Load shedding
Power saving … the DESY view
Sustainability

Christoph Beyer, Thomas Hartmann, Yves Kemp et al. DESY IT
Hamburg 15.3.2023

… speaking for Hamburg mainly

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

DESY.

The big picture a.k.a. Vision

Where we want to be in (hopefully) near future:

- Consume renewable energy when it is available.
- Reduce energy consumption when renewable energy is sparse.

This means: Flexibility:

- Flexibility in the consumption by users
- Flexibility in the infrastructure
- Flexibility in the contracts

Become more efficient: Optimize ~~Workload/money~~ ~~Workload/Energy~~ ~~Workload/CO2~~ Workload/Consumables

Be green, optimze, pay less

Recent history and upcoming future

- Assumption: There will be (frequent and) short-term interruptions in power provisioning
- Reality: Did not happen. At least not on short-term.
- Power consumption profile rather well known. Power production profile (RE) known up to 2 days in advance (TransnetBW “StromGedacht”)
- Assumption: Energy prices will kill us. Reality: Prices in 2022 not that exceptionnally high

- The time for immediate action is over
- ~~Time to relax and get back to business as usual~~
- Time to design and build really sustainable research infrastructures

Some ideas and work directions

- Analyze workloads and infrastructure
  - of: Grid / NAF / Maxwell HPC
  - from: WLCG, national HEP, photon science, machine R&D and ops, …
- Integrate storage into the big picture
- Integrate computing centers into the big picture
- Provide input to sustainability department
- e.g. in projects “EU Horizon Research Infrastructures 2.0”

- **Important: Redefine communication and interaction between:**

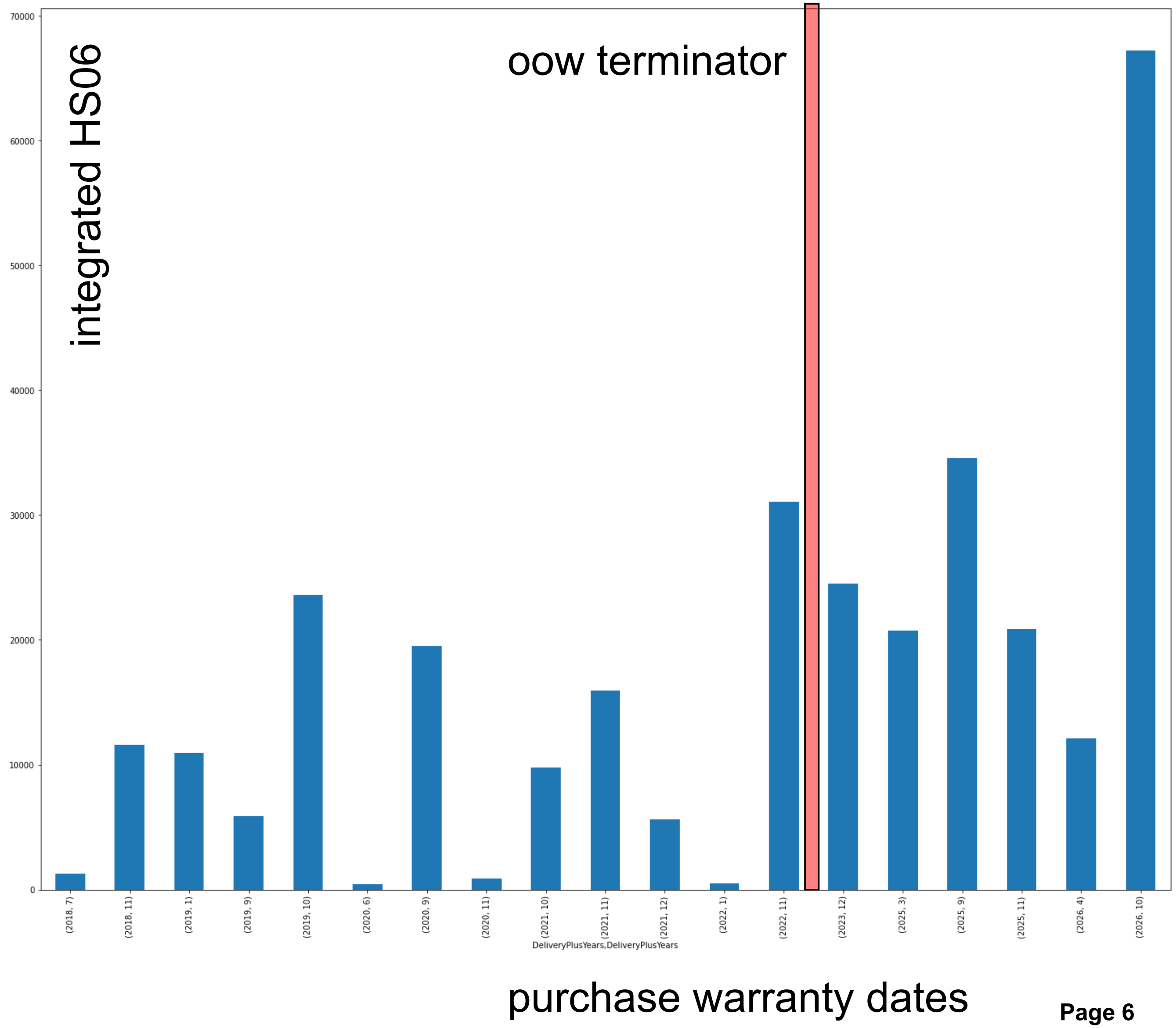
  Experiments ←→ Ressource provider ←→ Energy provider

