Infrastuctures for scientific computing at DESY

Sustainability at DESY-IT

Christian Voß, Martin Gasthuber, Yves Kemp, Philipp Neumann et al., DESY IT

FH Scientific Computing Workshop

Hamburg, 1&2 July 2024

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

DESY.

Users of the IDAF

- **Accelerator Data**

FLASH. Free-Electron Laser FLASH — European XFEL — FF

- **Accelerator Development Data**

SINBAD

- **HPC simulations**
- **Test-beam data**

Detector and Accelerator R&D

- **Facility User Data**

European XFEL — PETRA III — FLASH. Free-Electron Laser FLASH

- **Data of external Partners**

CFEL SCIENCE — CSSB Centre for Structural Systems Biology — EMBL — Helmholtz-Zentrum hereon

Research with Photons

- **Particle Physics Data**

CMS — ATLAS EXPERIMENT — LHCb — Belle II — ALPS II — ilc international linear collider

- **Astro-Particle Data**

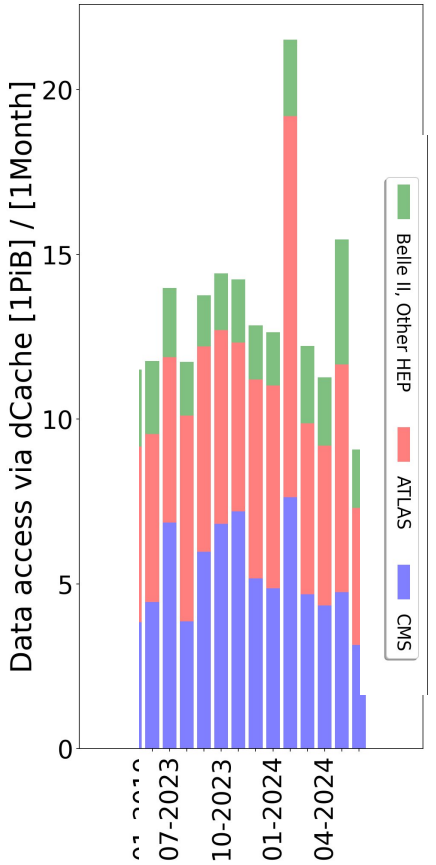
IceCube — H.E.S.S. — cta cherenkov telescope array

Astro- Particle Physics

IDAF: Paradigms … and an umbrella

(some) base concepts

- **DESY science is data intensive, infrastructure is data centric**
- **Generic setup for *all* participants**
- **Integration into experiments workflows**
- **Support & consulting is integral part of the infrastructure**



The IDAF (Interdisciplinary Data and Analysis Facility) is an umbrella over Grid+Maxwell+NAF

