Towards more Energy Efficient Compute Clusters

Experiences gained over the last year

DESY IT

Christoph Beyer, Stefan Dietrich, Martin Flemming, Manuel Hamurculu, **Thomas Hartmann**, Andreas Haupt, Yves Kemp, Maximilian König, Oliver Krüger, Krunoslav Sever, Alexander Trautsch

Hepix Spring 2023 Workshop
Taiphe, 2023.Mar.30

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

DESY.

The Last Year in Energy Worries

Development of European Energy

Situation one year ago in Europe

- Energy supply appeared fragile
  - dependencies on problematic external sources (gas, nuclear fuel,...)
  - nervous energy markets = significant higher prices

Stabilized situation today

- Climate Crisis becoming critical
- power outages have not realized
- high(er) electricity prices are to stay
- regenerative energy sources becoming much more significant
  - depending on regional conditions with short/mid term fluctuations
  - avoiding strategic dependencies with long term impacts

DESY Energy Usage

Hamburg and Zeuthen sites

- Hamburg Computing Centre
  - minor consumer compared to accelerators
  - but significant energy sink with respect to general consumers (~700 households)

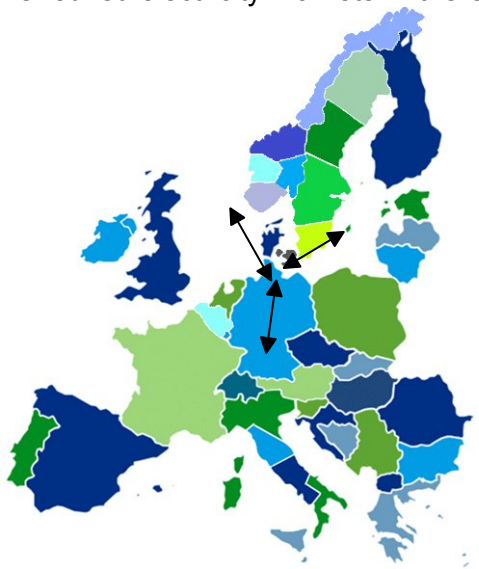
- Zeuthen Computing Centre
  - major consumer with respect to the whole site

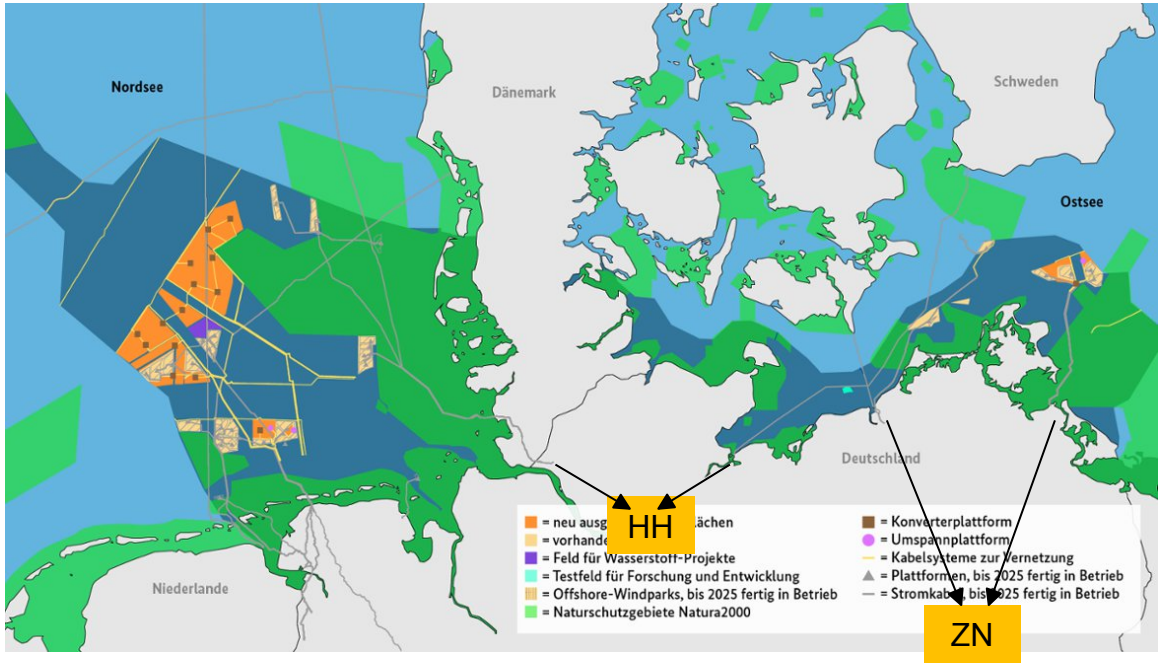
Regenerative Energy Sources

Nothern Germany Green Energy Supply

- offshore farms in Northern & Baltic Sea
- common German biding zone
  - Limited transport capacity between nothern and southern Germany
  - times of abundance wind energy not efficiently translated into cheaper regional prices
  - energy deficits in southern Germany when wind produces cheap green electricity in the north

European energy market bidding zones
https://fsr.eui.eu/electricity-markets-in-the-eu/





https://www.offshore-windindustrie.de/windparks/deutschland

Green Energy

Supply Implications for Nothern Germany

- become more efficient in the utilization of energy
  - react to lower & higher prices
  - reducing energy waste
  - better utilize the clusters and cut idle slacking