Optimizing Cluster Energy Efficiency

Cluster Energy Efficiency

HepSpec by Generation

- Grid pledge policy so far
  - Pledges with under warranty workers
  - Extra HS06s from oow workers

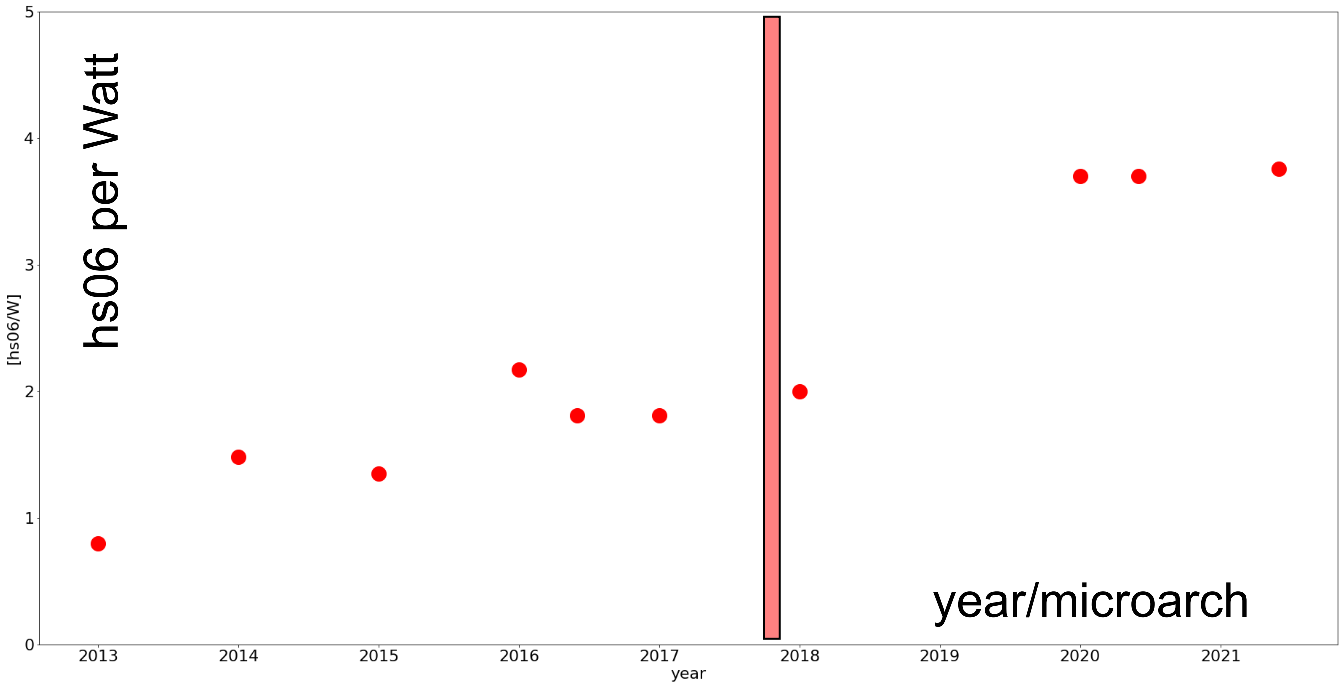