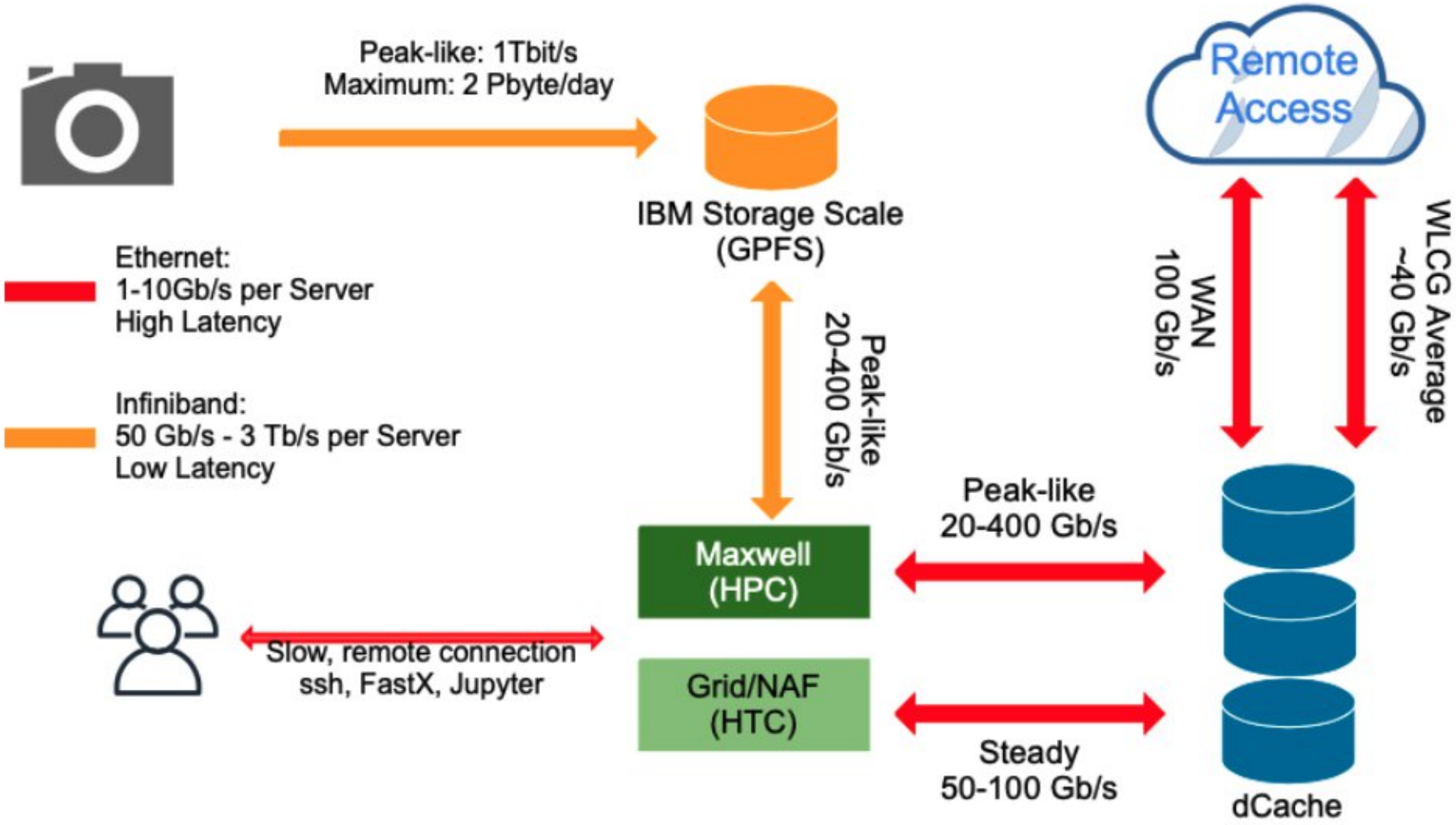
- and an Helmholtz LK-II facility

The IDAF is constantly evolving and subject to adaptions in requirements

- e.g. HEP ←→ HPC integration, new ML/AI workflows, …

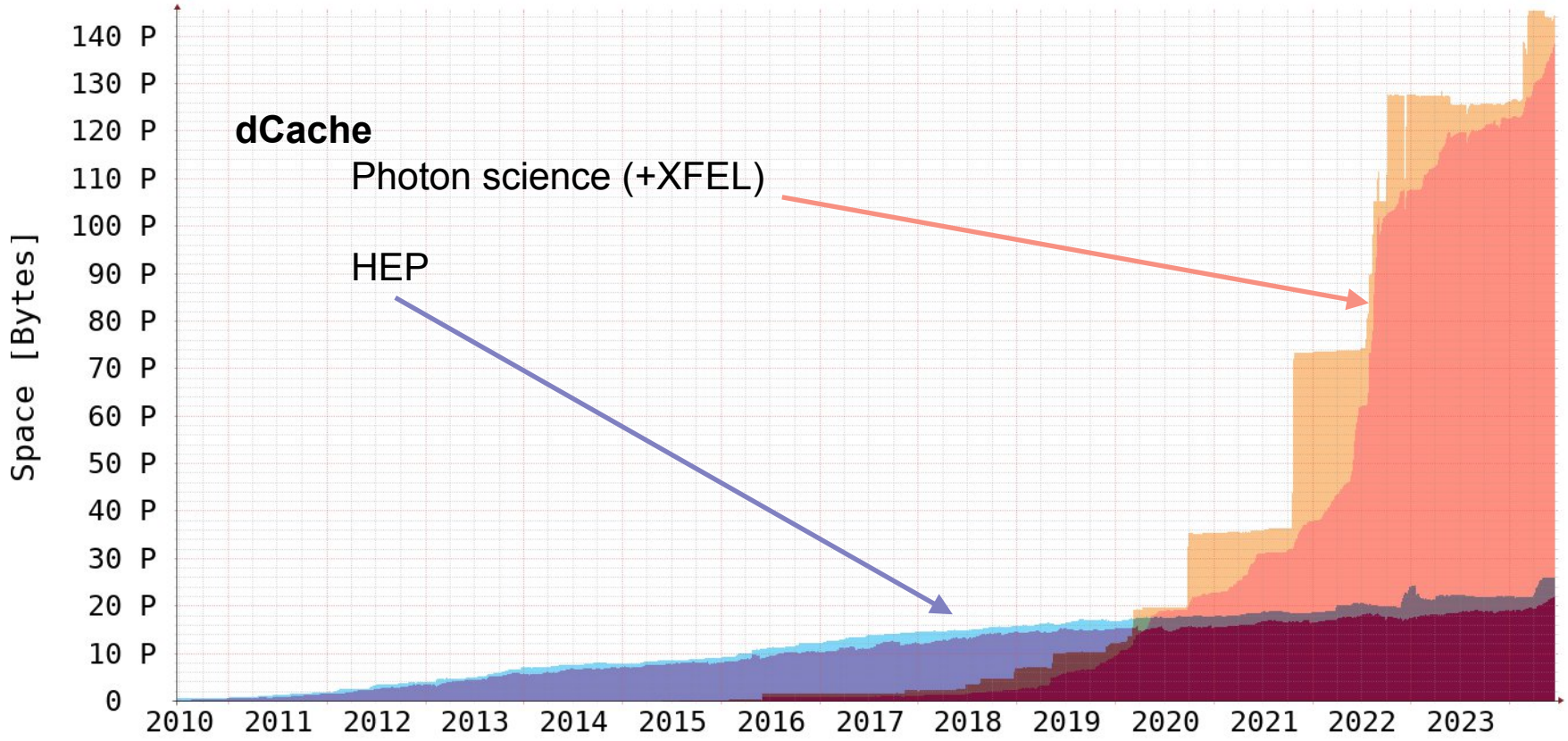
Data flows in the IDAF, and connection to computs

… approximate numbers



Facts and figures June 2024: IDAF

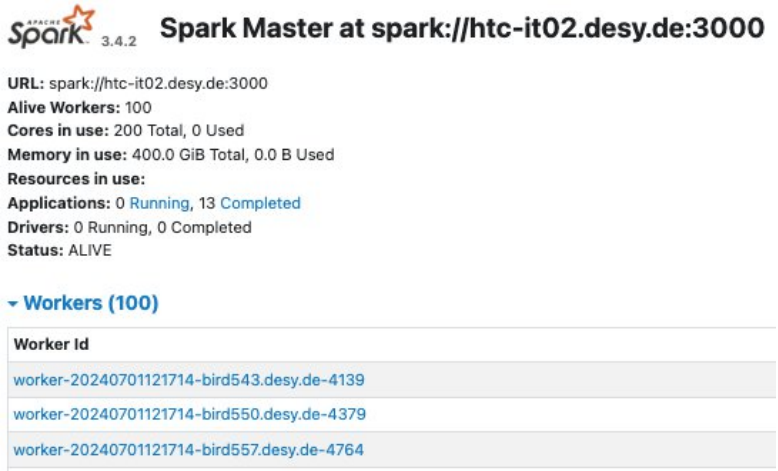
- **Maxwell + Grid + NAF**
- **~180 PB data on disk**
- **dCache + GPFS + BeeGFS**
- **~60.000 CPU cores, ~380 GPUs**
- **HTCondor, SLURM**
- **~2.700 server (compute, storage, management)**
- **~ >0.8 Megawatt (out of 1.3 MW)**



