Computer Clusters at DESY

- An overview
- Some looks under the hood
- Energy in all this

Yves Kemp et al., DESY IT
DESY 2.8.2023
Humboldt Highway II - computer cluster on renewable energies

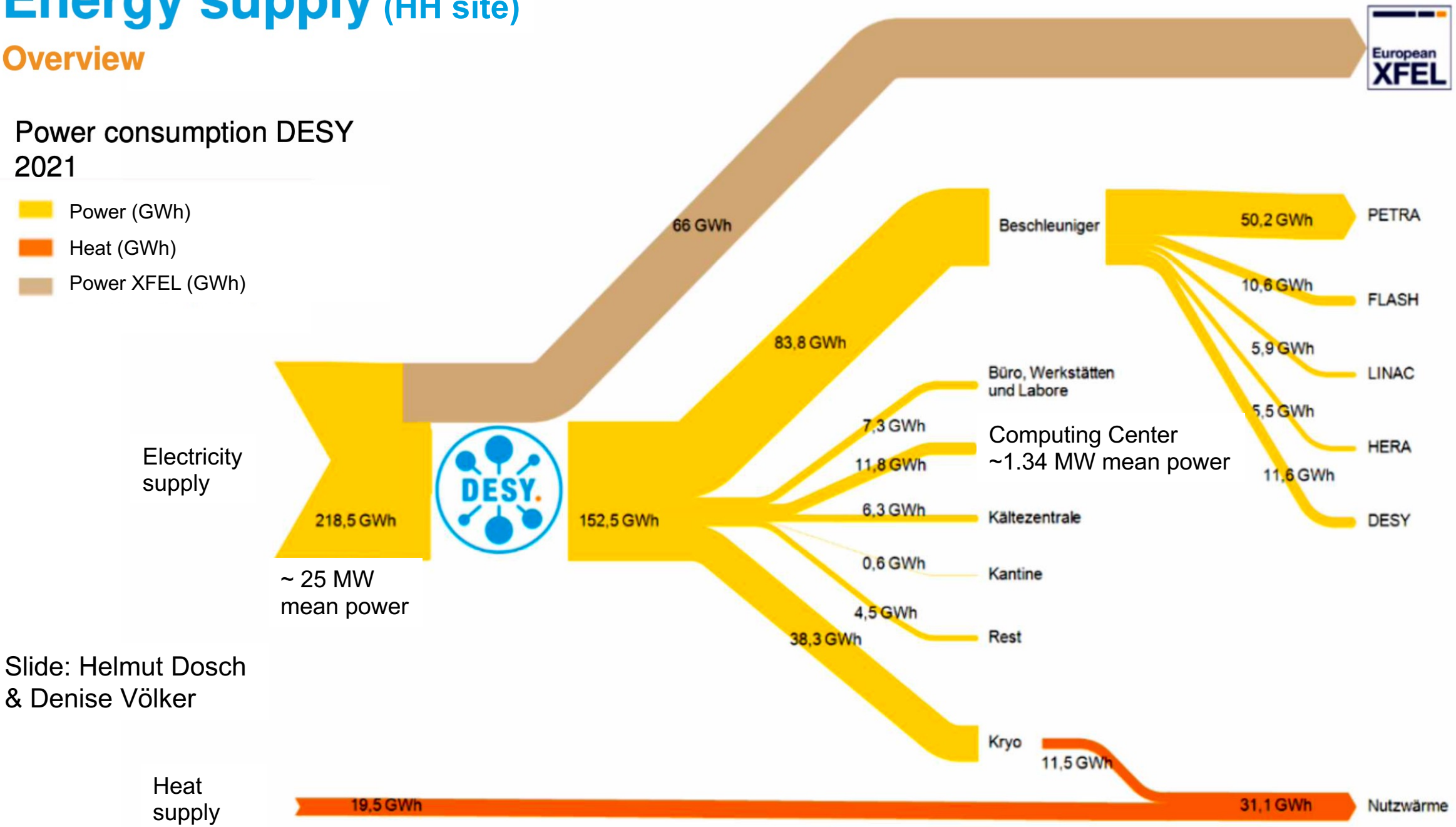
HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

DESY.

