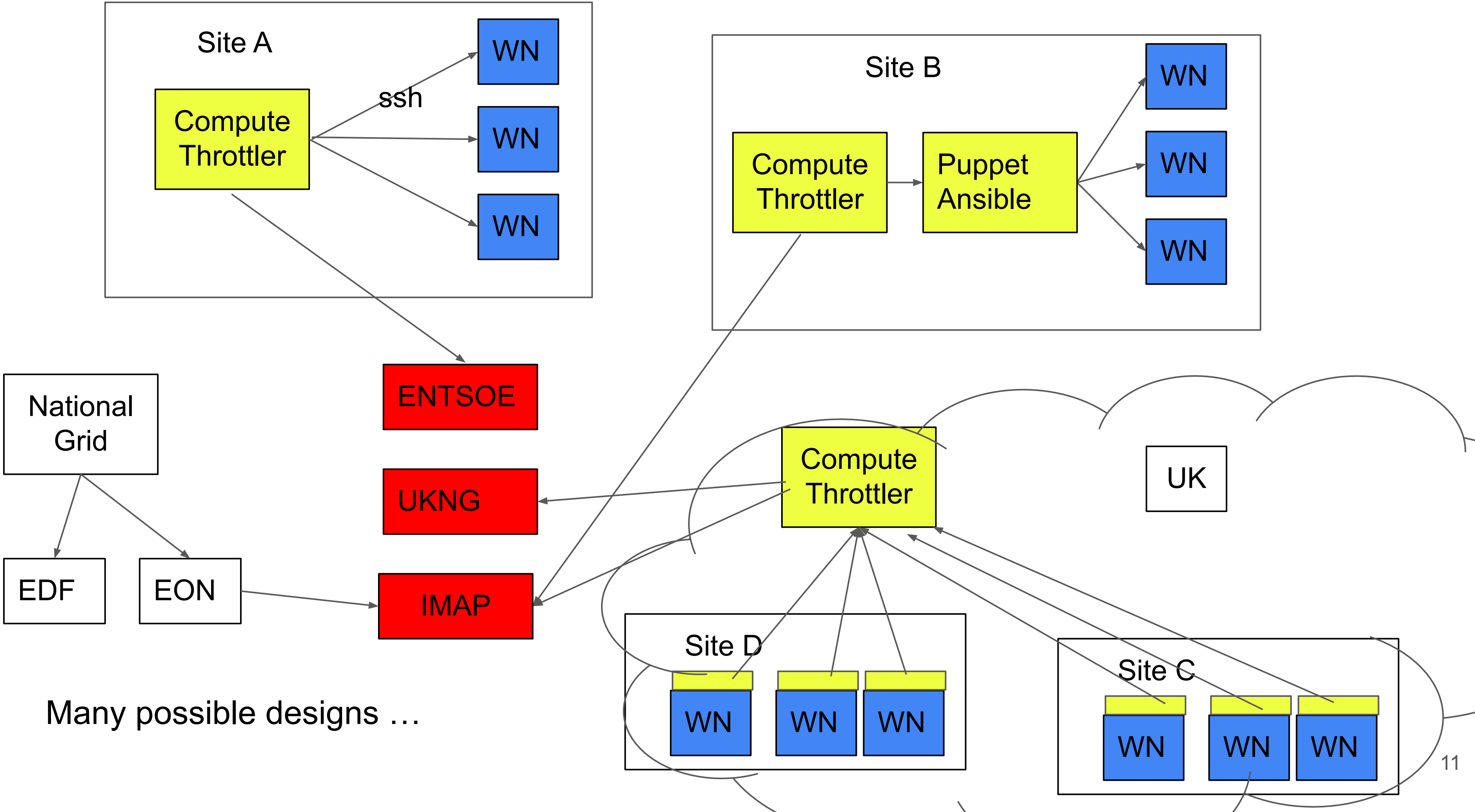
- Direct CPU frequency scaling to minimum: ~66% power saving
  - instant effect, also in reverse
  - transparent to workloads, apart from slowdown
- Or suspend processes, e.g. scontrol suspend [jobs], condor_suspend
  - SIGSTOP to all workload processes
    - then governor reduces the CPU frequency due to no load
    - might get into c-states, to save more power?
  - also stops pilot or overlay-BS startd
    - heartbeats not sent: ATLAS jobs would survive 90mins(configurable)
  - just SIGSTOP cpu-intensive processes? Pilot/Startd knows which is payload.
    - single core runs all pilot processes.
- Direct CPU frequency scaling has simplicity on its side
  - independent of BS, VO payload and WFMS
  - take the 66% for now.