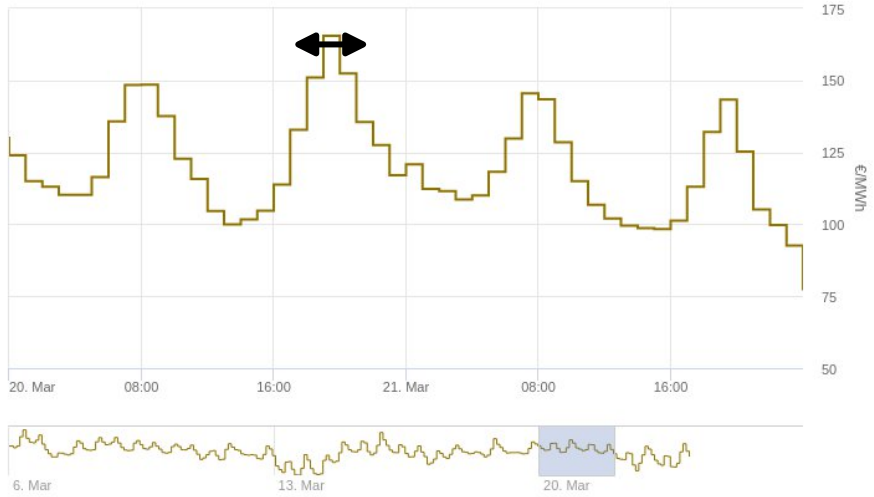
- become more flexible in the energy consumption
  - **short term:** hourly fluctuations
  - **mid term:** general weather conditions
    - reduce usage during lulls
    - increase usage during abundant green energy
      - “burning off” glut of cheap green energy
      - relieving energy market in the south

Short Term:
HH CPU Frequency/Energy Shaping

Short Term Optimization/Reaction

Dynamic CPU Scaling

German €/MWh market prices in 15m intervals 2023.Mar.20-21
https://www.smard.de/page/en/wiki-article/5884/5976



- current electricity contract: fixed price slice plus market price component
- how will contracts look like in # years?
  - more dynamic or still mostly fixed for large consumers?
  - e.g., DE day ahead prices last week [50€/MWh,200€/MWh]
    - fluctuations during the day

- ATLAS suggested to use CPU throttling to temporary lower energy consumption [↩]

- prepare for short term reaction to pricing peaks in O(15m)
- feedback mechanism required from supplier/local market conditions

Short Term Optimization/Reaction

Peak Usage Smoothing

- throttle CPU frequency governor to minimum frequency during peak minutes
  - stretching jobs over time for reduced momentary power load
  - transparent to Grid pilots/payloads
  - HS06/Watt similar for Zen2 for max (2.85GhZ) and min (1.5 Ghz) frequencies (3.10 kernel limited Zen support)
    - “uncapped” cluster: ~410 kWh for 1000 HS06 time 1x
    - “min freq” cluster: ~419 kWh for 1000 HS06 time ~2x
  - ~> cluster efficiency not affected - just more/less green depending on the power sources

- Questions
  - multi-VO site: Site decision to throttle? VO quorum on throttling?
  - how to account? pledges?
  - todo: Zen arch frequency stepping with newer kernels?