Energy supply (HH site)

Overview

Power consumption DESY 2021



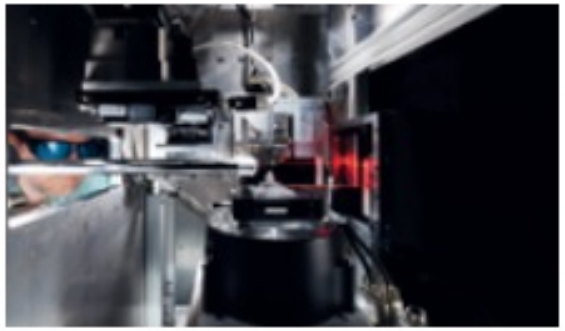
Slide: Helmut Dosch & Denise Völker

- DESY is a large electricity consumer
- Operating accelerators for Science main DESY mission
- DESY IT has some share of electricity consumption
  - most of it for helping DESY in its Science mission

DESY Science in a nutshell

Accelerators

DESY develops, operates and utilises state-of-the-art accelerator facilities. Scientists from all over the world use these facilities to investigate the structure and function of matter.

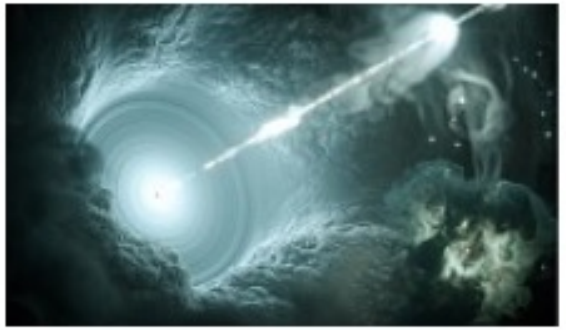


Photon science

Several of the world's best light sources are located at DESY. Their special X-ray radiation makes atomic structures and reactions in the nanocosmos visible.

Particle physics

In global cooperations and large teams, DESY scientists investigate the fundamental building blocks and forces of nature.



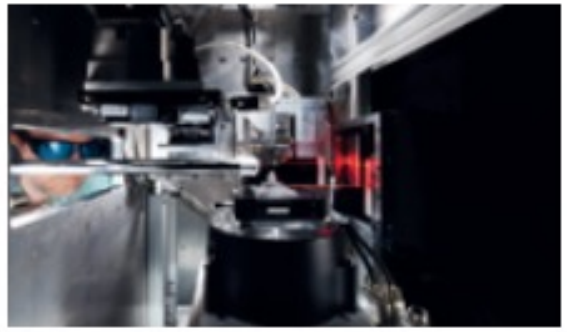
Astroparticle physics

Astroparticle physics uses various cosmic messengers, such as gamma rays or neutrinos, to understand high-energy processes in the universe.

DESY Science in a nutshell – and IT

Accelerators

- IT for Machine operations
- Accelerator R&D

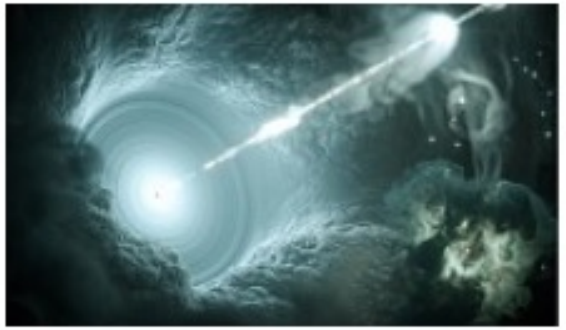


Photon science

- Online data acquisition
- Offline analysis and simulation
- Services

Particle physics

- WLCG commitment
- LHC Tier-2
- Belle Raw Data center
- National Analysis Facility
- Services



