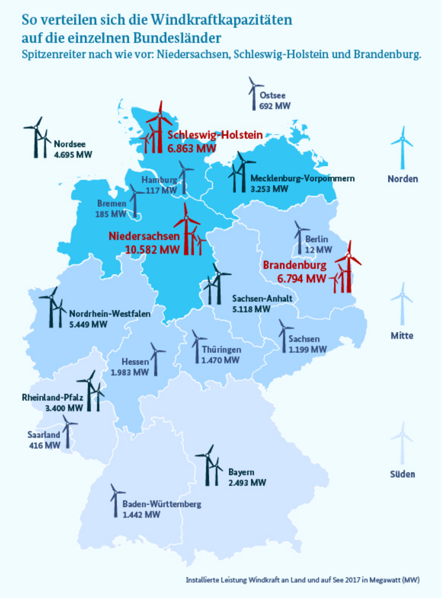
Energy Conservation Considerations @ DESY

Compute Cluster Preparations

Thomas Hartmann
DESY IT

Energy Consumers

Triage & planed energy usage modulation


So verteilen sich die Windkraftkapazitäten auf die einzelnen Bundesländer
Spitzenreiter nach wie vor: Niedersachsen, Schleswig-Holstein und Brandenburg.
Ostsee 692 MW
Nordsee 4.695 MW
Schleswig-Holstein 6.863 MW
Hamburg 117 MW
Mecklenburg-Vorpommern 3.253 MW
Norden
Bremen 185 MW
Niedersachsen 10.582 MW
Berlin 12 MW
Brandenburg 6.794 MW
Sachsen-Anhalt 5.118 MW
Nordrhein-Westfalen 5.449 MW
Mitte
Sachsen 1.199 MW
Thüringen 1.470 MW
Hessen 1.983 MW
Rheinland-Pfalz 3.400 MW
Saarland 416 MW
Bayern 2.493 MW
Süden
Baden-Württemberg 1.442 MW
Installierte Leistung Windkraft an Land und auf See 2017 in Megawatt (MW)

Short Term:

- depends on the winter
- load shedding if necessary
  - on-site instruments mayor consumers
  - IT ~5-10%

  - IT triage by relevance
    1. compute clusters
    2. storage instances
    3. central services
  - depending on local data taking

Mid Term:

- Hamburg with large renewable energy hinterland in northern Germany
- adaptable energy usage w/r to production conditions

- keeping clusters up at 100% with varying load has been an *extravagance*

- what latencies for cluster modulation w/r to power source modulations realistic?

User HTC Cluster

NAF HTCondor Cluster

National Analysis Facility - User Cluster

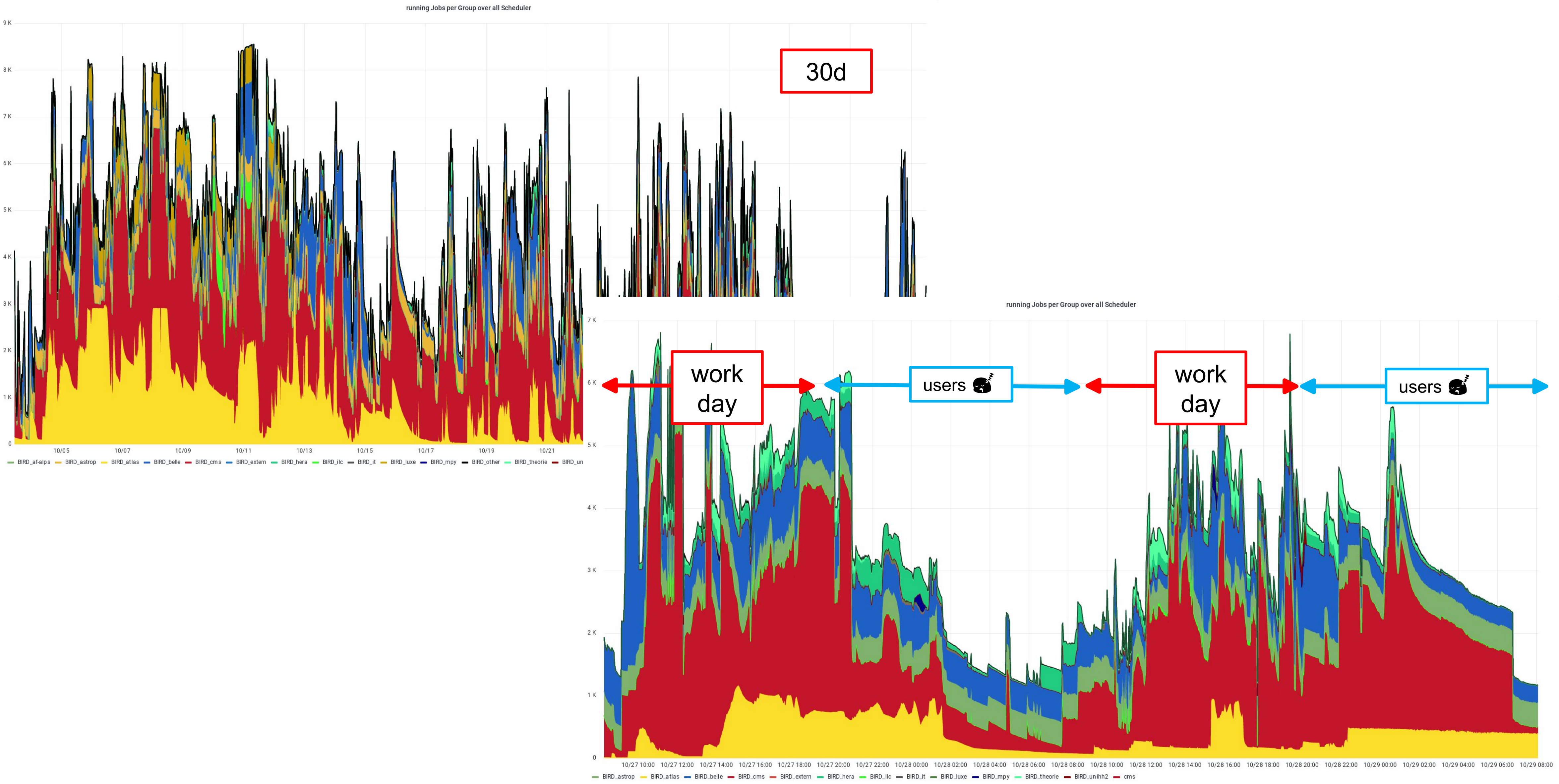
- NAF: complementary to the Grid for individual users' jobs