DESY.

Cluster Energy Efficiency

Arch HS06 per Watt

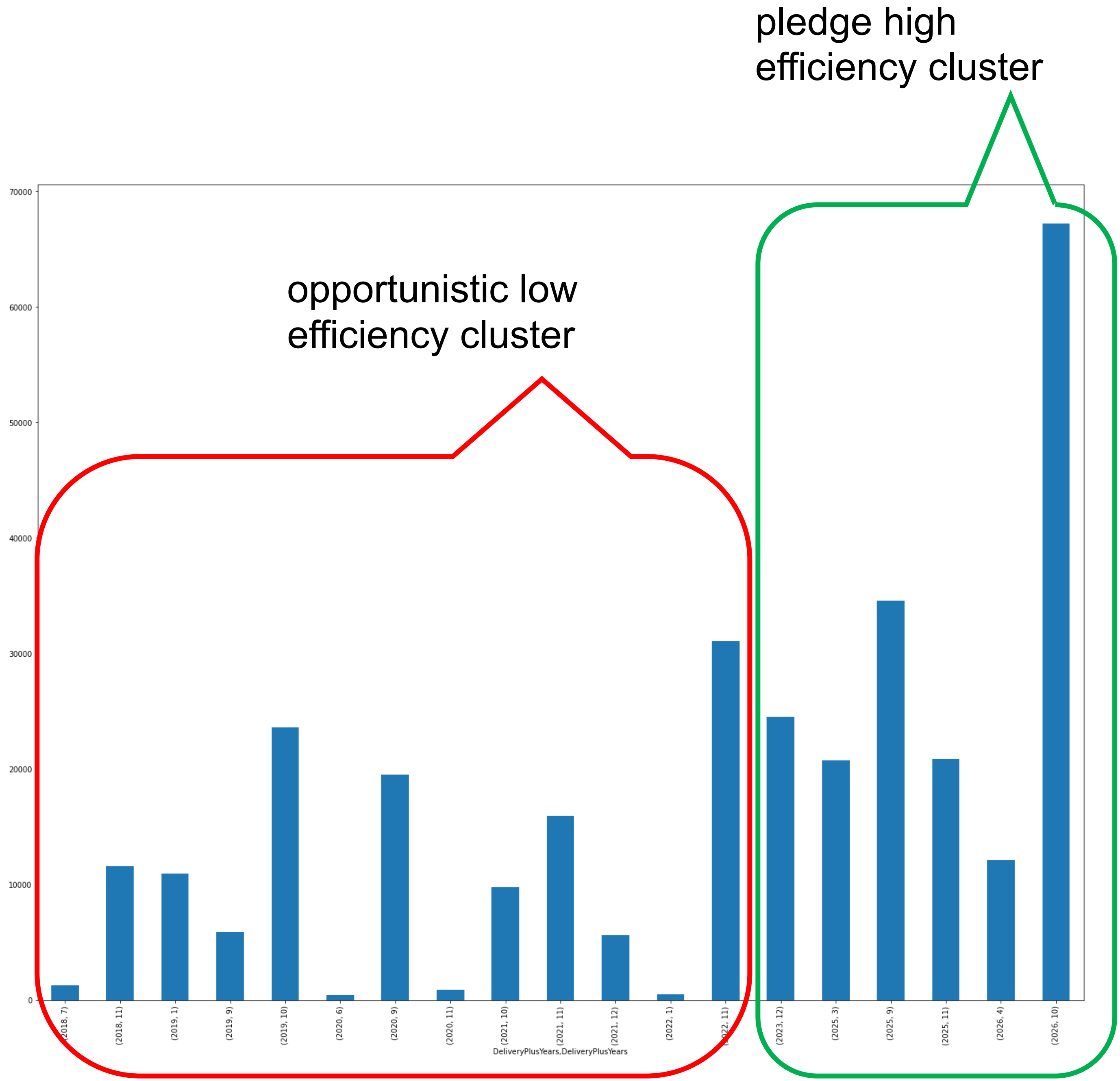
- Significant efficiency gains with recent microarchs (aka Zen)
- HS06 per Watt gain ~4x from oldest workers still in production