Astroparticle physics

- Zeuthen:
- IceCube, CTA, …

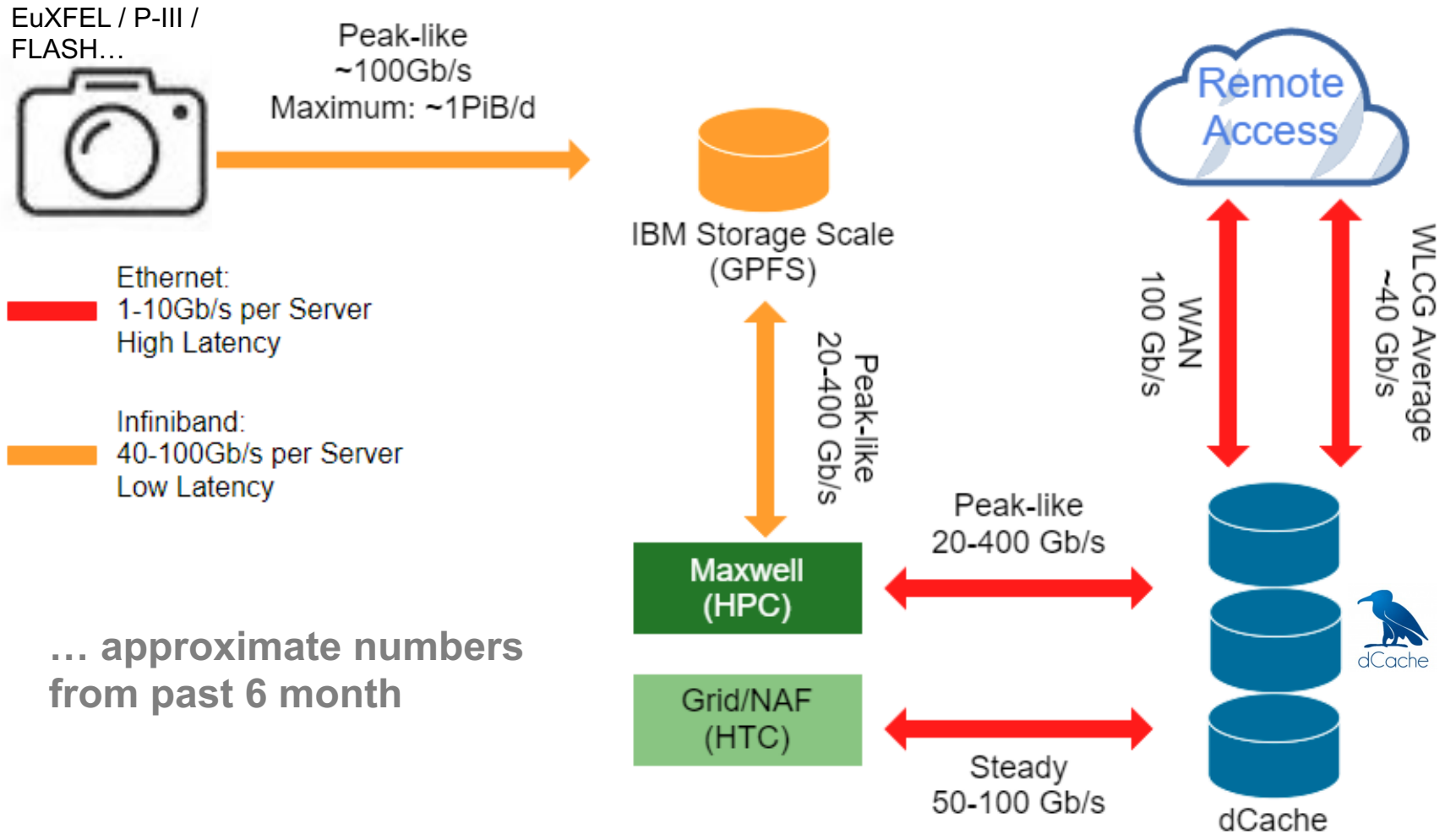
IT instruments for Science: Central Computer Clussters

- Different use cases – differnet clusters for compute:
  - Maxwell HPC cluster
  - Grid cluster
  - NAF cluster
- And storage: dCache computer clusters

IDAF:
Interdisciplinary
Data and
Analysis
Facility

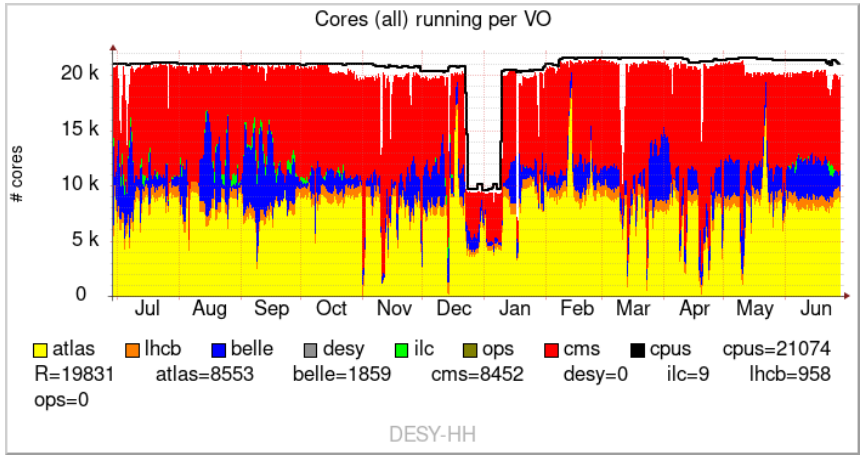
Paradigm: Scientific Analyses are Data Driven