- cluster utilization by the users fluctuating
  - day/night user behaviour + seasonable effects (aka conferences & holidays)
  - power consumption closely coupled

- had been keeping resources available 24/7
  - low job start latency pleases/placates users
  - now might become a noticeable cost

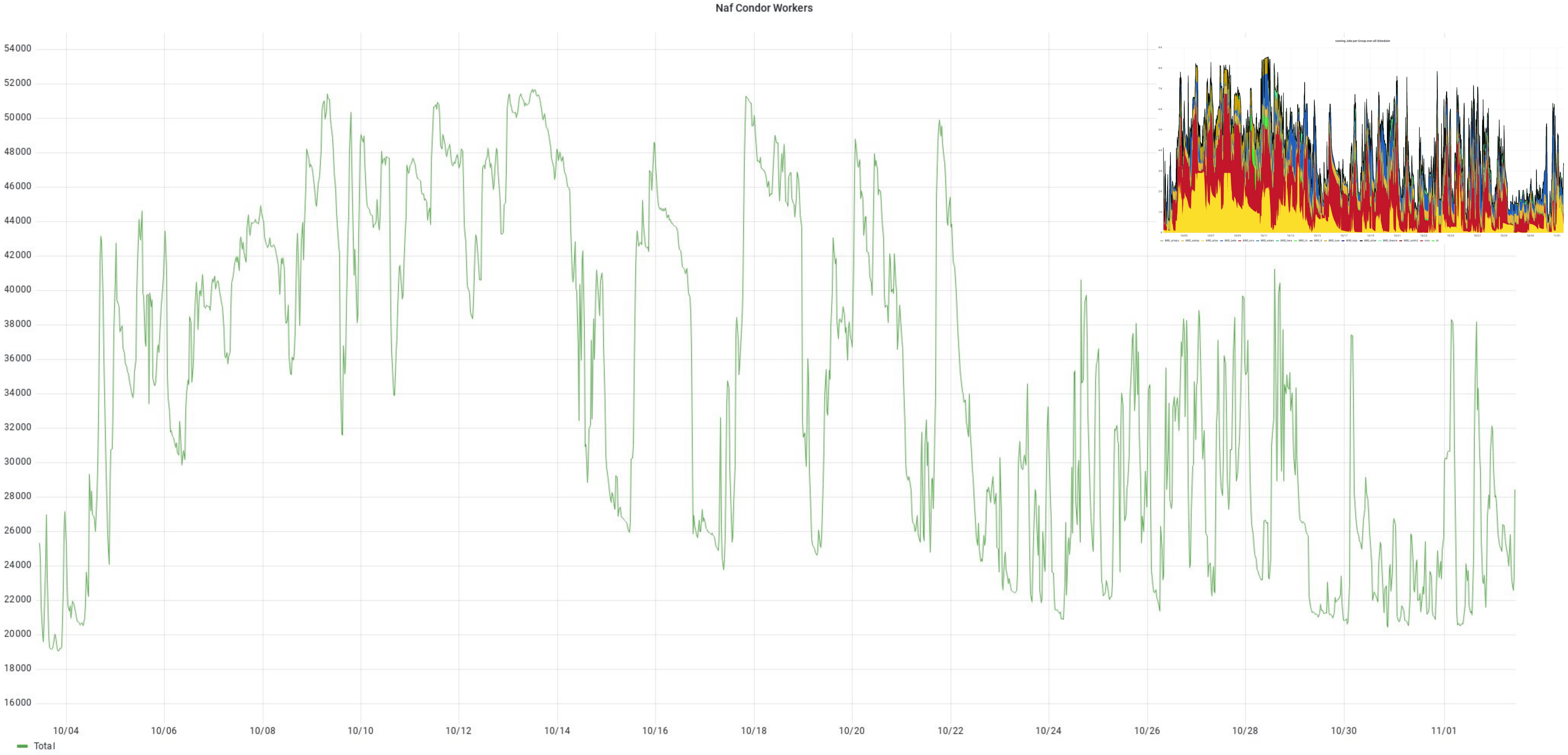
NAF HTCondor Cluster: Local User Jobs

Utilization over 30d/2d



NAF HTCondor Cluster: Load dependent Consumption

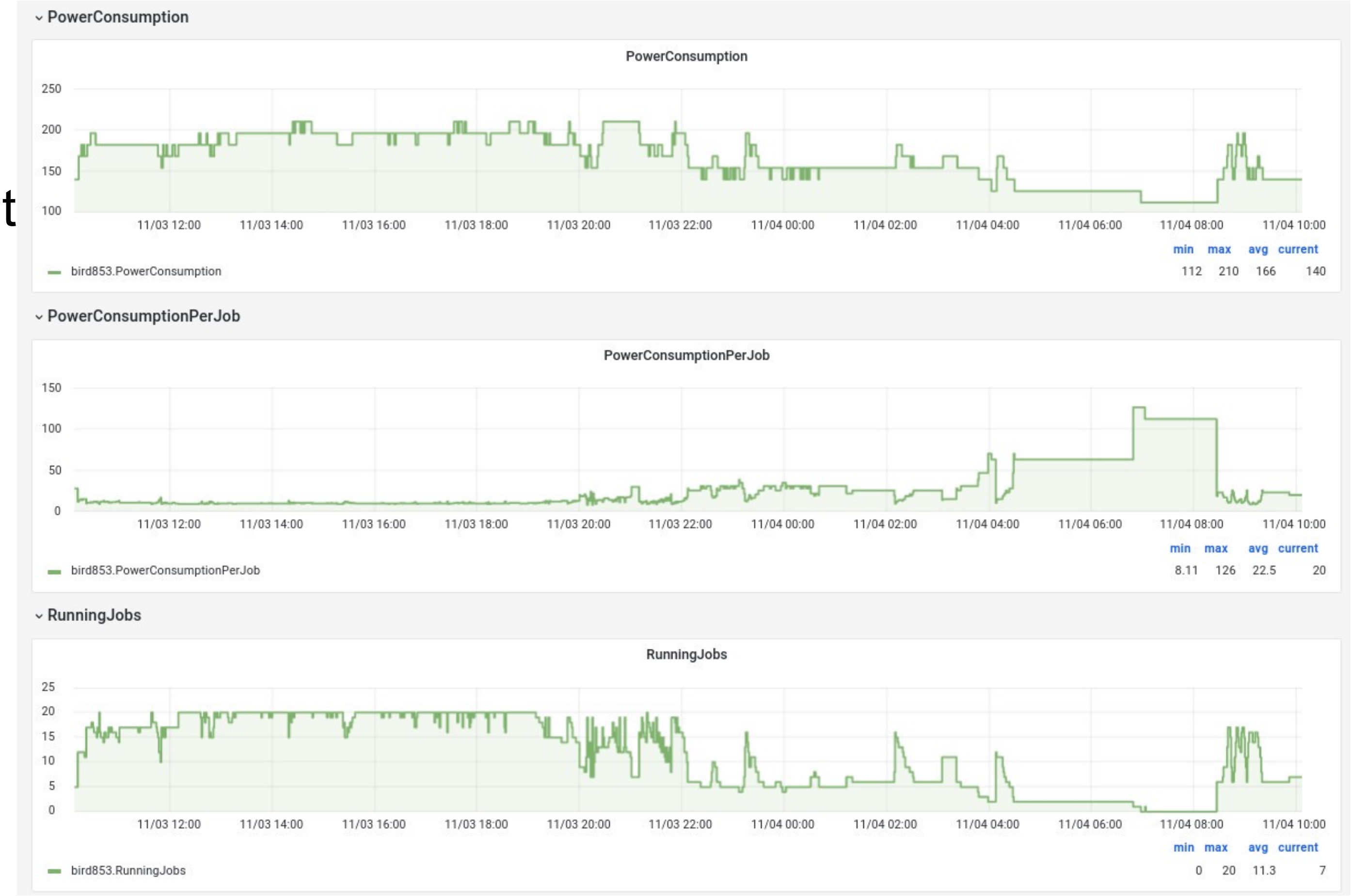
Power usage in kWh over 30d (incomplete, some older workers’ PSU do not report their consumption)