Cluster Energy Efficiency

Cluster sub designations

- Need to reconsider cluster operations with respect to efficiency
- Operating inefficient workers 24/7/365 still justifiable?
- Pledged high efficiency resources always online
- Low efficiency cluster as opportunistic resource
  - Load shedding when necessary
  - Scheduling needs to be adapted



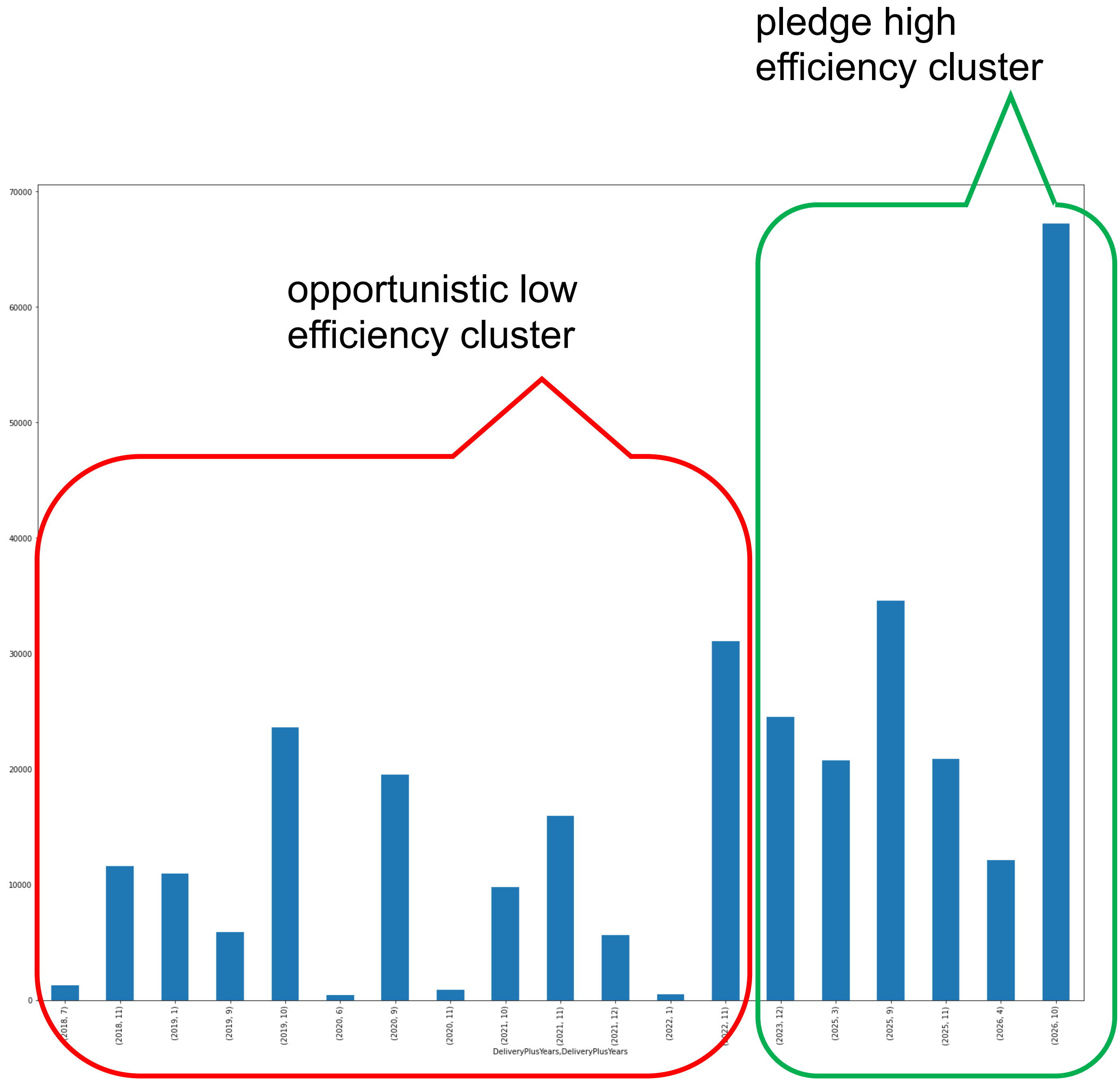
Cluster Energy Efficiency

WattHours consumed for HS06 delivered

E.g.

- target deliverable: 1000 kHS06

- “combo” cluster: ~410 kWh
- “high efficiency” cluster: ~298 kWh
- “low efficiency” cluster: ~587 kWh



Opportunistic Resource Utilization