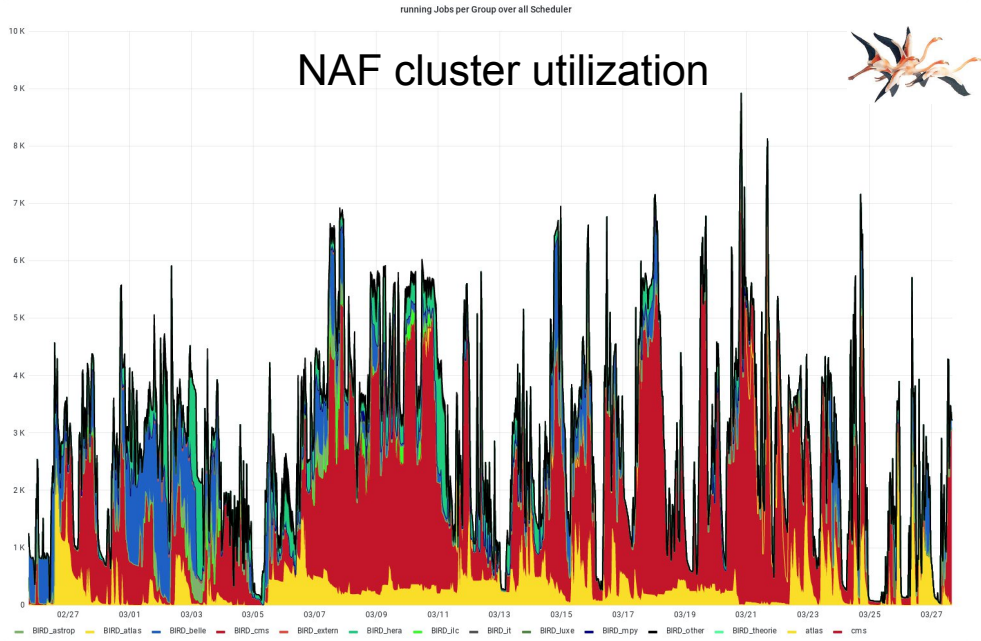
Mid Term:
HH NAF Energy Shaping

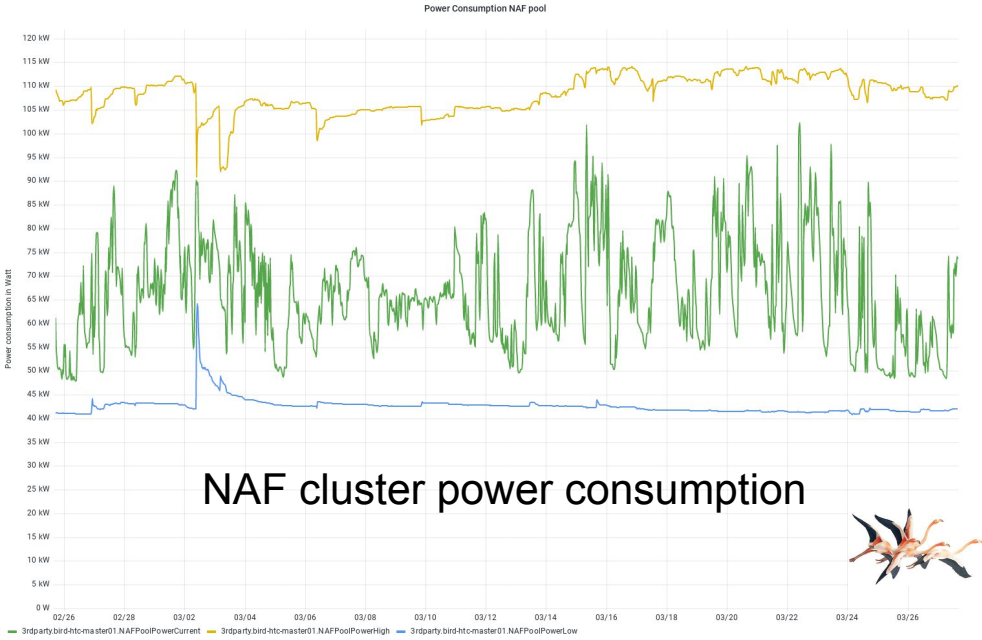
National Analysis Facility

Power Utilization Optimization

- NAF cluster: open to German HEP User
- HEP Users: job patterns ~workdays + larger trends(aka conferences)
  - Direct energy optimization potential!
- NAF power consumption follows the cluster utilization

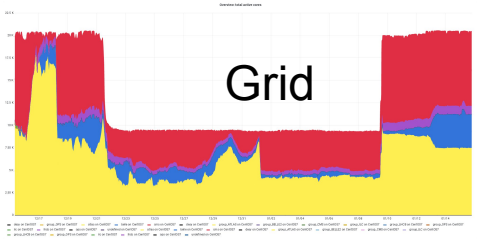
- horizontal -> vertical scheduling
- load shed not-requested cores/nodes



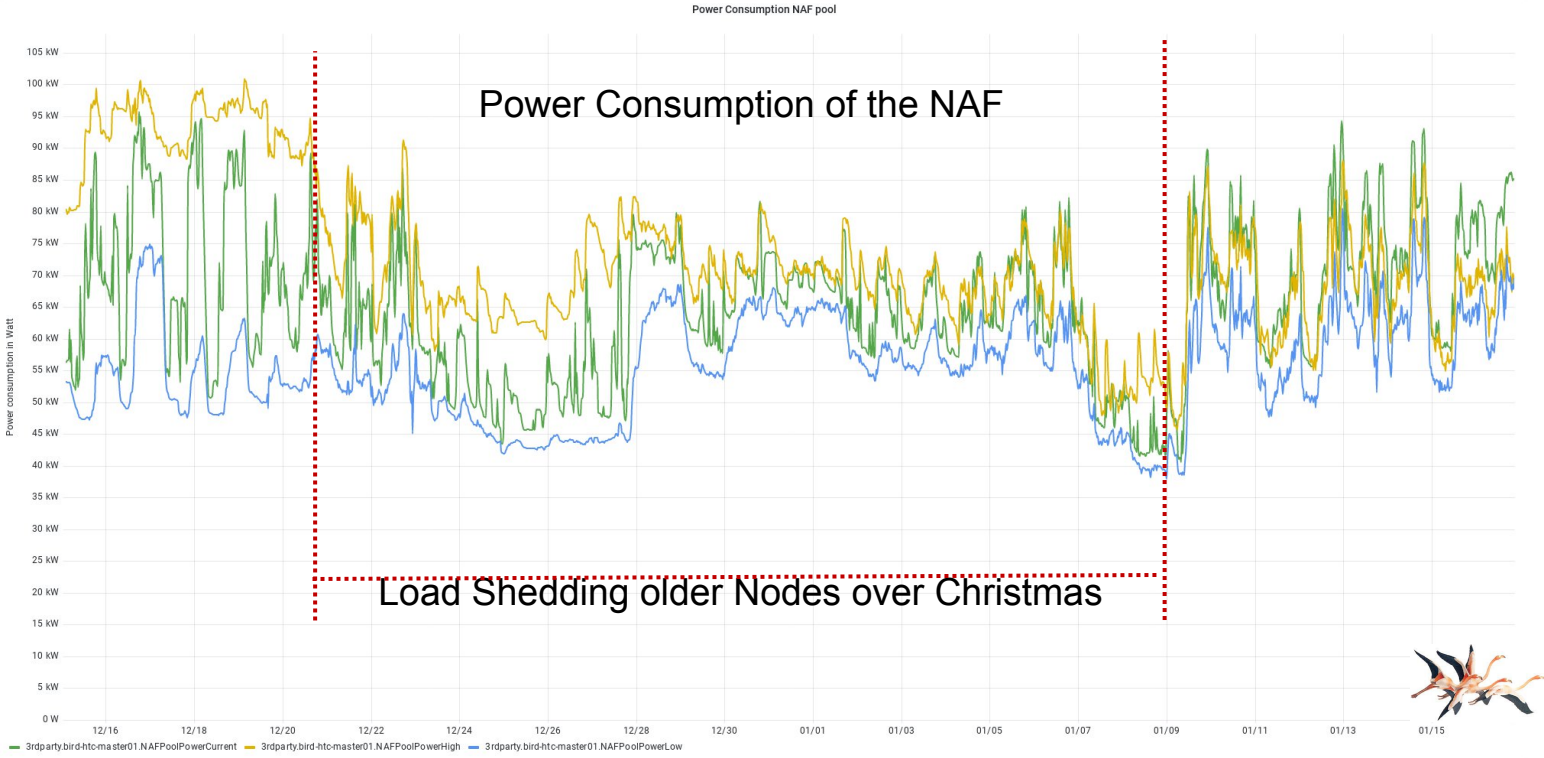


Reduced Power Consumption

Over x-mas



- PoC: shutting off old nodes over the Christmas break
- NAF Users lazy over the holidays
- Grid: shedding ~40% of HS06, saved ~60% of energy
- ~35 MWh saved over two weeks