Some ideas for the future

Interactive and Near-Real-Time Data Analysis

- Classic HEP analyses are batch oriented on larger scales with final steps done interactively
- Photon Science has needs for near-real-time analyses
  → react to shifting beams/probe position
- With larger data-sets classic interactive analyses on single WGS get more difficult
- With Jupyter notebooks (integrated into batch), interactive analyses easier to scale – but on limited scale
- Dask / Spark allow for interactive analyses at scale → but detrimental to efficiencies
- Status: Collected experience within IT, integrated into NAF as batch deployment.
- Find interest from the other FH groups



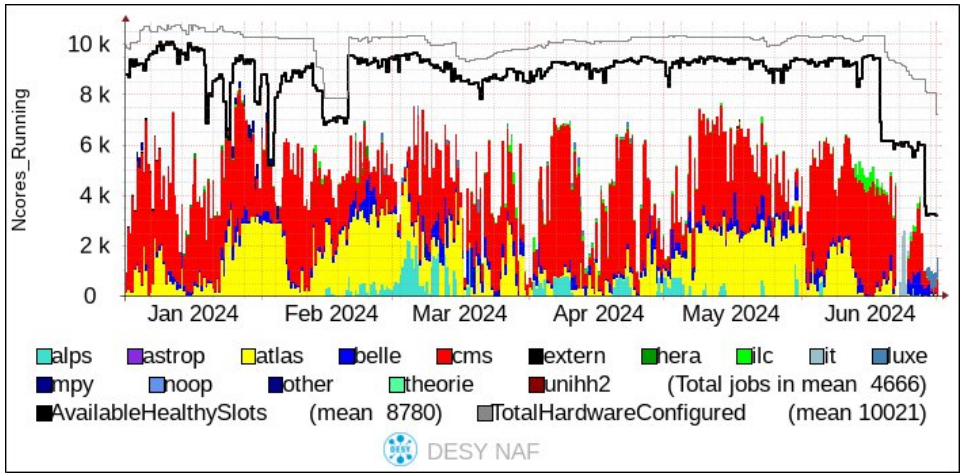
ALPS / (Baby-)IAXO / MADMAX: Data taking and analysis

- Small compared to LHC experiments, but challenging computing requirements
- Cannot draw from well established setups and external contributions

ALPS II

- Integration into IDAF data management and analysis required person-power from the smaller experiments & IT
- Data-format based on tool-set from machine division incompatible with modern Python based tools
- Inclusion of meta data

- Some technical highlights:
  - Working e.g. on automated CI/CD pipelines with CVMFS container deployment on NAF
  - Data conversion at scale on NAF
  - ToDo: E.g. automated conversion of DOOCS files during DAQ



Preparation for HL-LHC

- How to run analysis with O(x5) more data?
- Which tools? How will Analysis Facilities look like in the future?
- → Lots of discussions going on in HSF, WLCG, ERUM, …

Research aspects of these within the NAF.

- repeat OSG Analysis Grand Challenge (“200 Gbit/s Analysis”) with the NAF
- pioneering new tools, e.g. Coffea Casa