Burning off surplus green energy

- Complementary to load shedding
- Opportunistic green energy sink
- Offshore wind farms
  + limited transport capacity in the south
  = potential to "burn off" energy with opportunistic computing
- ATLAS: additional compute resources
- Our benefit: increased utilization/efficiency of Maxwell + NAF

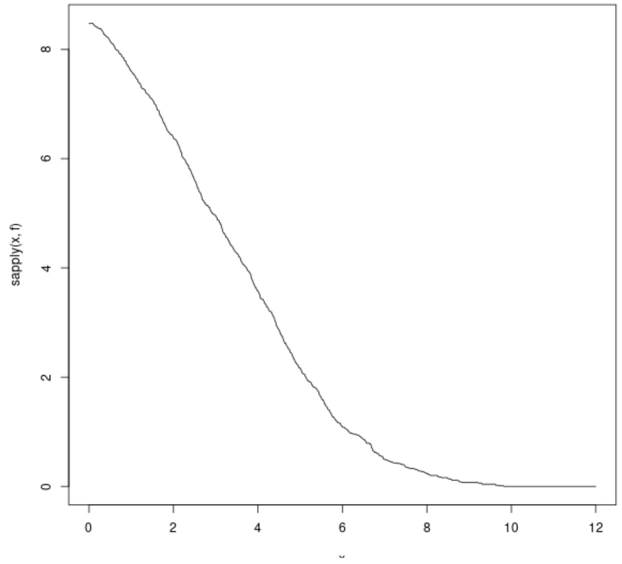
- **Need scheduling information**
  - scheduling jobs to opportunistic slots to kill them not constructive
  - Job run times ~= O(peak load times)
  - else need effective SIGTERM → SIGKILL job wrap ups
- **Need elasticity on the job supply side**