Forecasts to schedule power saving pauses

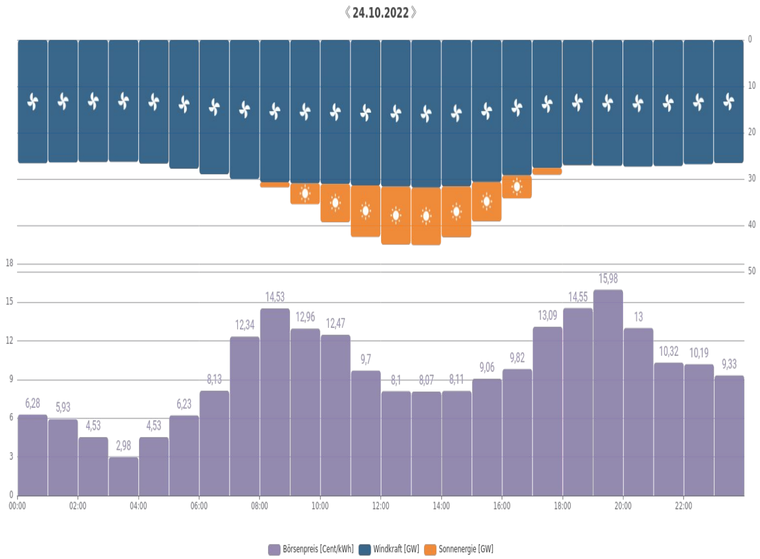
- Can assume day-ahead market price reflects physical need
  - EU wording specifically says to use this to identify the 10% peak hrs.
  - includes demand, weather forecasts(wind, solar), power station schedules
  - misses sudden deviations in weather, failures
- Available for most EU countries
  - https://transparency.entsoe.eu/ with API to retrieve JSON
  - https://www.awattar.de/tariffs/hourly same information for DE/AT, convenient without token
  - I can’t find it for UK, but has https://data.nationalgrideso.com/carbon-intensity1
- Direct signal from National Grid or energy provider
  - UK Demand Flexabilty Service sends mail/SMS with start time and duration to reduce power
    - pays 3GBP/kWh saved c.f. baseline. Business customers included.
    - need smart meter and participating provider, e.g. EON, EDF

Tool to schedule and trigger power saving actions

- Use forecast and some algorithm to schedule actions
  - ENTSOE, Awattar, NGESO or direct signal(IMAP)
  - find local maximum, or sliding window to maximize value, EU algorithm(TODO)
- Actions supported:
  - ‘scontrol suspend/resume [jobs]’ with reservation to block new jobs
  - ‘cpupower **f**requency-set -g powersave/schedutil’
    - either by parallel_ssh or sharedFS control file read by cron on WNs
  - TODO: puppet, ansible?
- https://gitlab.cern.ch/walkerr/computethrottler
  - past ‘proof of principle’ to ‘usable demo’ level, but still rough.
  - config, logging, systemd service but no rpm.