Recent portal developments

Providing applications for Maxwell users

General purpose

Julia apps

Generic utilities

NoVNC EL9

Virtual Desktops

Debian + CrystFEL

mlExchange

Scientific Applications

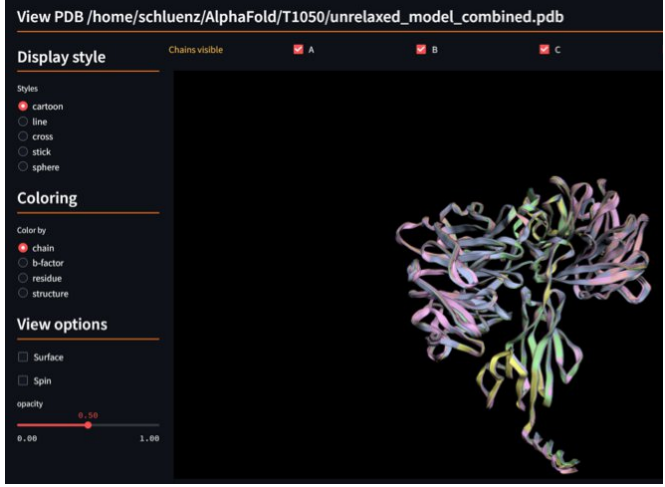
HDF5 Viewer

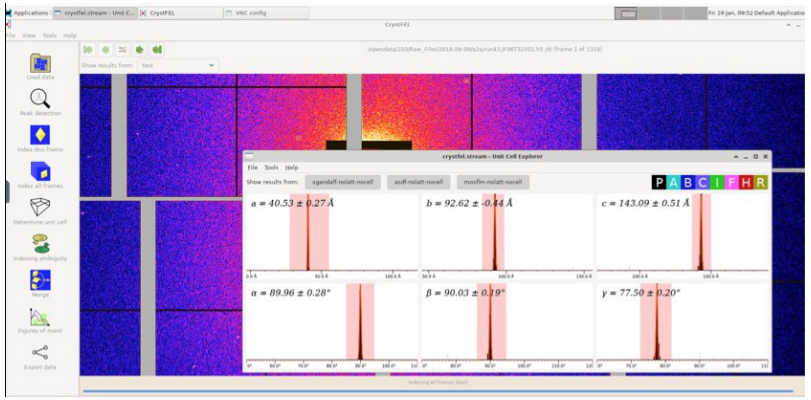
SciCat Frontend

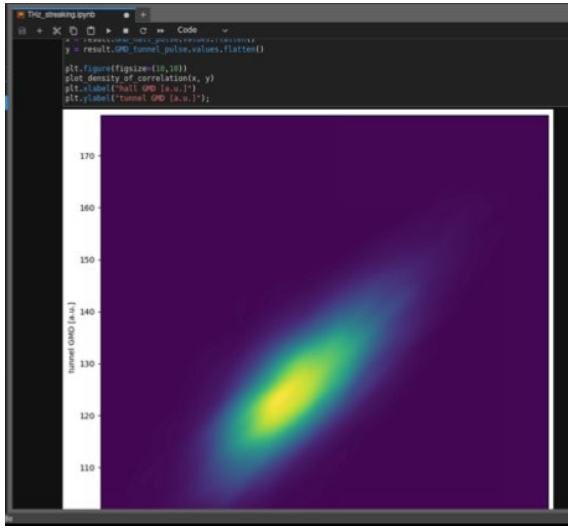
Named Servers

Note: this is a test setup. It won't work most of the time! Start a named server, or use one of the pre-configured applications.

Server name	URL	Last activity	Actions
Name your server Add New Server			
hdf5		a month ago	start delete
scicat		a month ago	start delete







Sustainability: Green-IT

How to Make the Infrastructure more Sustainable

Constant improvement on PUE in DESY CC and infrastructure on DESY Campus … ongoing since years

- Hardware life cycle under close watch

Compute: Adapt hardware availability to power availability and/or user needs

Storage: Unused data on tape → Tape?

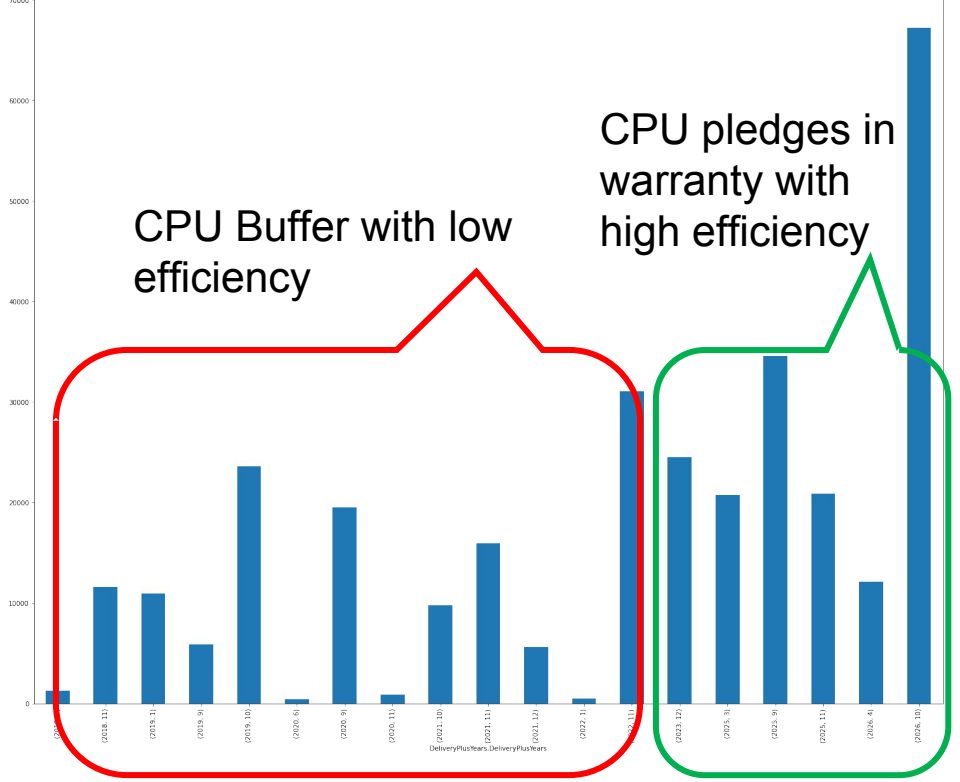
Raising **awareness** of users

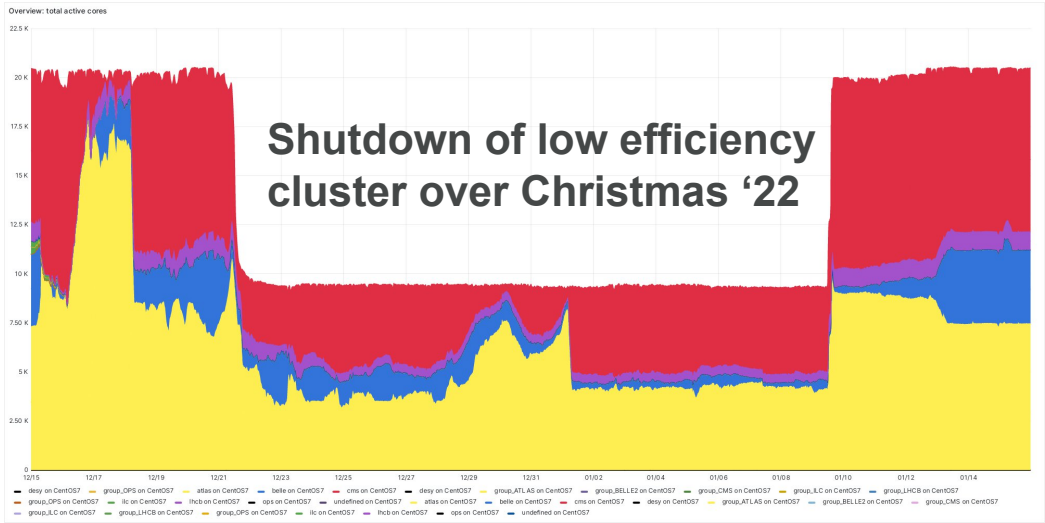
Train users on most efficient use of IDAF

Train users on tooling and optimal algorithms

Interactivity and fast reaction come with inefficiencies:

- Re-evalute how much fast response is needed
- Eventually tax users
- Work on scheduling and availability





Provided by T. Hartmann

Project facts & figures

- research facility 2.0 – short RF2.0
- EU project – HORIZON-INFRA-2023-TECH-01-01
- from 01/01/2024 until 31/12/2026
- could employ 2 scientific FTE – hiring shown being ‘not easy’

For further information and to follow our project progress visit **www.rf20.eu**

and our Social Media accounts: RF2.0 Project @rf20_project

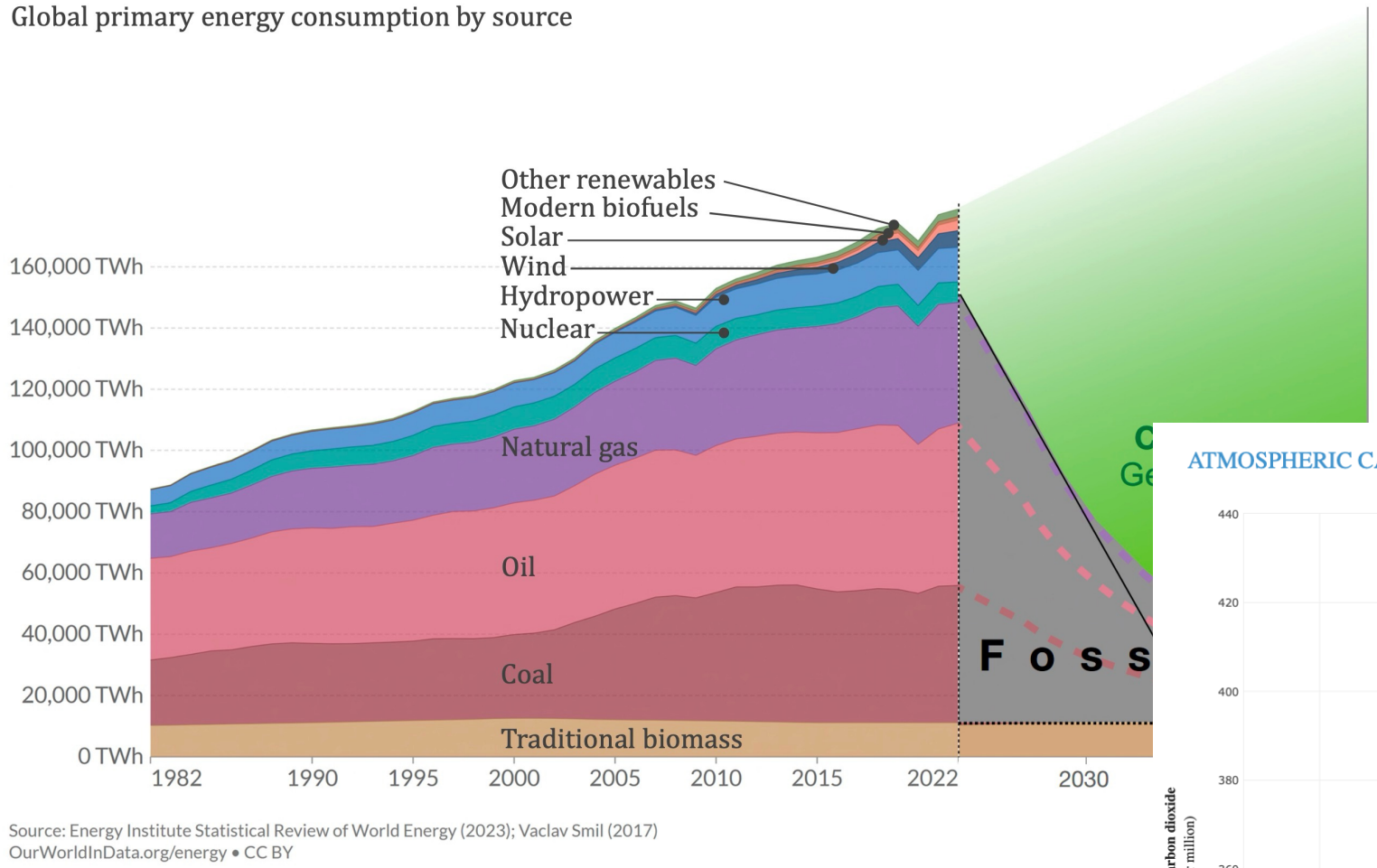
Funded by the European Union

This project has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No. 101131850 and from the Swiss State Secretariat for Education, Research and Innovation (SERI).