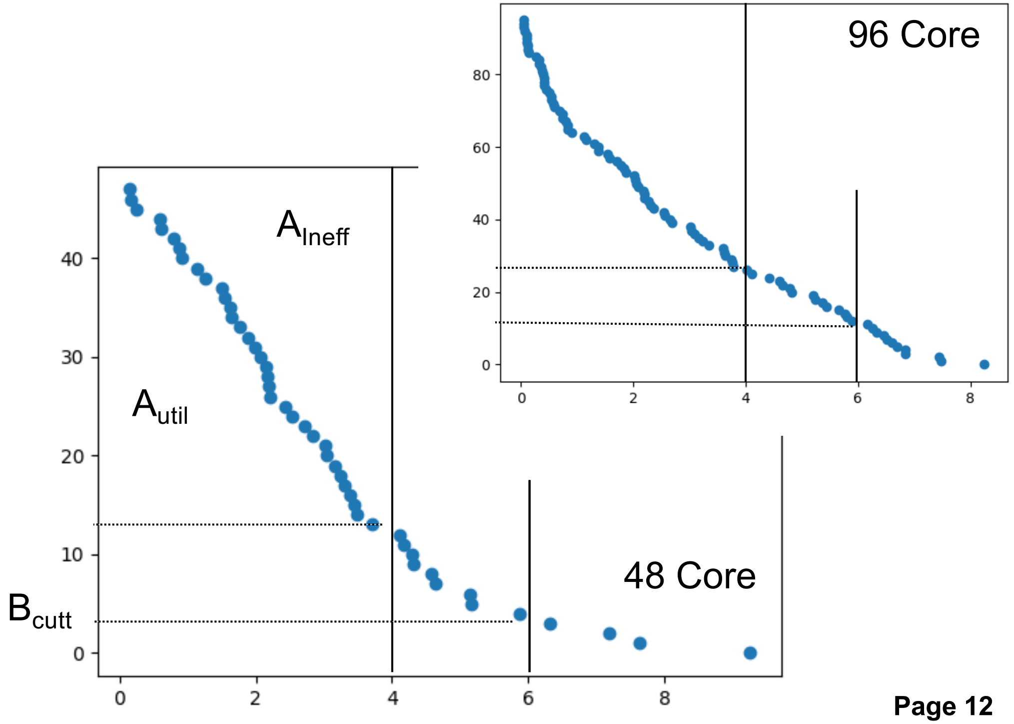
Scheduling aspects

Load Shedding: Worker Draining

Worker Draining Projections

- Without scheduling information only statistic estimates
- Efficiency stochastic draining vs. scheduled draining
- Load shedding efficiency
  - Utilization drain start and shut/cutt off
- Simulation + analytic projections
  (K. Severin, L. Mansur, L. Janssen)
- ATLAS jobs ~= 6h lifetime Gaussian + 2h width
- Old 48 core workers & new 96 core workers
- What draining inefficiency acceptable?
- Going more for vertical scheduling?





Making scheduling green

- Partitioning cluster in new and old hardware, and switching on/off older hardware depending on availability of RE is more efficient than throttling the whole homogeneous cluster
- Our goal: Make the “old hardware cluster” as efficient as possible
  - We need information about job lifetimes → we have to do real scheduling
  - Or jobs that can finish themselves within a defined TTL → experiments have to do real scheduling based on our TTL
  - Or jobs can stomach preemption without significant workload loss (*)
  - Multi-VO sites need a standardized solution!

- (*) Opens the way to opportunistic use of other resources!