NAF HTCondor Cluster

User jobs with runtime requirements

- already enforcing user jobs run times request
  - makes scheduling/planing possible
- currently: horizontal scheduling to distribute job entropy
- going for more vertical scheduling condensing short jobs on workers
  - easier & faster draining for projectable load shedding
- rough & ready user education:
  power consumption + CO2 load summaries



NAF HTCondor Cluster

User jobs with runtime requirements

- setting up worker drain/shutdown/wake flows
- central wake via Foreman solution already in place
- currently manual steps ~> automation upcoming
- investigating *rctwake* for power napping (S2/S3 sleeps problematic)

- drain and wake on power sources as well as cluster load
- either central wake or individual worker wake

- cluster power ceiling as midterm aim
  - max total cluster power consumption as tunable

Grid HTC Cluster

Grid HTCondor Cluster

Grid prod jobs

- cluster utilized 24/7

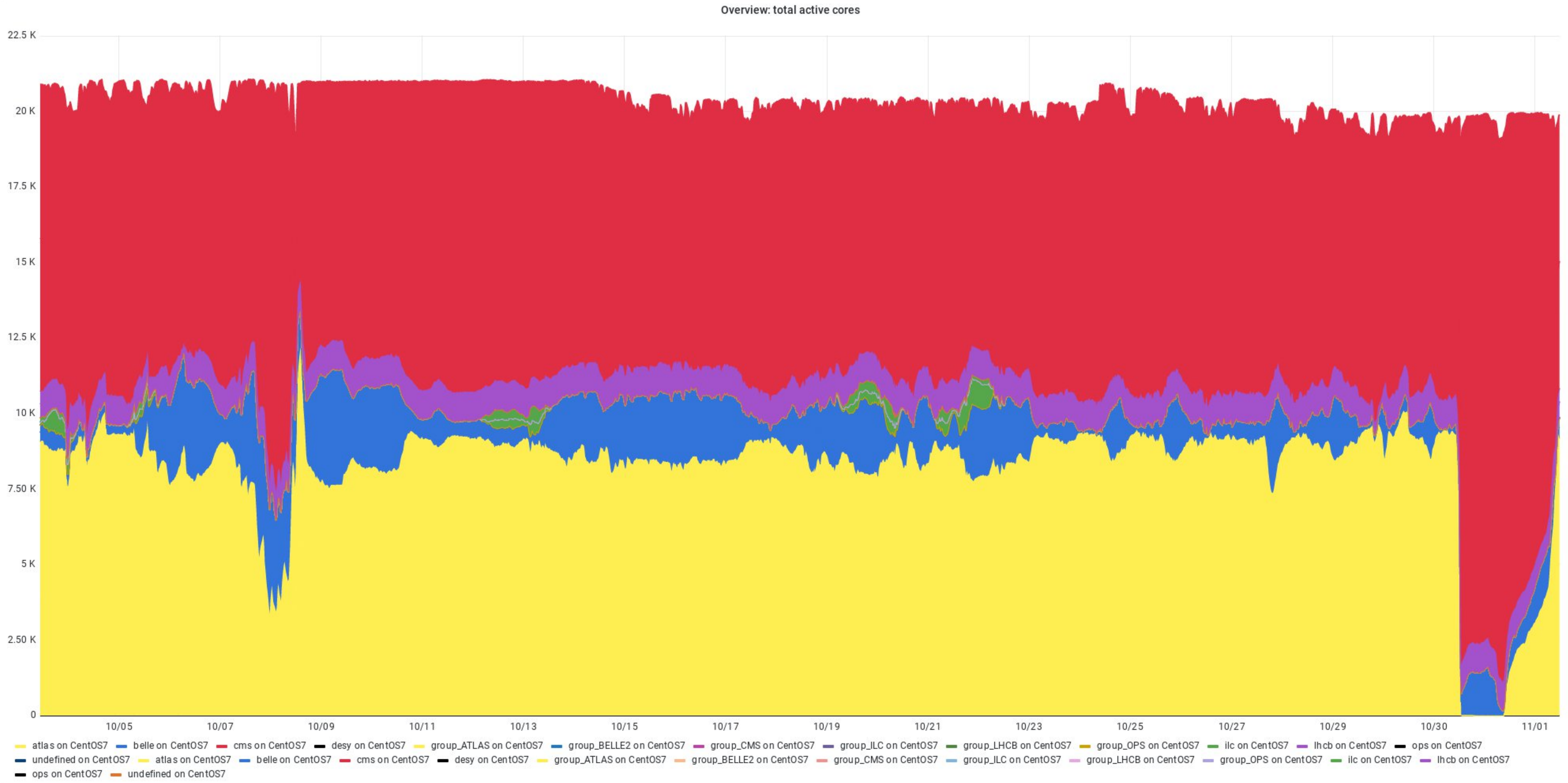
- high utilization - more *efficient/effective* than the NAF user cluster
  - w/o respect to job start latency
  - much higher inertia...
  - dynamic adaption to power provisioning only on longer time scales

- some sensitivity on payload efficiency (wall vs cpu time)