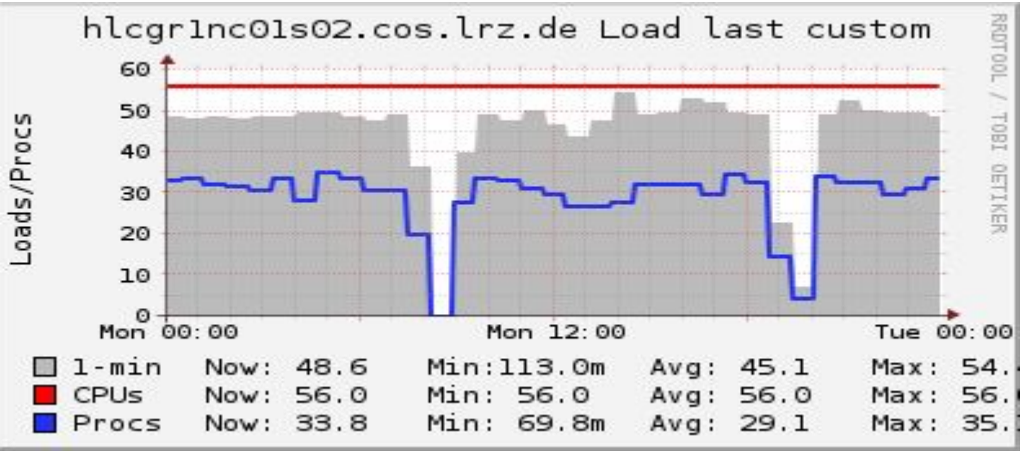


Many possible designs …

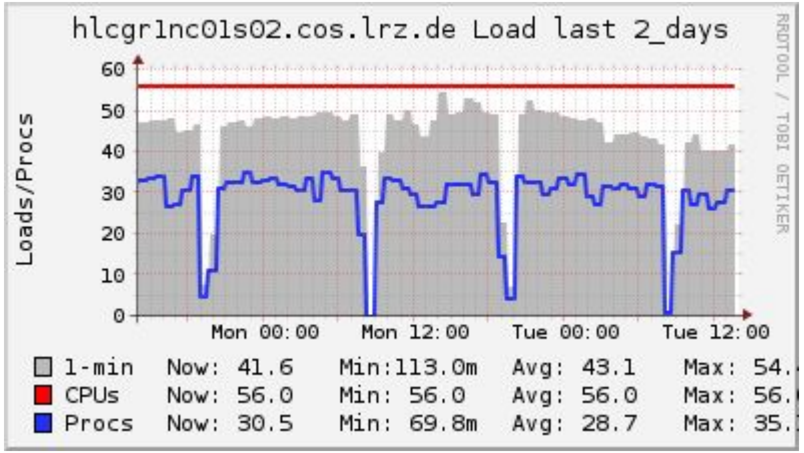


Throttling 1 WN at LRZ-LMU for 2 weeks

- Variable tariff based on spot-price
- 1hr power save at peaks
  - slurm suspend in this case



```
□ 1-min   Now:  48.6   Min:113.0m   Avg:  45.1   Max:  54.
■ CPUs    Now:  56.0   Min:  56.0   Avg:  56.0   Max:  56.
■ Procs   Now:  33.8   Min:  69.8m  Avg:  29.1   Max:  35.
```



```
□ 1-min   Now: 41.6   Min:113.0m   Avg: 43.1   Max: 54.
■ CPUs    Now: 56.0   Min: 56.0    Avg: 56.0   Max: 56.
■ Procs   Now: 30.5   Min: 69.8m   Avg: 28.7   Max: 35.
```

Monitoring & Accounting

- Would like to see effect on power consumption, e.g. MONIT
  - kW reduction per site, region, forecast used
    - can be an estimate, based on 1-off measurement
  - store forecast data for plotting, archive & uniform access (for the throttler)
- Show idle HS06, due to load-shaping and temporary power down
- APEL accounting uses single HS06 rating per cluster
  - job on frequency throttled node HS06s=HS06_nom(wall_full + wall_pause*0.33)
  - job takes longer: short job with tight maxwalltime might suffer
- This Winter: no need to do accounting properly
  - 4hrs pause per day ~ 10% HS06s - within errors from non-homogeneity/HT.
  - monthly correction based on monitoring?
  - ensure no bad consequences for contributing sites.

Conclusion

- Power down old hardware for Winter to get 10% flat reduction in Europe
    - do out of solidarity: regardless of power bill problem. VOs will accept this..
- Can easily shed 66% load from compute cluster during peaks
    - motivated sites needed to improve and harden service
        - leading to simple service for WLCG deployment
    - lack of financial benefit(due to flat tariff), missing monitoring or accounting NOT good reasons not do this.
- Tools and lessons will be useful when variable tariffs are available/standard
    - overdue and unavoidable as more intermittent renewables in mix