A broader view

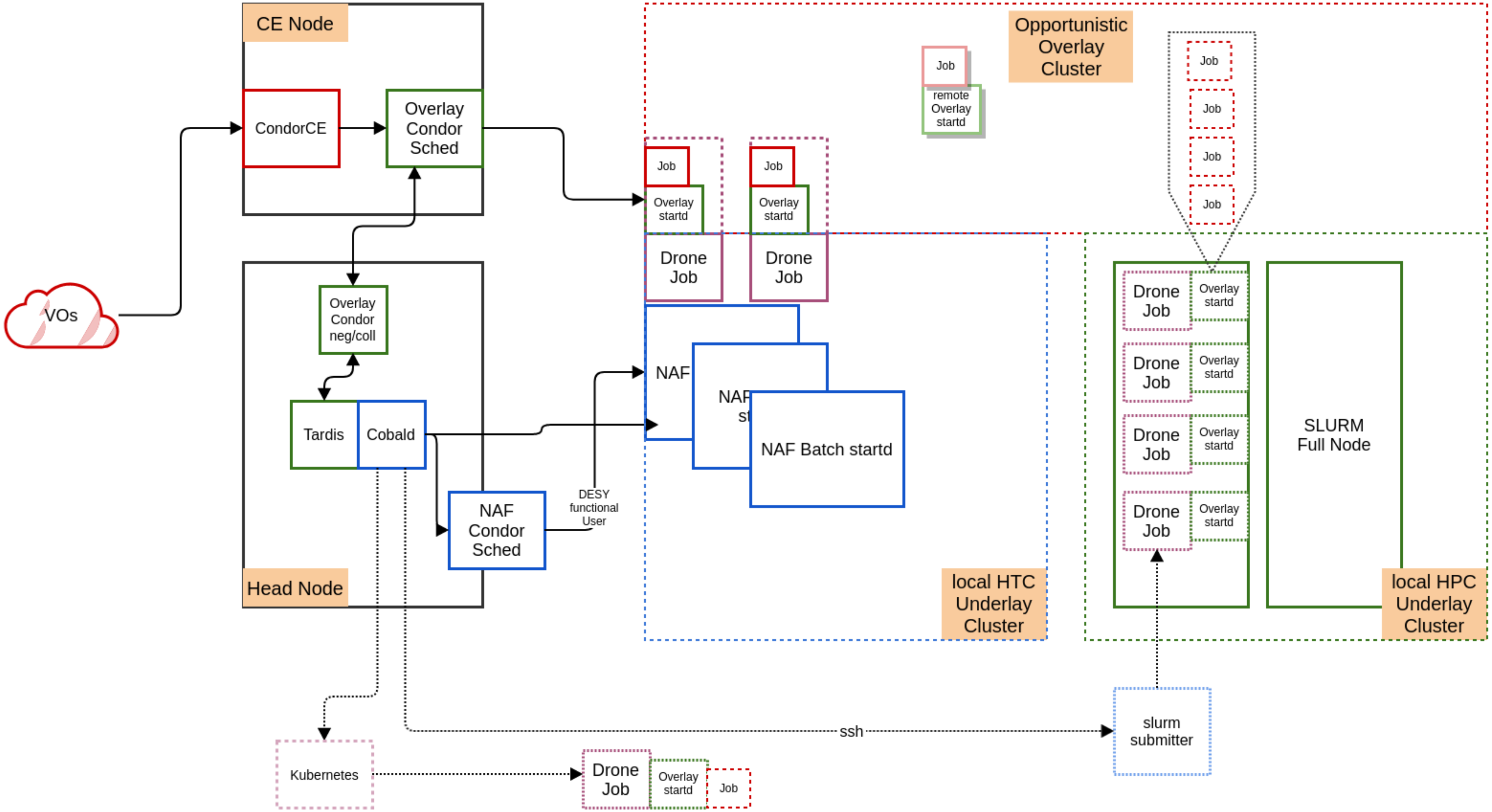
DESY infrastructures: Use cases and plans

- Grid: Basically 100% occupied the whole time. Partitioning of cluster and adaption to RE availability
- NAF: time dependent usage profile. Increase utilization by vertical filling and load shedding. Add dependency on RE availability.
- Maxwell HPC: time dependent usage profile. Singe node scheduling. load shedding. Add dependency on RE availability.
- dCache: Investigate in more efficient CPU utilization. Can we think about cold storage (aka. tape) … and an associated data management plan?
- General:
    - ARM seems to be more efficient. Are the VOs prepared for ARM deployment at sites?

Appendix

Opportunistic Resource Utilization

Draft Dynamic Overlay Cluster



Worker Frequency Scaling

CPU Govenor Scaling vs/ Sub-Clusters

- Zen only three freqs with 3.10
- Idle offset ~150W
- Normalized to HS06 benchmark runs
- Efficiency sweet spot at mid freq

- recap: 1000 kHS06 delivered
- “uncapped combo” cluster: ~410 kWh
- “min freq combo” cluster: ~419 kWh

- “high efficiency” cluster: ~298 kWh