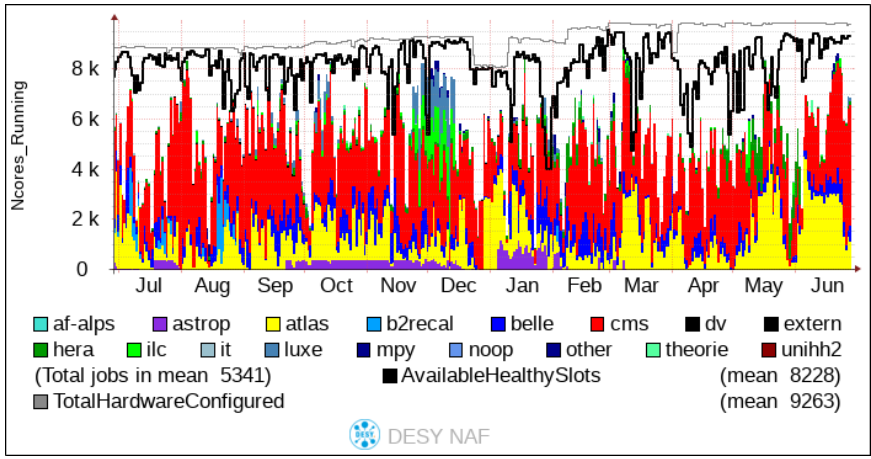
- Example: Traffic pattern in IDAF, approximate numbers from 2023H1



Ressource and usage status IDAF

- **High Performance Cluster: Maxwell**
  - ~900 nodes (inkl. ~250 GPU), ~50k Cores. 2700 users (~1000 active in past 3 month)
  - Storage: GPFS, dCache, (BeeGFS). InfiniBand, SLURM scheduler
- **High Throughput, Production: Grid**
  - 400 nodes, 20.000 cores
  - Storage: dCache, CVMFS. Ethernet, HTCondor Scheduler – Integration in WLCG/Experiment frameworks.
- **High Throughput, Interactive: NAF**
  - 350 nodes, 8.000 cores.
  - Storage: dCache, DUST (GPFS/NFS), CVMFS, AFS. Ethernet, HTCondor Scheduler.





Why different compute clusters?

- Maxwell HPC:
  - Needs dedicated high-performance network: InfiniBand: Inter-process-communication (MPI), GPFS storage access
  - Whole-node scheduling fits best the jobs on these systems (well, most jobs)
  - Buy-in-model, reservations for special purposes (online, …)
- Grid Cluster vs. NAF Clusters
  - Both: 1 or 10 GE Ethernet is sufficient. No inter-node-communication. No GPFS (natively)
  - Both: Smaller jobs … per-core-scheduling
  - Both: Purchase organisation by IT, fairshare model for scheduling
  - Grid: Production jobs, integrated into larger workflow engines … Optimization strategey: Increase throughput
  - NAF: User jobs, interactive or small user production … Optimization strategy: Increase responsiveness
- One size does not fit all
  - Aiming at consolidating as much as possible