- investigated transparent job/CPU throttling as stop gap

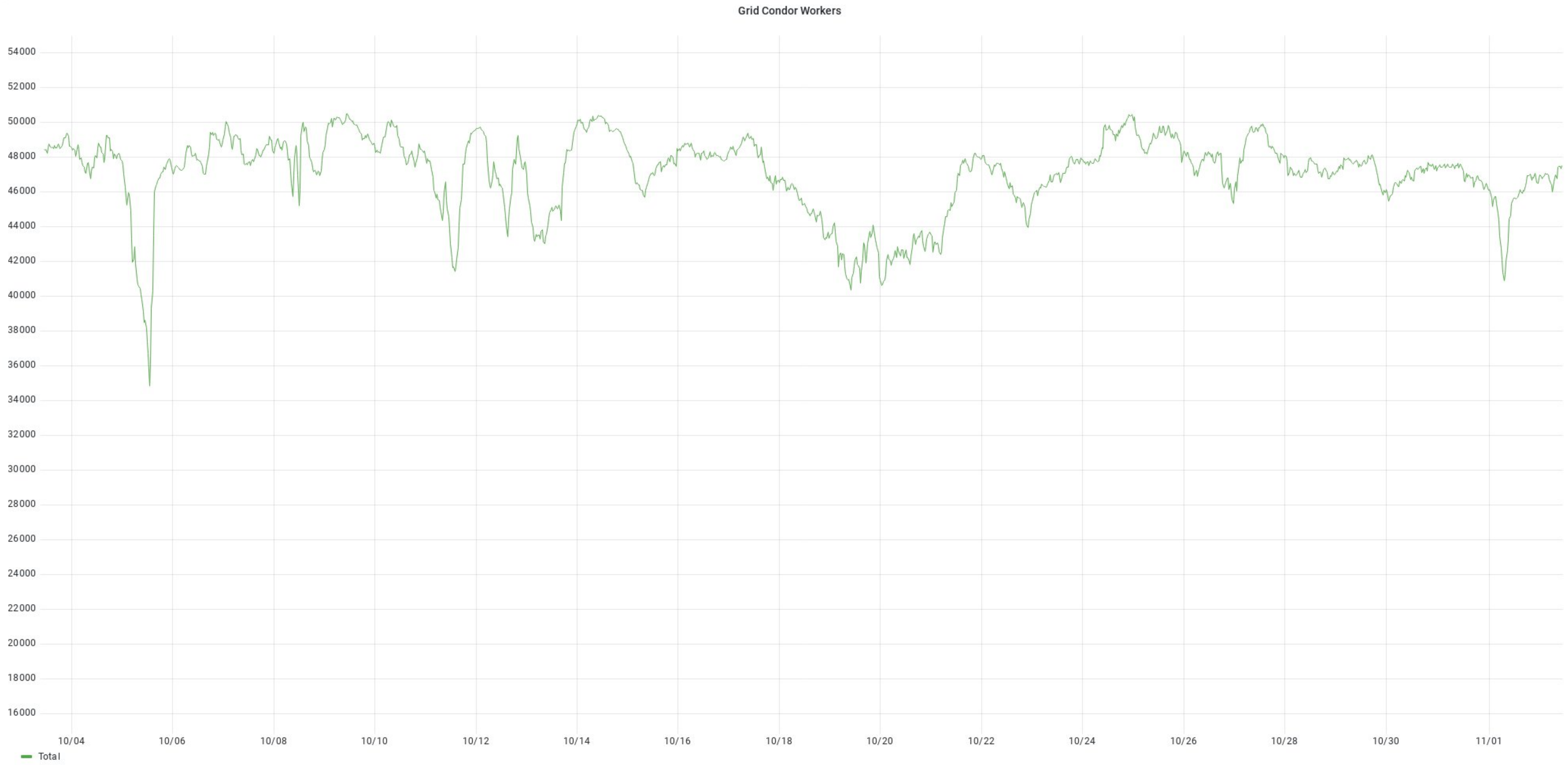
Grid HTCondor Cluster

Utilization over 30d



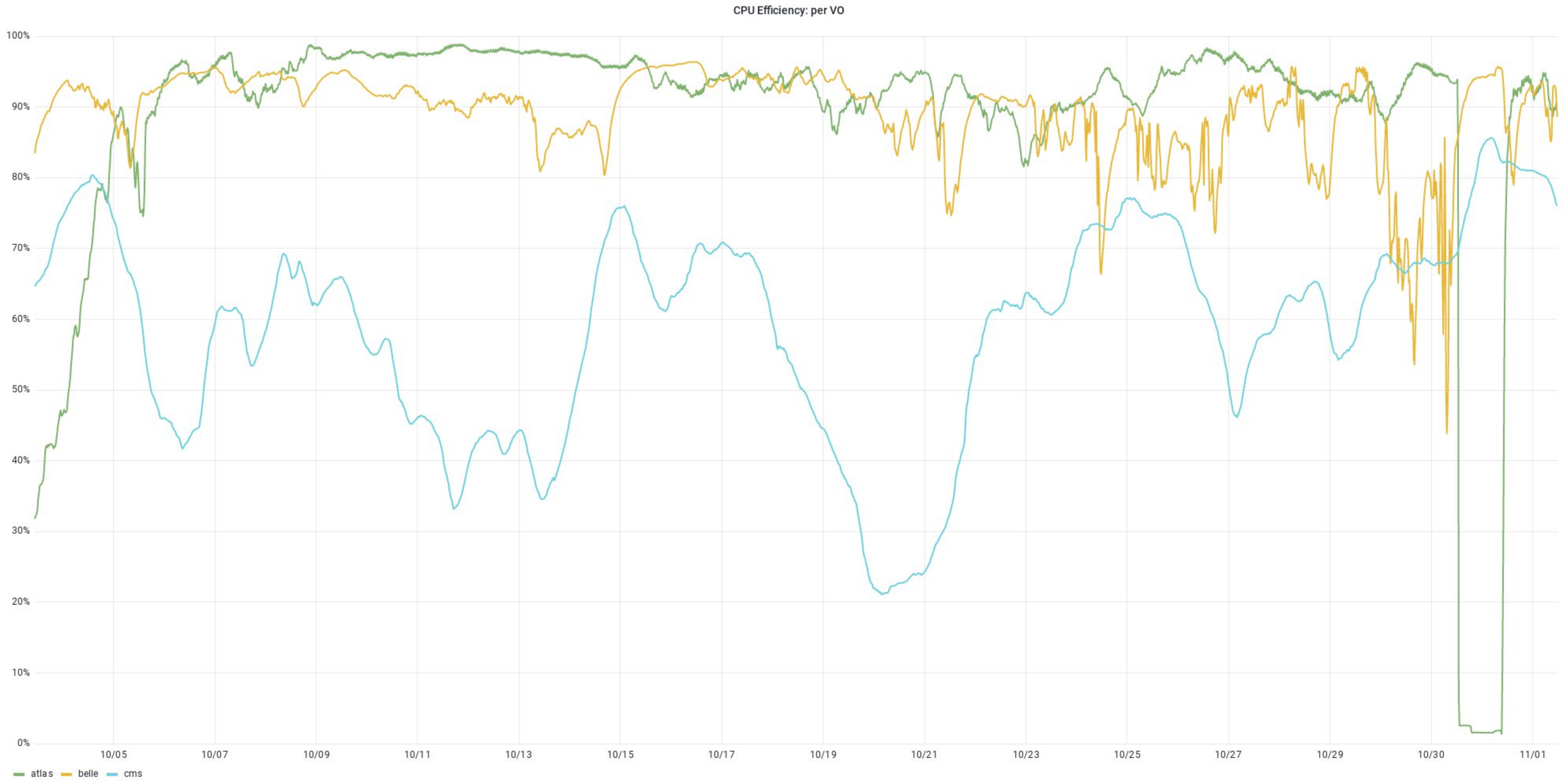
Grid HTCondor Cluster: Power Usage