Things we learned on the way & future plans

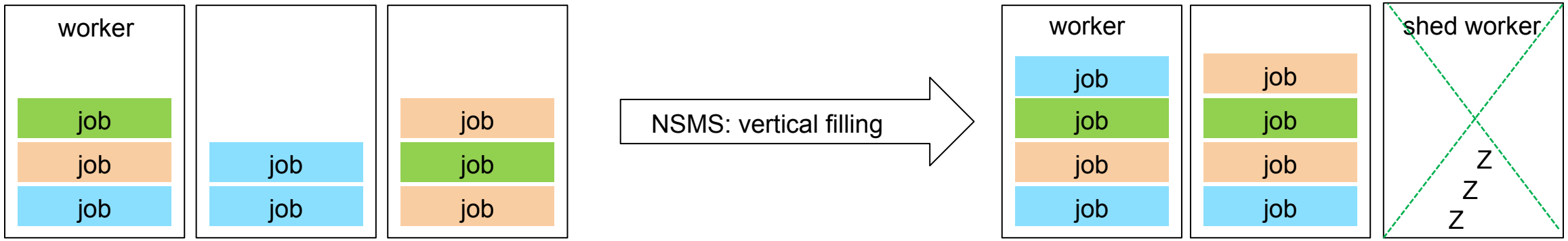
it’s complicated & idle workers are evil

- Measuring power consumption can be tricky
- The base power consumption is much higher than we anticipated
- We developed a reliable monitoring of actual and possible max & min power consumption
- Older machines use more power in every sense
- Idle cores and even more idle worker nodes are a big luxury
- Horizontal filling of the pool is to be reconsidered
  - was debatable before as more the jobs are spread the more IPs are stressing the storage
- We need a more dynamic worker node management in order to shut-off idle worker nodes and ‘wake’ them when needed
  - Preparation is done and a set of test-workernode behaves accordingly to the plans
  - Some integration in monitoring etc. needs to be tested

Future plans

Making the NAF more sustainable

- Make sustainability of worker nodes a hardware fact and part of the host classadd (HS06/W)
- Horizontal filling of the ‘sustainable’ part of the pool
- Opportunistic usage of the ‘*not-so-much-sustainable*’ aka ***NSMS*** part of the pool
  - vertical filling and little headroom
  - shutdown idle worker nodes (maybe hibernate once more recent hardware becomes part of the NSMS part)
  - use ATLAS GRID jobs for backfilling NSMS part of the pool in case of green energy being available
  - fullfill all pledges at any time with the sustainable part of the pool



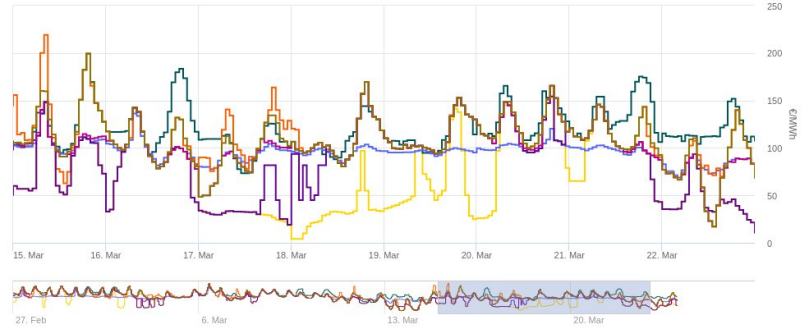
Mid Term:
HH Grid Cluster Energy Shaping

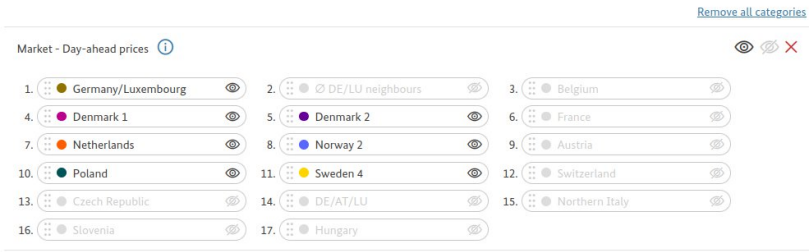
Efficiency Optimization

Optimizing More Compute for the Watt

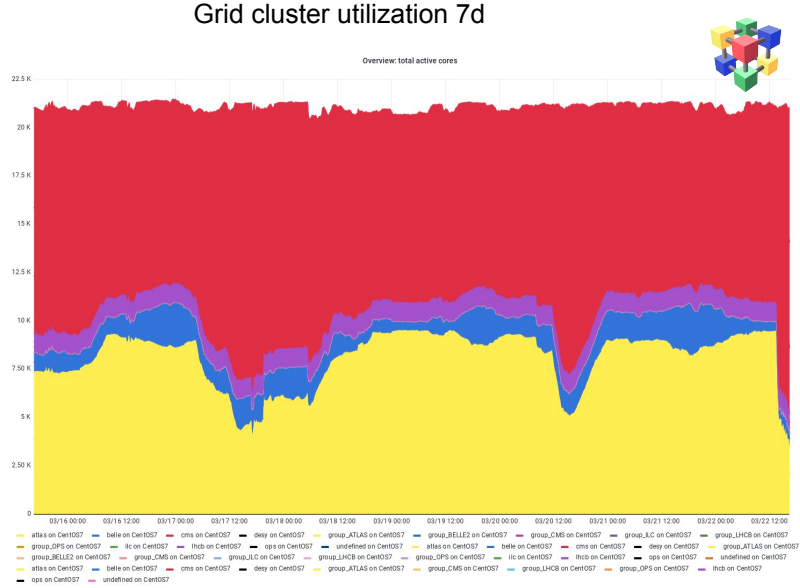
- Grid complementary to NAF
- Grid cluster: production 24/7
  - Cluster utilization optimization when already ~100% utilized?
- -> Power source orientated optimization
  - Efficiently(!) load shed when expensive non-green power dominates
  - Opportunistic fill idle resources when cheap green power abundant
- Power day ahead prices fluctuating ~O(day)
  - Still mostly fixed prices – but situation in X years...?