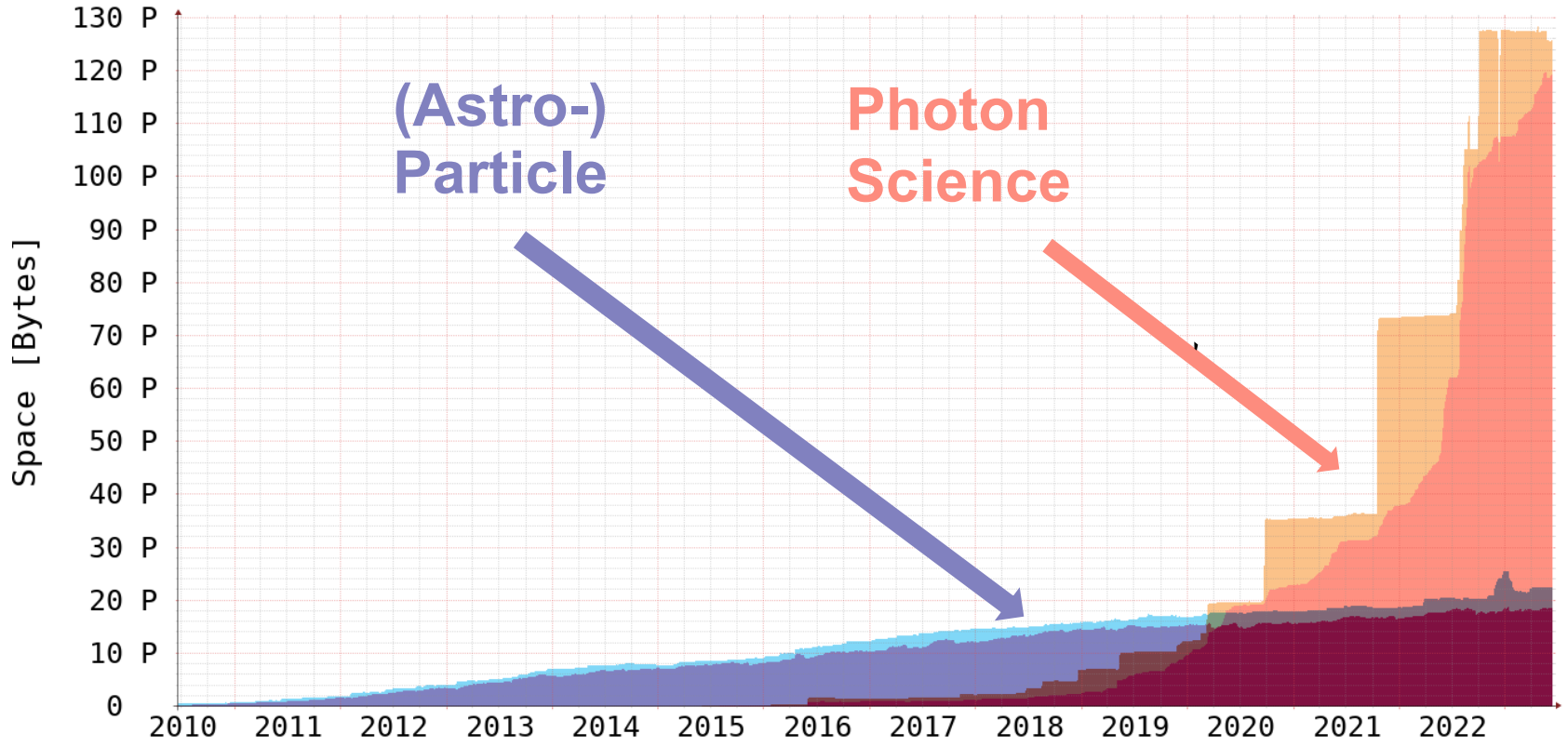
Storage at DESY: Largest working horse: dCache

- The infrastructure view on dCache: Different instances:

January 2023

Instance	Storage-Server	Overall Disk Capacity [PiB]
XFEL	369	109.4
CMS	87	10.1
ATLAS	52	6.0
DESY	34	4.5
Photon	26	3.8
Cloud	22	2.97