Power usage in kWh over 30d (incomplete, some older workers' PSU do not report their consumption)



Grid HTCondor Cluster: Job Efficiencies

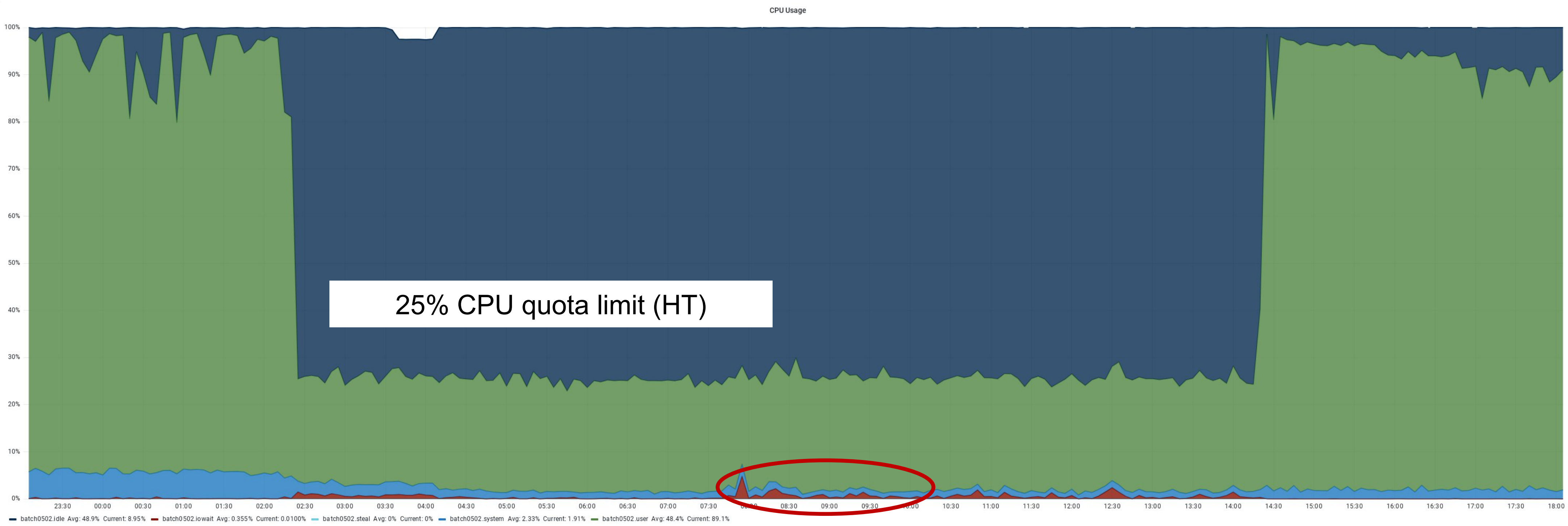
Power usage correlated with VO Jobs Wall Time Efficiency