day ahead market prices Germany and neighboring countries
https://www.smard.de/page/en/wiki-article/5884/5976





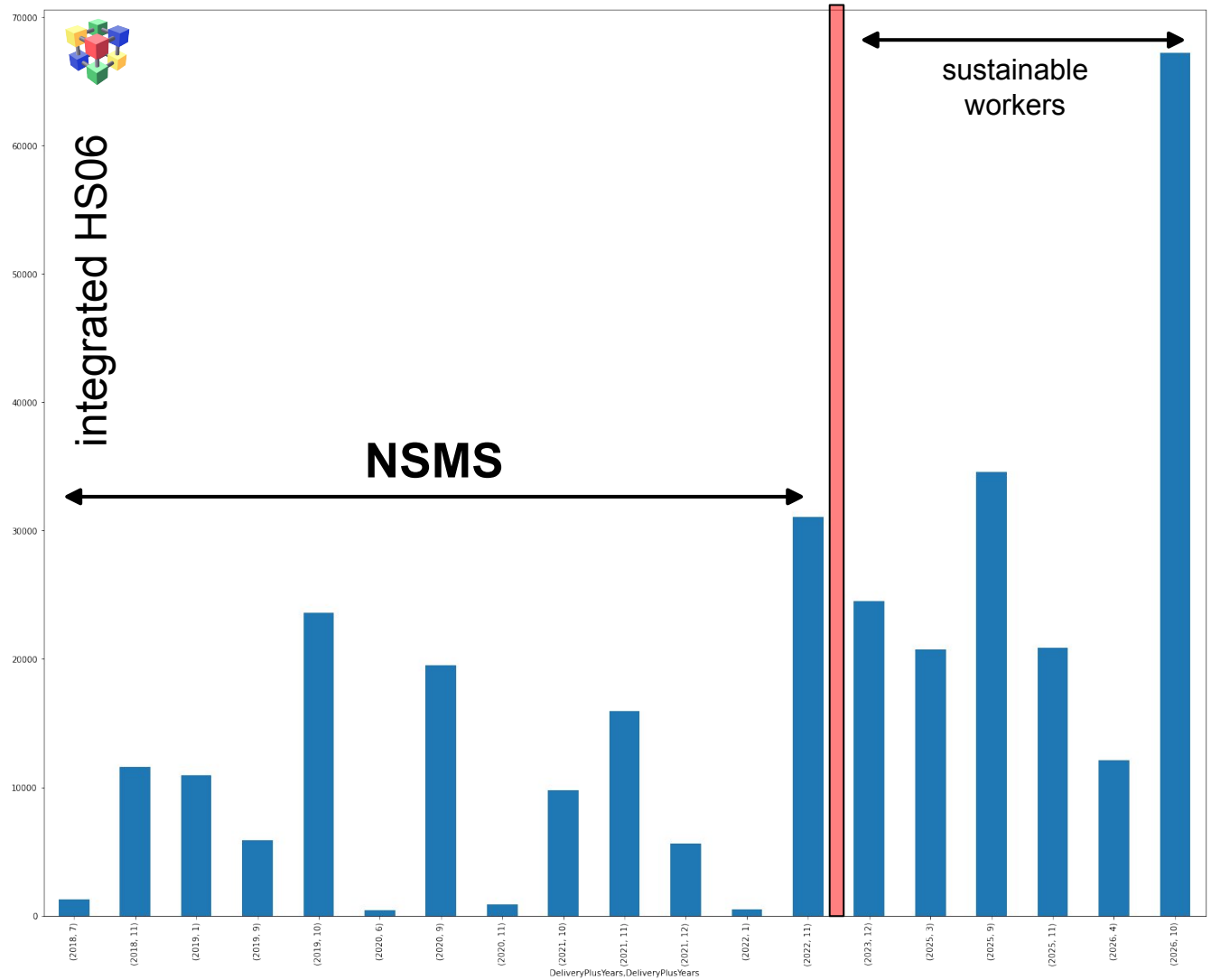
Grid cluster utilization 7d



Cluster Energy Efficiency

HepSpec by Generation

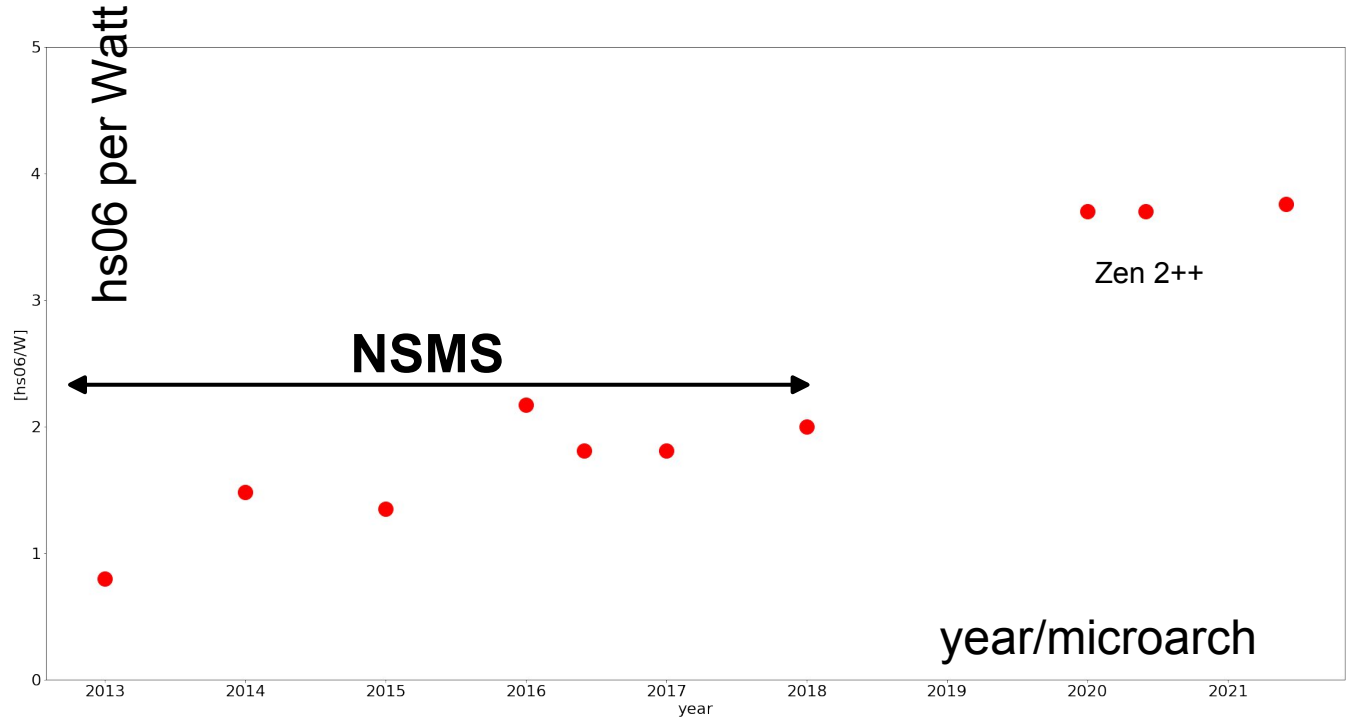
- Grid pledge policy
  - Pledges covered by sustainable workers
  - older generation worker nodes: **NSMS**
  - opportunistically extract green extra HS06s from NSMS workers