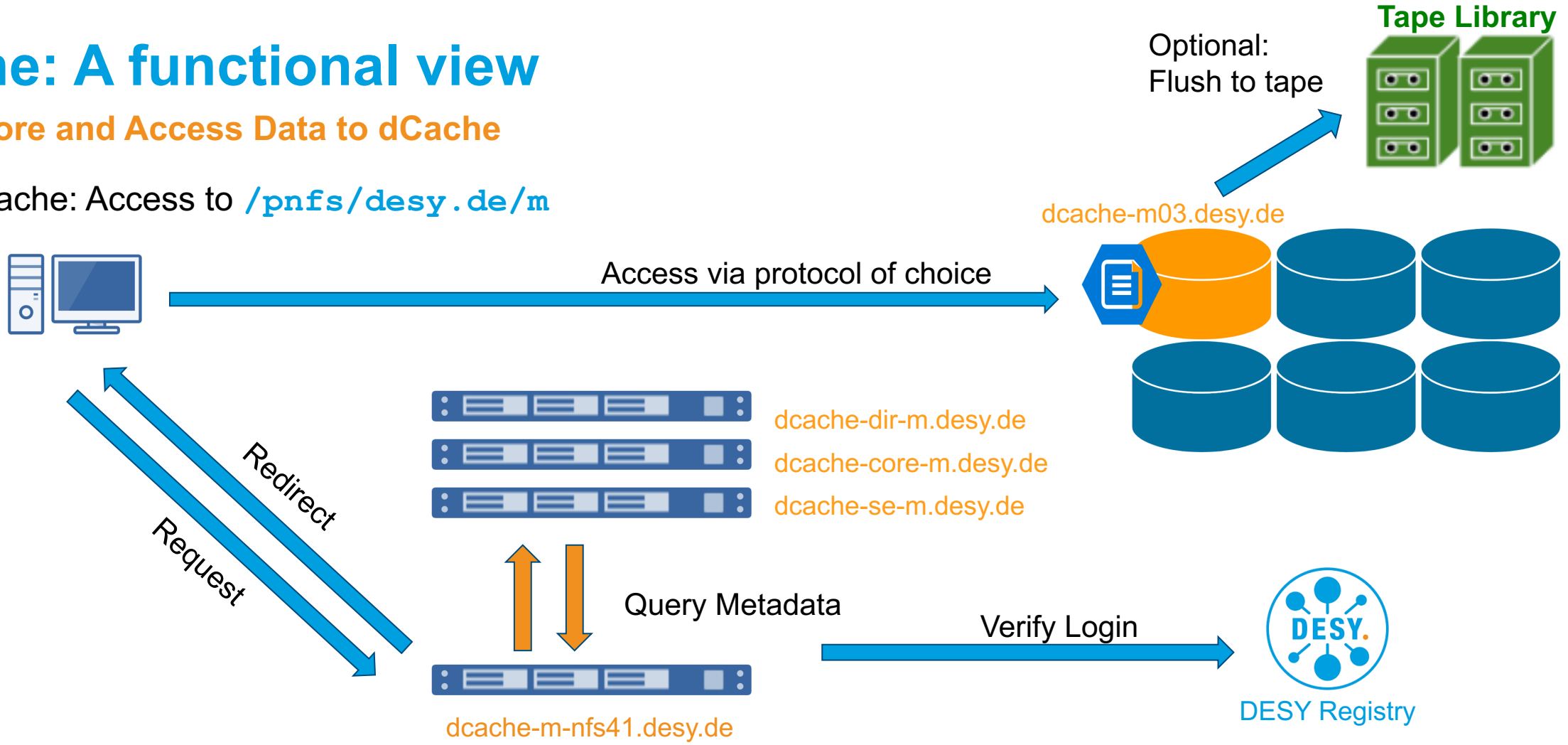
dCache: Historic evolution Photon Science vs. HEP



dCache: A functional view

How to Store and Access Data to dCache

- Use dCache: Access to `/pnfs/desy.de/m`



- Access done through doors: several load balanced door for each protocol to ensure availability
- Access controlled via Grid-certificates or by tokens and POSIX (NFS@NAF and Maxwell)
- Data streamed to/from pool, never through doors: allows horizontal scaling
- Namespace is uniform and independent of protocol

DESY computer management system: Puppet & Foreman

- Why a configuration management system?
  - Easy, reproducible, automated installation of hardware servers → Foreman
  - Automatization, enforcement and reproducibility of configuration → Puppet
  - Audits of configurations, enables distributed development workflows → Git (+Puppet)
- Others tools to achieve the same goal: Infrastructure-as-Code
  - Choice depends on specific requirements

- DESY key facts:
  - Puppet/Foreman managed nodes: ~6.500
  - Developers: 20-30 active
  - 380 modules, 112k LOC (incl. 3rd party)
  - Platforms:
    - RHEL, CentOS, Alma Linux (7,8,9)
    - Ubuntu 20.04 & 22.04 LTS
    - Debian 10 & 11
    - x86-64, ARMHF, POWER

Interface to users: The batch & scheduling system

- Maxwell HPC:

- Grid & NAF:

Common workflow:

- Log-in to a Work-Group-Server (via SSH)
- Do some development, compile, small test job,…
- Submit the larger task to the batch system
  - scheduling of your jobs
  - scheduling of jobs of other users
- The batch system will handle optimal node, optimal time for execution, including specific requirements