Job Throttling

Limiting job usage by cgroup CPU quota

- HTCondor constrains jobs in cgroups
- CPU time quota can limit all/individual job payloads in their CPU walltime
- (mostly) transparent to the jobs
- depends on the CPU governor, HT/SMT, freq settings,...
- energy savings limited by base load



25% CPU quota limit (HT)

Grid HTCondor Cluster

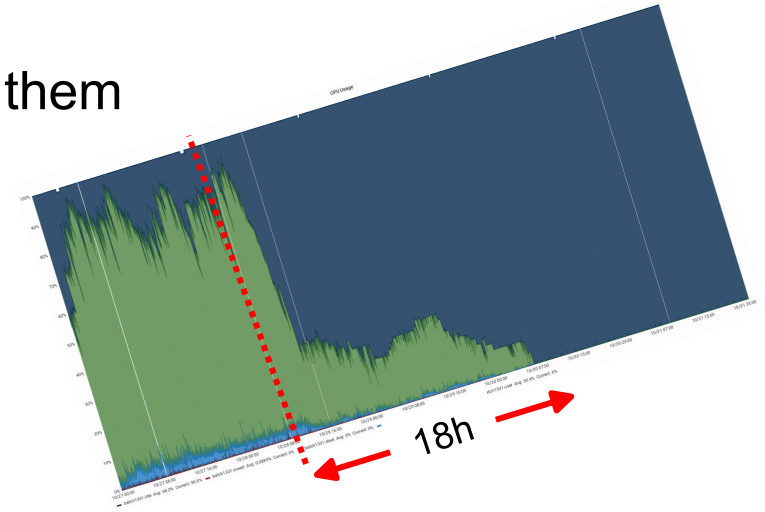
Pilots make projectable scheduling impossible, Payloads not preemptable

- not really feasible to adapt, i.e., drain, with respect to energy source modulations
- longest running jobs force the min frequency for draining
  - CMS 48h pilots with horizontal scheduling
  - vertical guess scheduling aka ATLAS, Belle, ILC,... would penalize them
    - i.e., effectively segmenting cluster by VOs

- w/o preemption, checkpointing,... no transient load shedding possible

- hard load shedding would waste significant consumed power

- transparent CPU throttling in principle possible
  - significant power fine by offset base consumption
  - draining and keeping off/load shed sections of the cluster more economical



Job/CPU Throttling

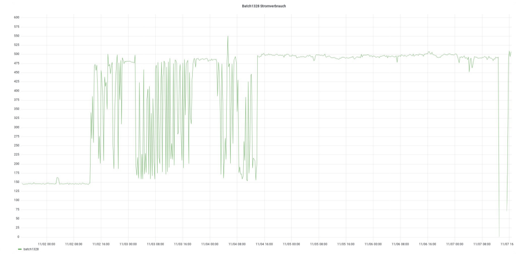
On demand throttling

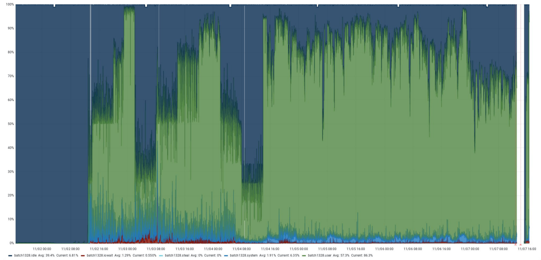
- run a few tests
  - throttling node to [100%, 75%, 50%, 25%] CPU time + [0 load, off]
  - PSU & PDU power consumption(s)
  - ~75W per 25% steps (@25% extra *savings* due to IOwait...)

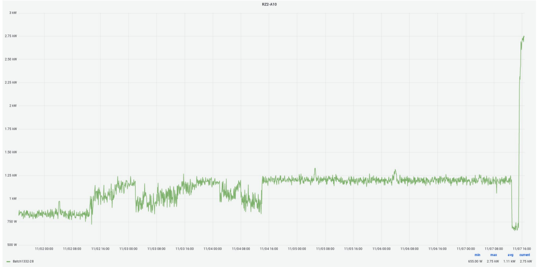
- base idle load ~150W incl. PSU ~10% inefficiency

- realistically 1/3 of the power consumption might be saved by throttling...
- ...with a ~150W base offset
  - not very efficient (effective??) for a nearly 100% utilized HTC cluster

- **conclusions** for power savings or cluster power ceiling
  - load shedding nodes for good...