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

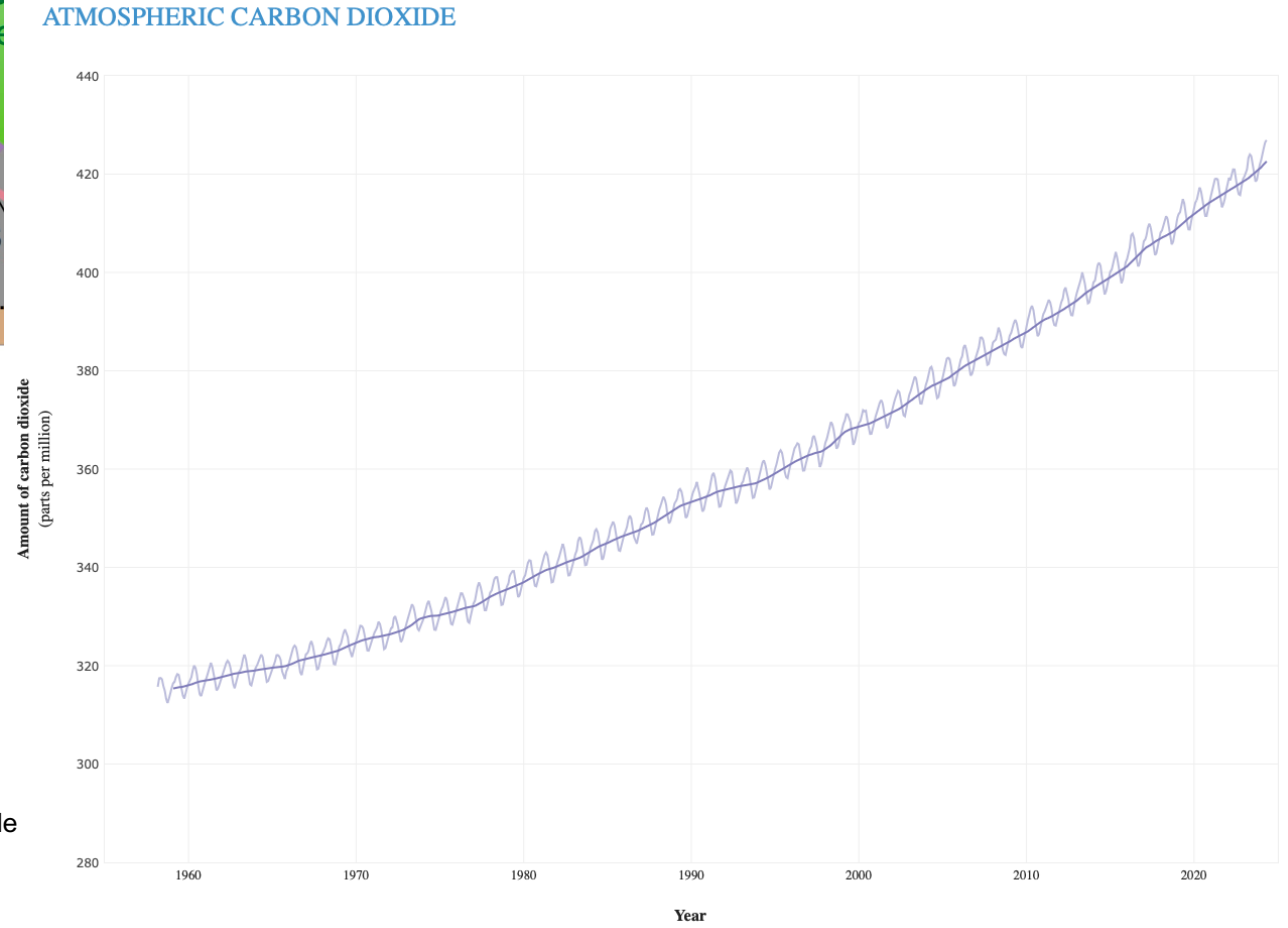
Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI



Target is pretty simple to locate ;-)

the current real effect too



https://www.climate.gov/news-features/understanding-climate/climate-change-atmospheric-carbon-dioxide

topics to look at – real GHG reduction is primary goal

the infrastructure to be looked at - *IDAF*

- **software efficiency** – has been and will be a contigous topic – responsibility shared between IT and *YOU*
  - has some ‘dilemma’ inbuilt – high efficiency -> high vendor/technology lock-in – technology will change
    - so do programming language selection (see “No computing on a dead planet”, Chris Edsall, Sept. 2021)
- **hardware efficiency/new architectures** – similar responsibily distribution – strong links to previous topic
- more core infrastructure related
  - **longer lifetime** (RE usage mandantory, puts emmbedded carbon into the game)
    - with changed purchasing – reduced direct service costs, adapted contracts, skill change for operating crew
  - **waste heat usage** – sounds easy – its not.
- the above topics are expected to bring 10-20 % gains in the long run and on average – won some battle but not the war

- higher potential – with *additives* from above
  - **active control of compute load to ‘only consume’ renewable**
- 25/75 rule – IDAF equippment has ~25% embedded vs. ~75% usage CO_2eq share – personal stuff just reversed

- elephant in the room - **Rebound Effect** – simple to understand, complicated to overcome

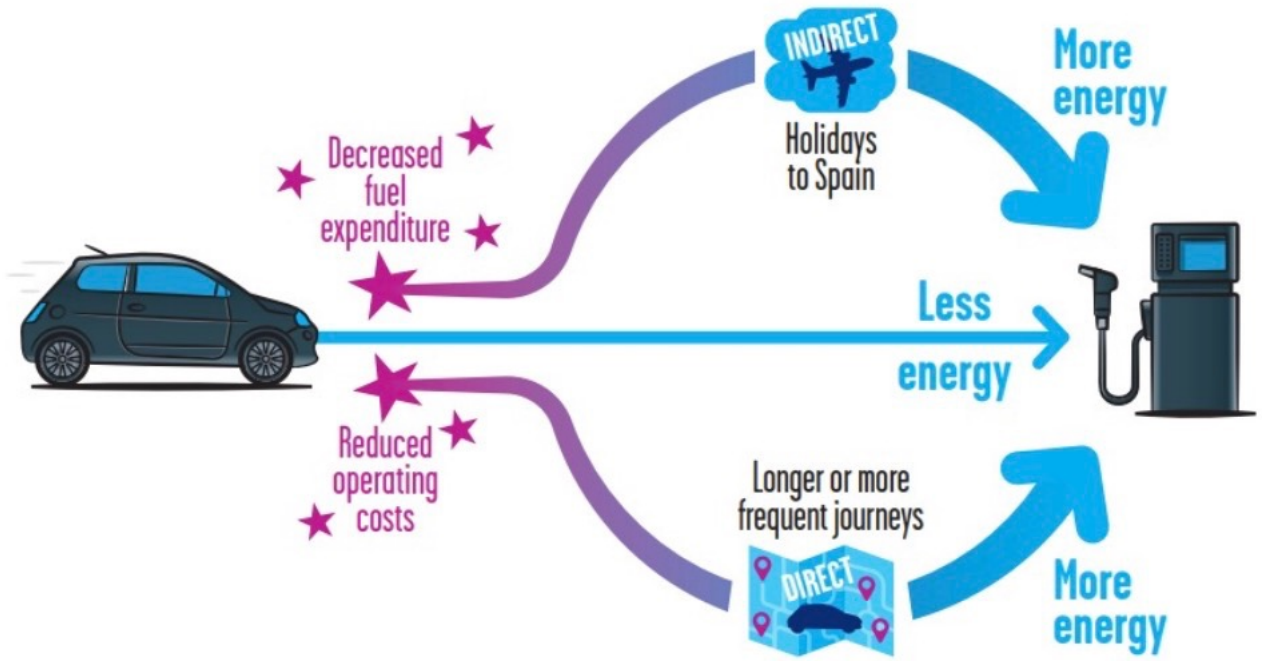
“Energy efficiency across programming languages:
how do energy, time, and memory relate?”