Alternative access methods: Interactivity

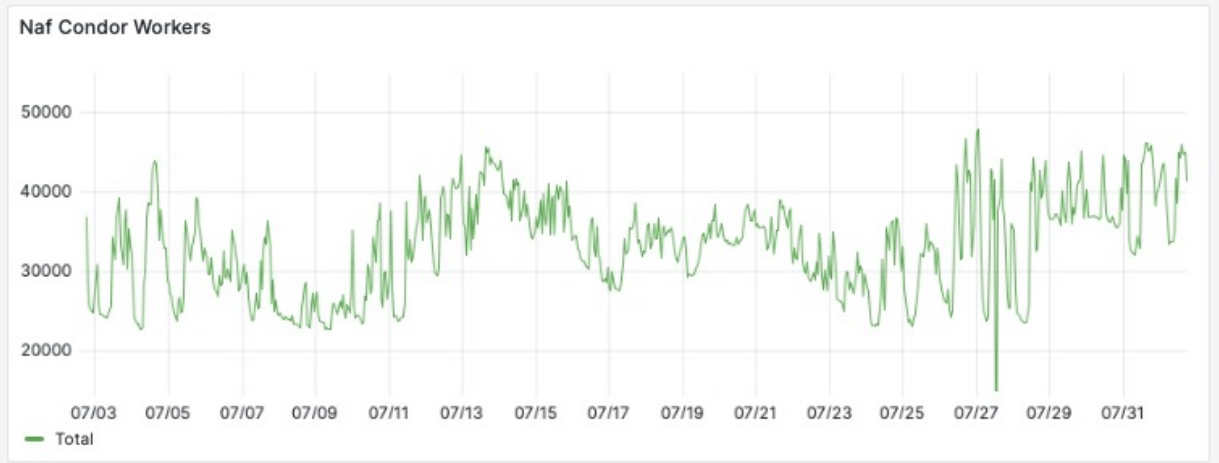
- FastX: Graphical access to WGS

- Jupyter: Integrated into Batch systems

- DASK: Different ansatz: “Memory scheduling”

Cluster: Monitoring

- Extensive node-level monitoring
- Extensive cluster-level aggregation
  - User cores by experiment, ….
  - Recently: Power consumption added



Looking at the energy side of things

- Until 2022: Energy just was there.
- 2022: Awareness increased. Profited from the Christmas holidays to shutdown some machines in a more controlled way
- 2023: Working on user awareness: NAF: Sending weekly reports about consumed time → kWh → CO2
- 2023: Green energy availability is fluctuating. So are electricity prices. First ideas:
  - Shutdown (some) nodes at different times … in a scheduled way, no job losses
  - Throttling nodes at differnet times
  - Shutdown (some) nodes at different times … using preemption (=job losses)

  → More by Thomas and Rod
- Planning of a sustainability workshop: How to best use compute clusters?
- Participating in Horizon Europe Project “Research Facility 2.0”

Energy supply

- DESY (HH) buys so-called futures in tranches and thus minimizes price risk
  - 2022: 100% fixed, 2023: 80%, 2024, 60%, 2025 40%
  - Extreme price developments: Electricity futures for 2023 on 28.9: ~5x working price for 2022
- Additional shortage possible

- Heating:
  - Hamburg: Using district heating, can be reduced by using waste heat (if accelerators are running)
  - Zeuthen: Gas, contract ends 2022, new provider needed for 2023