Cluster Energy Efficiency

HS06 per Watt consumed depending on architecture generation

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



Cluster Energy Efficiency

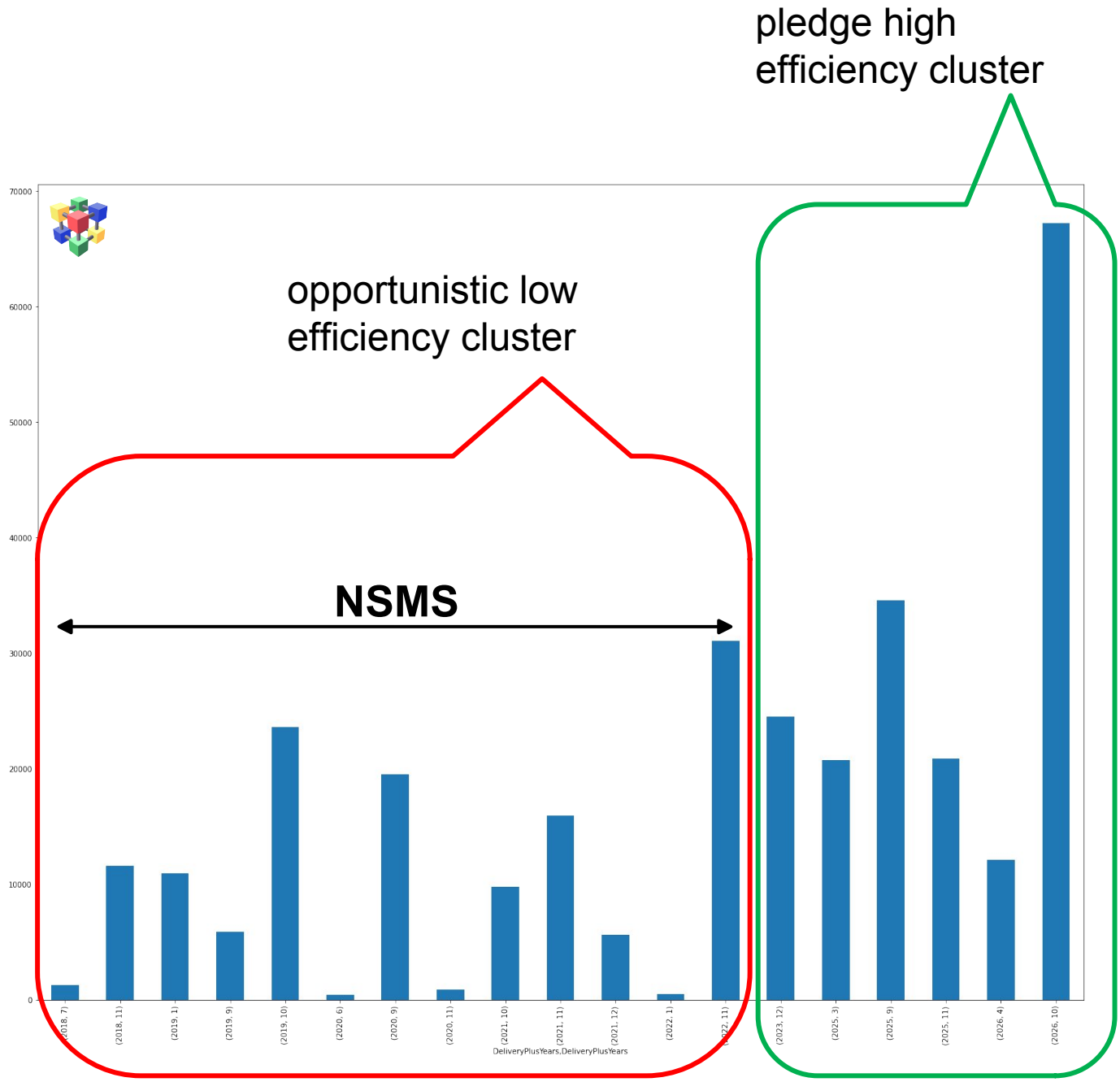
Cluster sub designations

- Reconsider cluster operations with respect to efficiency
- Pledged high efficiency resources always online

- Low efficiency cluster as opportunistic resource
  - Load shedding when necessary
  - Scheduling needs to be adapted

E.g.