DOI:10.1145/3136014.3136031

abstraction and ease of use
has its costs !

Table 4. Normalized global results for Energy, Time, and Memory

Total					
	Energy		Time		Mb
(c) C	1.00	(c) C	1.00	(c) Pascal	1.00
(c) Rust	1.03	(c) Rust	1.04	(c) Go	1.05
(c) C++	1.34	(c) C++	1.56	(c) C	1.17
(c) Ada	1.70	(c) Ada	1.85	(c) Fortran	1.24
(v) Java	1.98	(v) Java	1.89	(c) C++	1.34
(c) Pascal	2.14	(c) Chapel	2.14	(c) Ada	1.47
(c) Chapel	2.18	(c) Go	2.83	(c) Rust	1.54
(v) Lisp	2.27	(c) Pascal	3.02	(v) Lisp	1.92
(c) Ocaml	2.40	(c) Ocaml	3.09	(c) Haskell	2.45
(c) Fortran	2.52	(v) C#	3.14	(i) PHP	2.57
(c) Swift	2.79	(v) Lisp	3.40	(c) Swift	2.71
(c) Haskell	3.10	(c) Haskell	3.55	(i) Python	2.80
(v) C#	3.14	(c) Swift	4.20	(c) Ocaml	2.82
(c) Go	3.23	(c) Fortran	4.20	(v) C#	2.85
(i) Dart	3.83	(v) F#	6.30	(i) Hack	3.34
(v) F#	4.13	(i) JavaScript	6.52	(v) Racket	3.52
(i) JavaScript	4.45	(i) Dart	6.67	(i) Ruby	3.97
(v) Racket	7.91	(v) Racket	11.27	(c) Chapel	4.00
(i) TypeScript	21.50	(i) Hack	26.99	(v) F#	4.25
(i) Hack	24.02	(i) PHP	27.64	(i) JavaScript	4.59
(i) PHP	29.30	(v) Erlang	36.71	(i) TypeScript	4.69
(v) Erlang	42.23	(i) Jruby	43.44	(v) Java	6.01
(i) Lua	45.98	(i) TypeScript	46.20	(i) Perl	6.62
(i) Jruby	46.54	(i) Ruby	59.34	(i) Lua	6.72
(i) Ruby	69.91	(i) Perl	65.79	(v) Erlang	7.20
(i) Python	75.88	(i) Python	71.90	(i) Dart	8.64
(i) Perl	79.58	(i) Lua	82.91	(i) Jruby	19.84



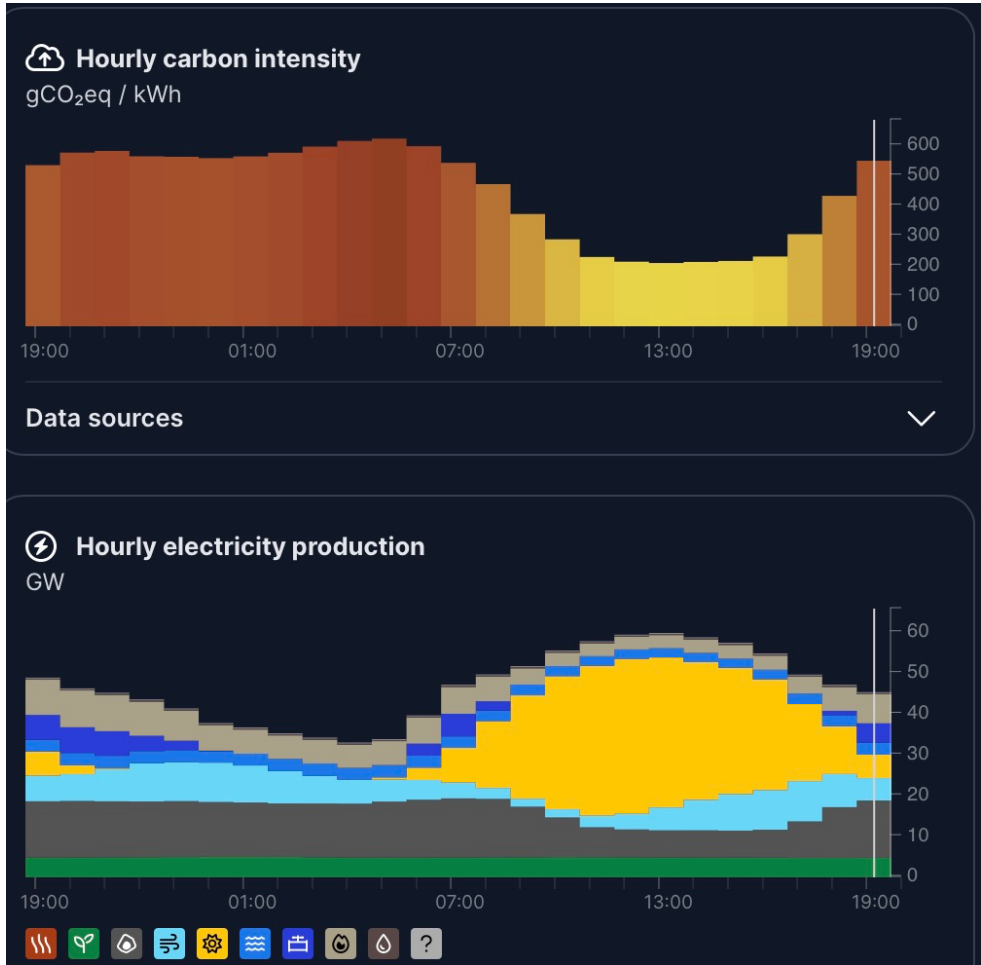
Source: Steven Sorrell: The Rebound Effect, UK Energy Research Centre, 2007