Image: https://www.tnn.ltd/energy-supply/

Waste Heat Usage 1

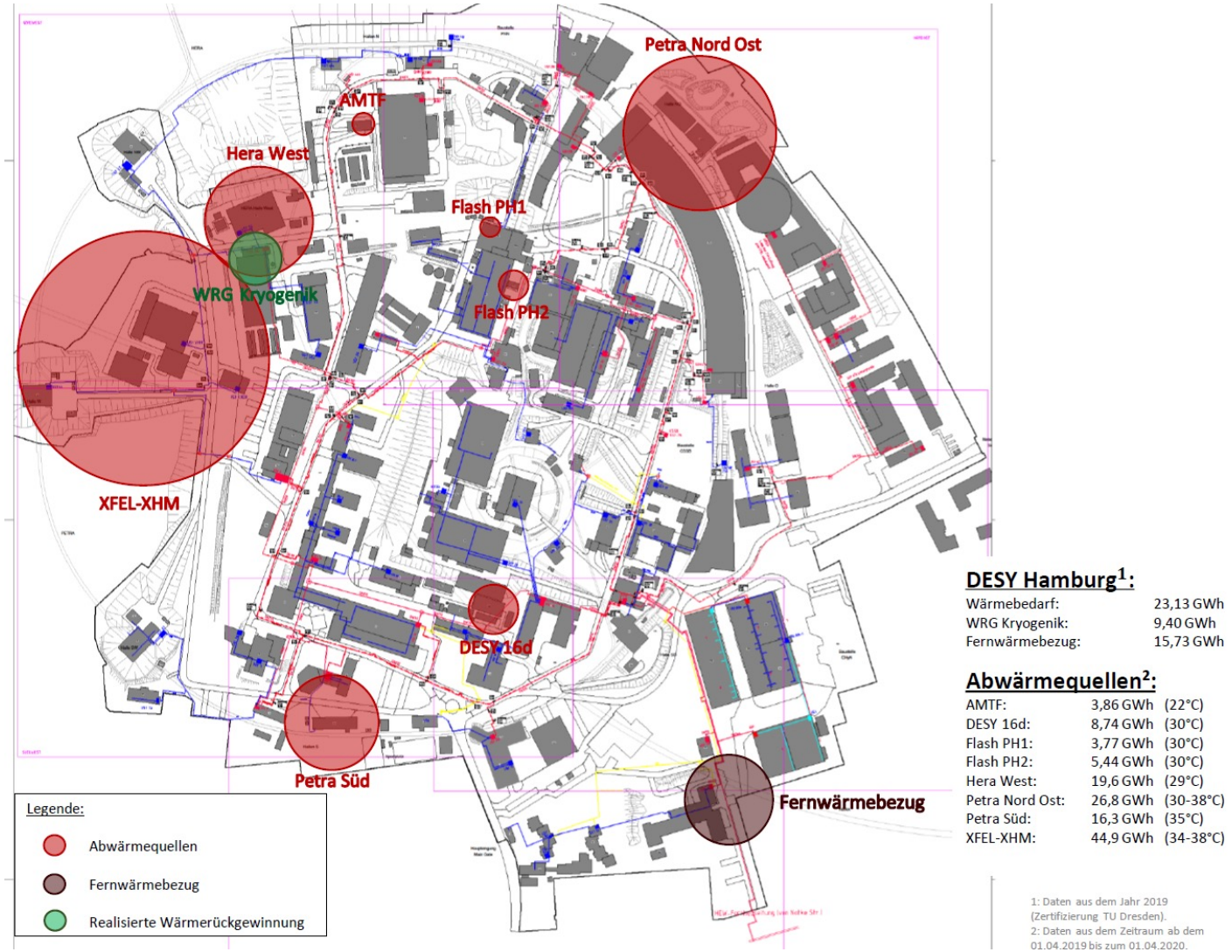
Cryogenic Plant - in use since 2017

- waste heat recovery from cryogenic plant commissioned in June 2017
- helium cooled down to minus 271 degrees Celsius (2 Kelvin) for operation of the superconducting accelerators
- process generates waste heat at a temperature of about 70°C
- two heat exchangers that connect the cooling system of the oil-cooled screw compressors with the local heating network
- Yearly heat extraction of >10 GWh (gigawatt hours)
- >1/3 of the total heat demand at DESY

Waste Heat Usage 2

Unused Potential at DESY Campus in Hamburg

- Currently heating is DESYs 3rd biggest CO2 emission source
- Waste heat from accelerators not yet in use
- project with University of applied science in Hamburg (HAW) to identify potential
- Result: 129 GWh/y of waste heat available at a temperature level of 30°C - 40°C
- Can cover all of DESYs heat demand (ca. 12 GWh/y) – usable in existing and new buildings
- possible CO_2 savings at DESY campus of about 4.000 tons/y
- Surplus can be used in neighborhood; if we get the 129GWh in use saving will be up to 40.000 tons CO_2/y



Summary and outlook

- DESY is experienced in operating large computer clusters
- DESY is well connected with developers of management tools
- DESY is adressing the energy challenge – also in computer clusters
- Looking forward to share knowledge – and gain new insights!

http://dsphotographic.com/photos/cuba-part-i/