- target deliverable: 1000 kHS06
- “combined” cluster: ~410 kWh
- “high efficiency” cluster: ~298 kWh
- “low efficiency” cluster: ~587 kWh

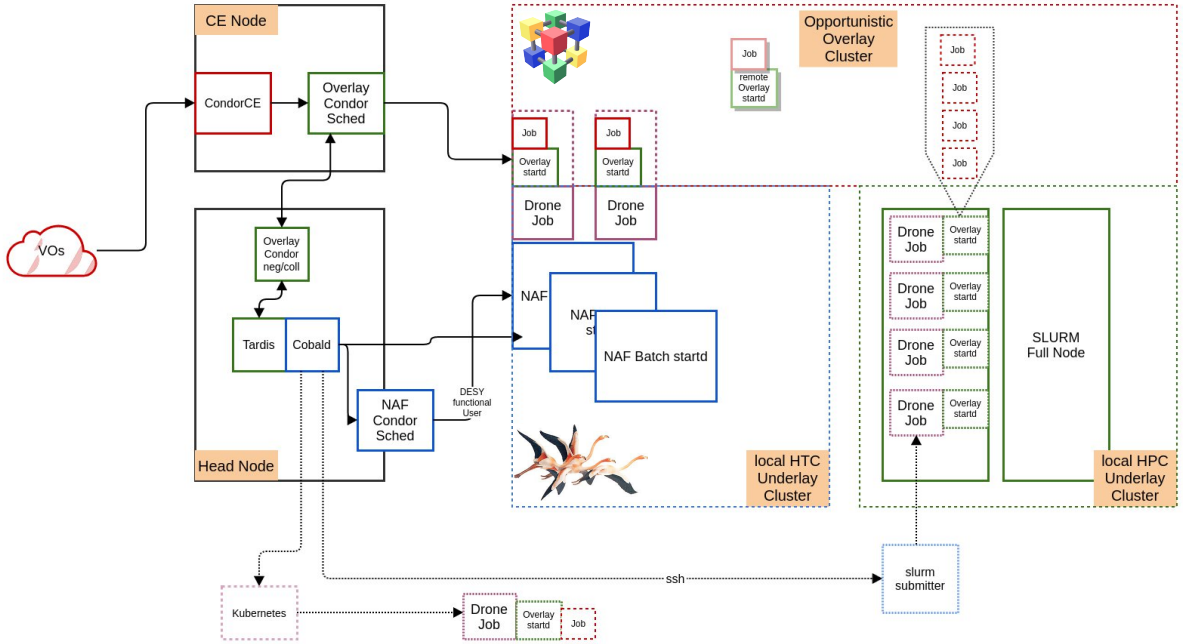


Opportunistic Resource Utilization

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
- Investigating Cobald/Tardis to backfill our other clusters like the NAF with Grid jobs
- **Need scheduling information**
- **Need elasticity on the job supply side**

opportunistic backfilling of local clusters



https://github.com/MatterMiners
https://cobald-tardis.readthedocs.io/en/latest/

Karlsruher Institut für Technologie

Summary

Summary

Making most of Green Energy

- time frame and use case specific optimizations

- optimize NAF cluster resources utilization by condensing jobs
- expand and contract Grid cluster with green energy/prices
  - Short term O(minutes): CPU frequency throttling
  - Mid term O(hours): expand and retract from opportunistic resources
- energy efficiency per hardware generation/architecture becoming more critical