known facts and valid expectations

for us (DESY) at Hamburg within Germany

- renewable power has 13-35 gCO_2eq / kWh (called: carbon intensity) – not ZERO and not UNLIMITED
  - production costs around 6ct/kWh or lower
- other ‘low carbon’ – nuclear and/or fusion
  - >=16ct/kWh and growing fast – even reasonable fusion planning
- the ‘cable’ in between costs ~5ct/kWh
- adding ‘smart buffers’ will add some ‘smart’ 5-10ct/kWh’ (and more) on top

https://app.electricitymaps.com/zone/DE?wind=false&solar=false&page=highscore&countryCode=CA-PE



- the German e-power market is a mess and not in a shape to start the ‘change’ – ‘Strom-Mafia’ – exempted from physics ;-)
  - DESY grid example

- the ‘where’ and ‘when’ to draw power are the important factors
  - the ‘where’ assumed to be fixed ;-) - DC close to RE generation (m not km)
- using ‘direct RE power’ with just a ‘cable’ to DESY
- dealing with the ‘when’ – riding the ‘carbon intensity wave’

today

~2000 servers
~200 PB of storage
~few hundred switched

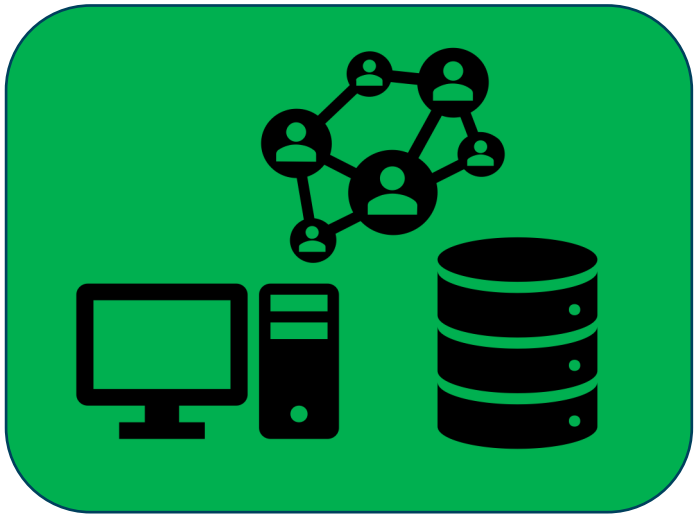
all in ‘non-simple’ environment
cooling, power, access, etc.

- steady power demand (1.4MW)
- no active power control

ASAP – As Amall As Possible
same mode as previous option:

- run new HW (~2 years)
- all (nearly) network
- all storage

final / some day



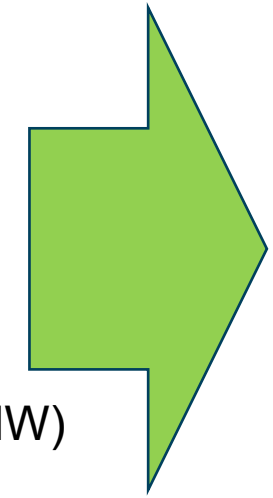
- majority of compute systems
- required network stuff

- lifetime >= 8 years
  - in total, more boxes !
- runs only on ***direct renewable power***
- use old Exp. Hall ?
- E-Storage – UPS extension
  - district integration

today

~2700 servers
~180 PB of storage
~few hundred switched

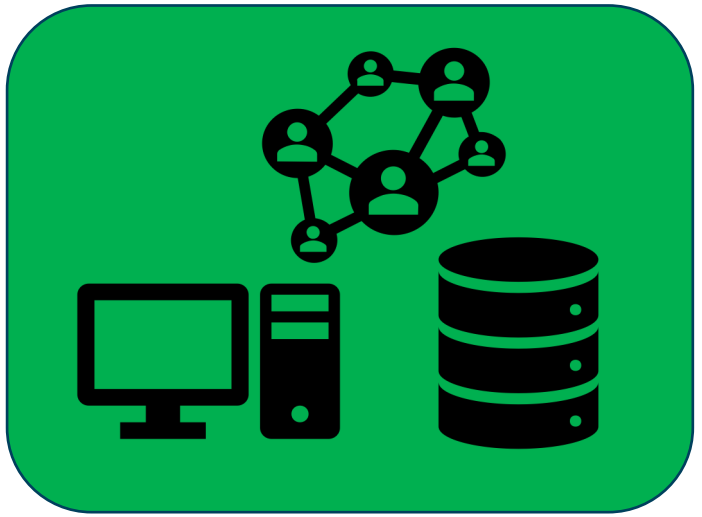
- constant power demand (1.4MW)
- no active power control



ASAP – As Amall As Possible
same mode as previous option:

- run new HW (~2 years)
- all (nearly) network
- all storage

some day



- majority of compute systems
- required network stuff
- runs only on ***direct renewable power***
- optionally
  - lifetime >= 8 years
    - in total, more boxes !
  - bigger chunks of the same
  - use old Exp. Hall ?
  - E-Storage – UPS extension
    - district integration

- **motivation (individually)**
- **expectations – your perspective**
- **require**
  - **quality metrics**
  - **WLCG adoption (pledges, job submission, ...)**
- **things already happened – training courses, XMAS switch off, ...**
- **other brilliant ideas ? its time to publish ;-)**

Visibility: Presentations, papers are possible!

Computing groups / workshops of the experiments

…