Summary

HTC Cluster energy saving outlook

NAF

- dynamic user HTC cluster with realistic saving options
  - horizontal —> vertical scheduling
  - compacting schedulable short jobs/nodes for quick draining/shedding
    - utilization management
  - investigating power ceiling / capping
    - dynamic max cluster power consumption with automatic shedding

Grid

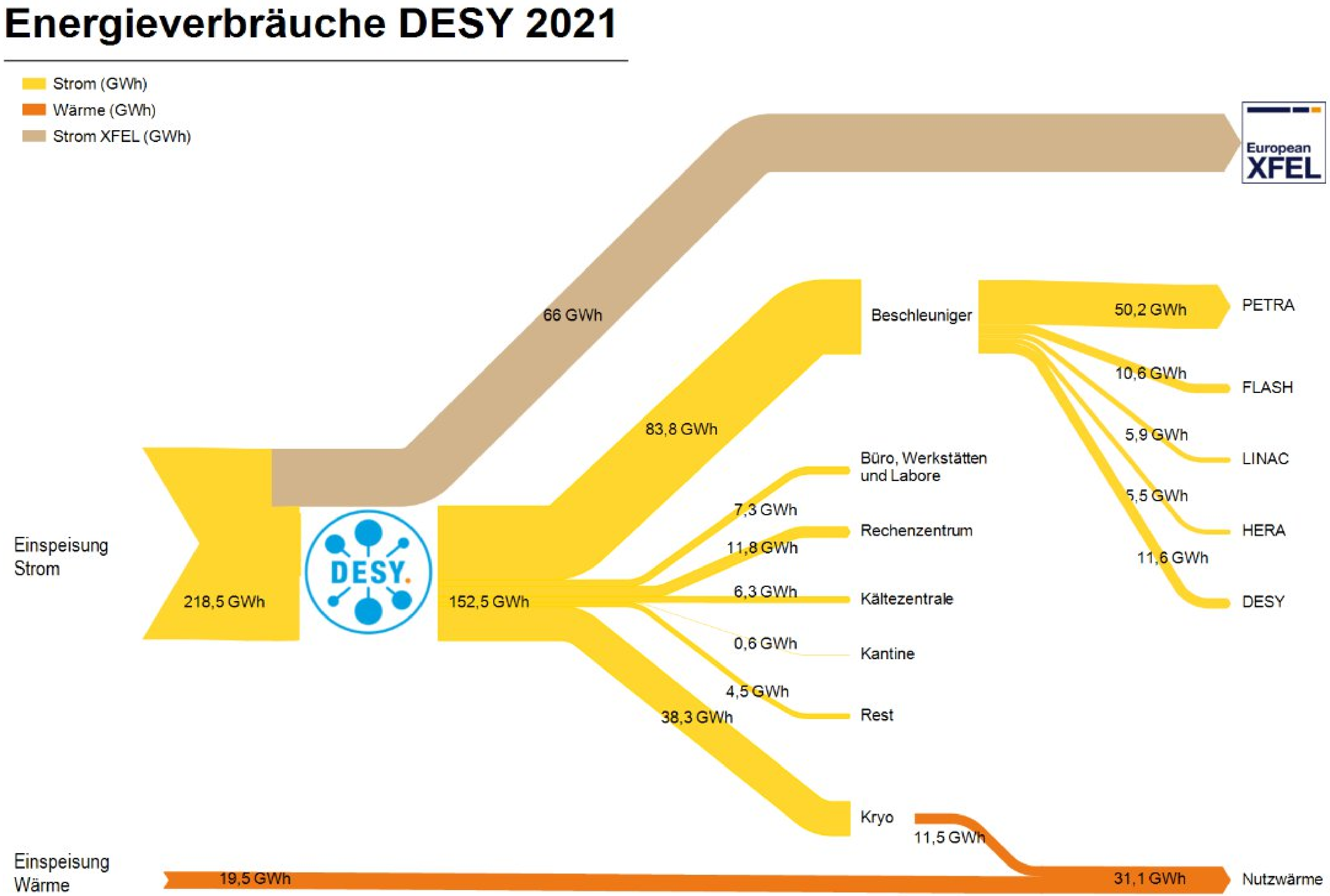
- more static Grid HTC cluster already ~100% utilization
  - dynamic load shedding constricted by w/o scheduling info
  - job CPU time or CPU freq throttling prohibitive base idle load

Thank You

Questions?

Appendix

Hamburg Campus Total Power Consumption



Energieverbräuche DESY 2021

Contact

DESY. Deutsches
Elektronen-Synchrotron

www.desy.de

Thomas Hartmann
DESY IT
https://naf.desy.de
https://bird.desy.de