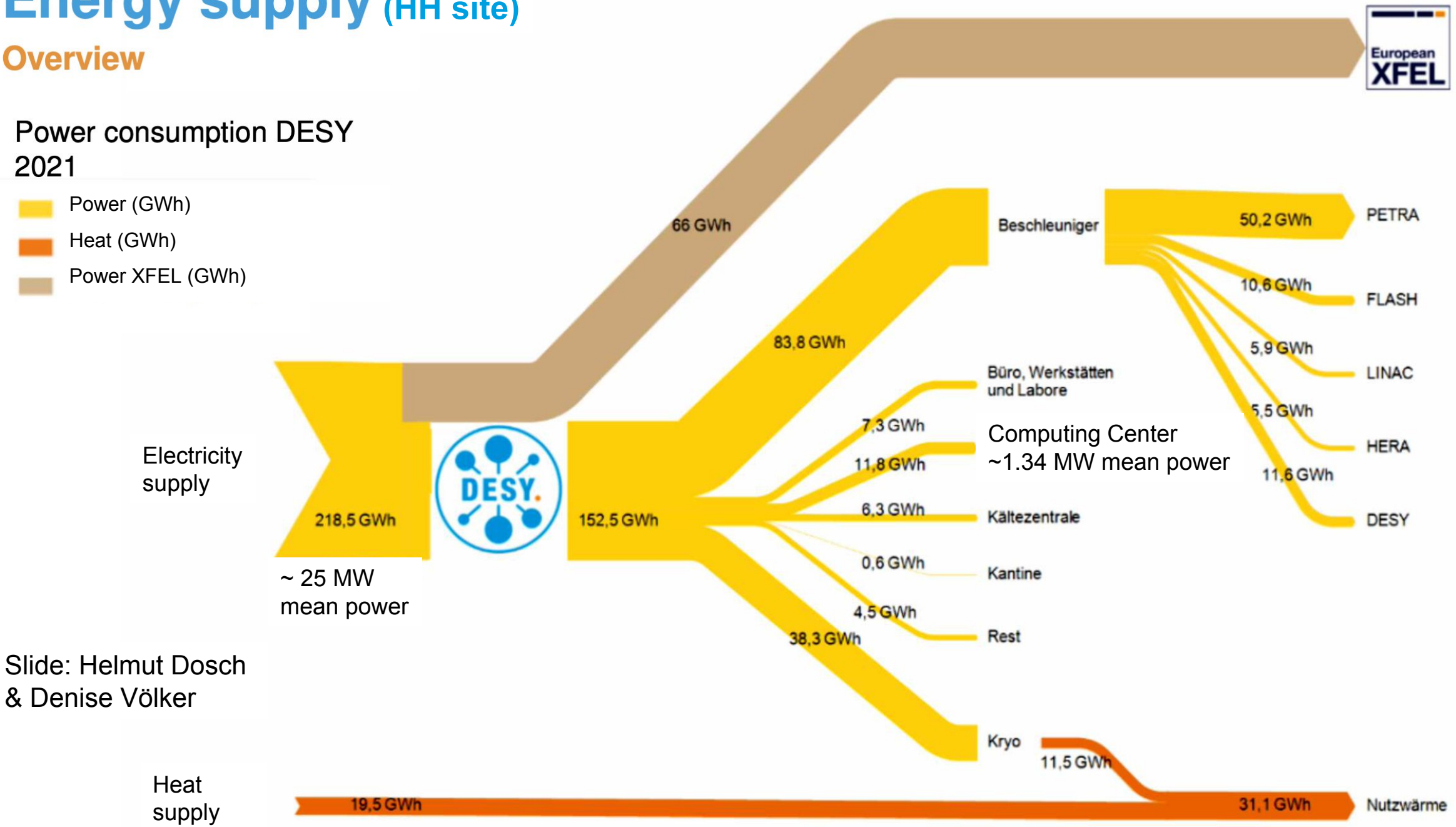
Karlsruher Institut für Technologie

Appendix

Energy supply (HH site)

Overview

Power consumption DESY 2021



Slide: Helmut Dosch & Denise Völker

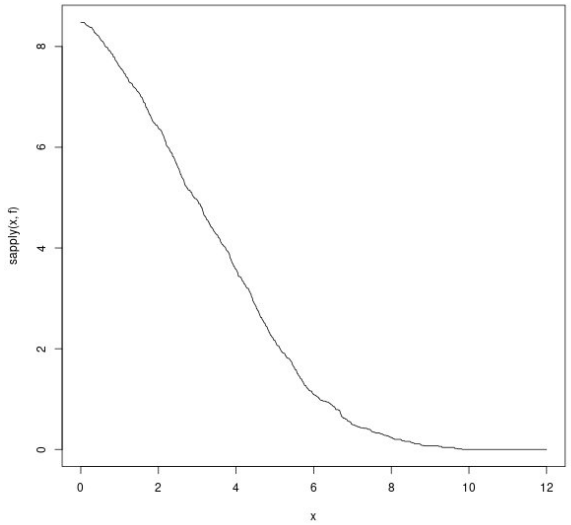
Load Shedding: Worker Draining

Worker Draining Projections

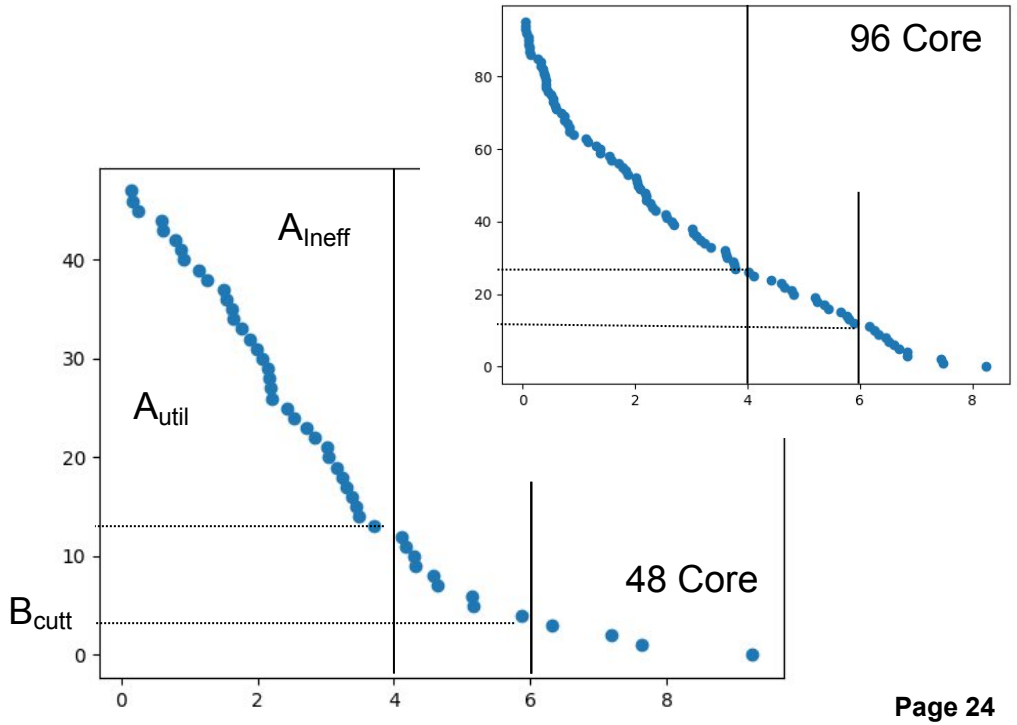
- Without scheduling information only statistic estimates
- Efficiency stochastic draining sub-optimal wrt. scheduled draining
- Load shedding efficiency
  - Utilization between drain start and shut/cut off
- Simulation + analytic projections
  (K. Severin, L. Mansur, L. Janssen)

- Reducing A_{ineff} by scheduling short gap jobs
  - ATLAS already sending short aux jobs

- Next steps: monitoring node shedding turn around and overall utilization



simulated & analytic models for nodes with various cores
assuming jobs with Gaussian runtimes 6h +- 2h



Load Shedding Workers: Low Poer State & Switch off

Hardware Stress

- Workers in low efficiency oow Grid cluster: HDDs
- More HW failures to be expected with more frequent off/on cycles

- First generation with SSD workers entering low efficiency pool end of this year

Worker Frequency Scaling

Appendix