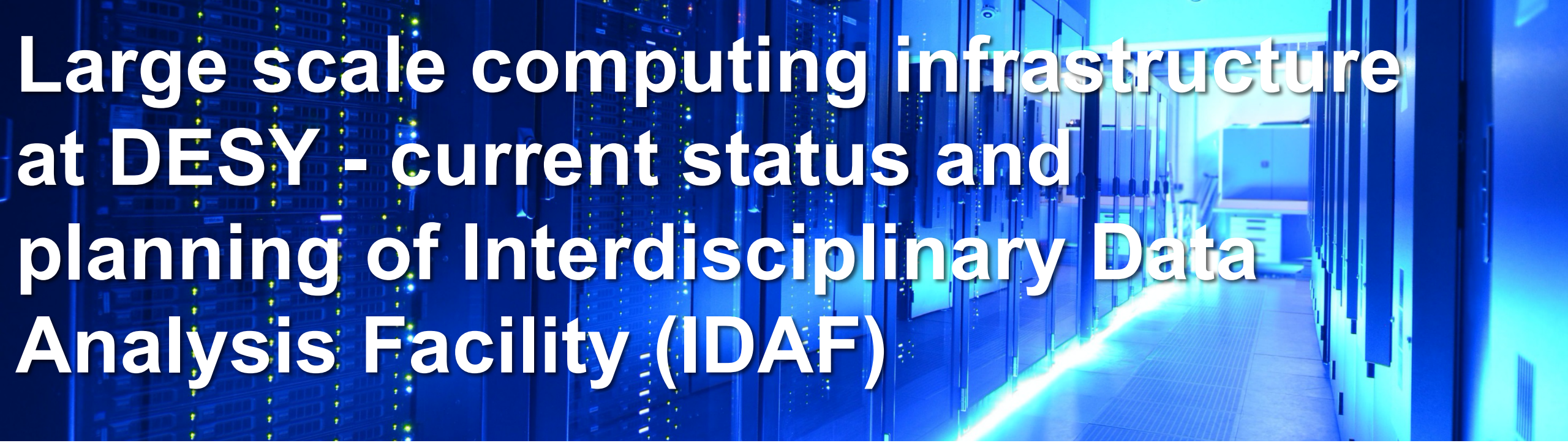


Large scale computing infrastructure at DESY - current status and planning of Interdisciplinary Data Analysis Facility (IDAF)

Christian Voss, Yves Kemp, for DESY IT
IPC seminar
03.11.2020 DESY

With slide contributions from DESY and XFEL people

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

DESY.

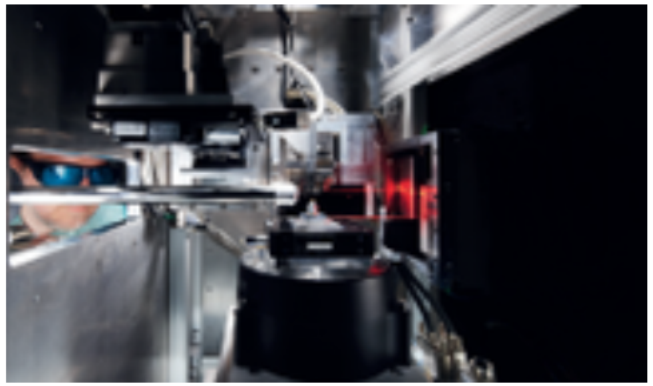
DESY research divisions ... In a nutshell

Accelerators »

Running / Operating:
- Petra III, FLASH, XFEL, ...

Planning:
- Petra IV

General Accelerator R&D



Photon science »

Petra III, FLASH, EXFEL, CFEL, CSSB, EMBL, HZG

Particle physics »

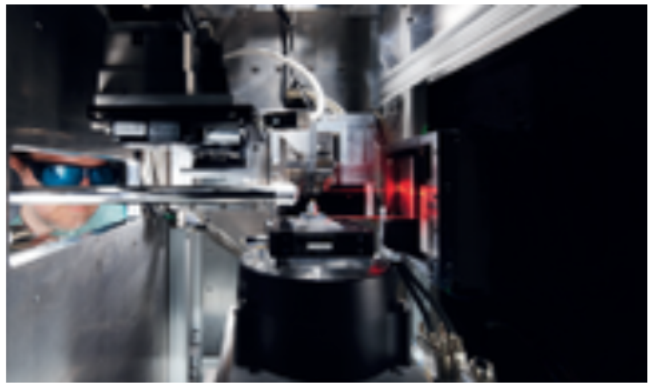
- LHC, HL-LHC
- Belle II
- ILC, ALPS, ...
- Theory division

DESY research divisions ... IT involvment in scientific computing

... An incomplete view

Accelerators »

- Storage operational data
- Simulation & computational infrastructure for R&D
- Support



Photon science »

- Online DAQ
- Offline storage & analysis infrastructure
- Simulation & computational infrastructure for R&D
- Support

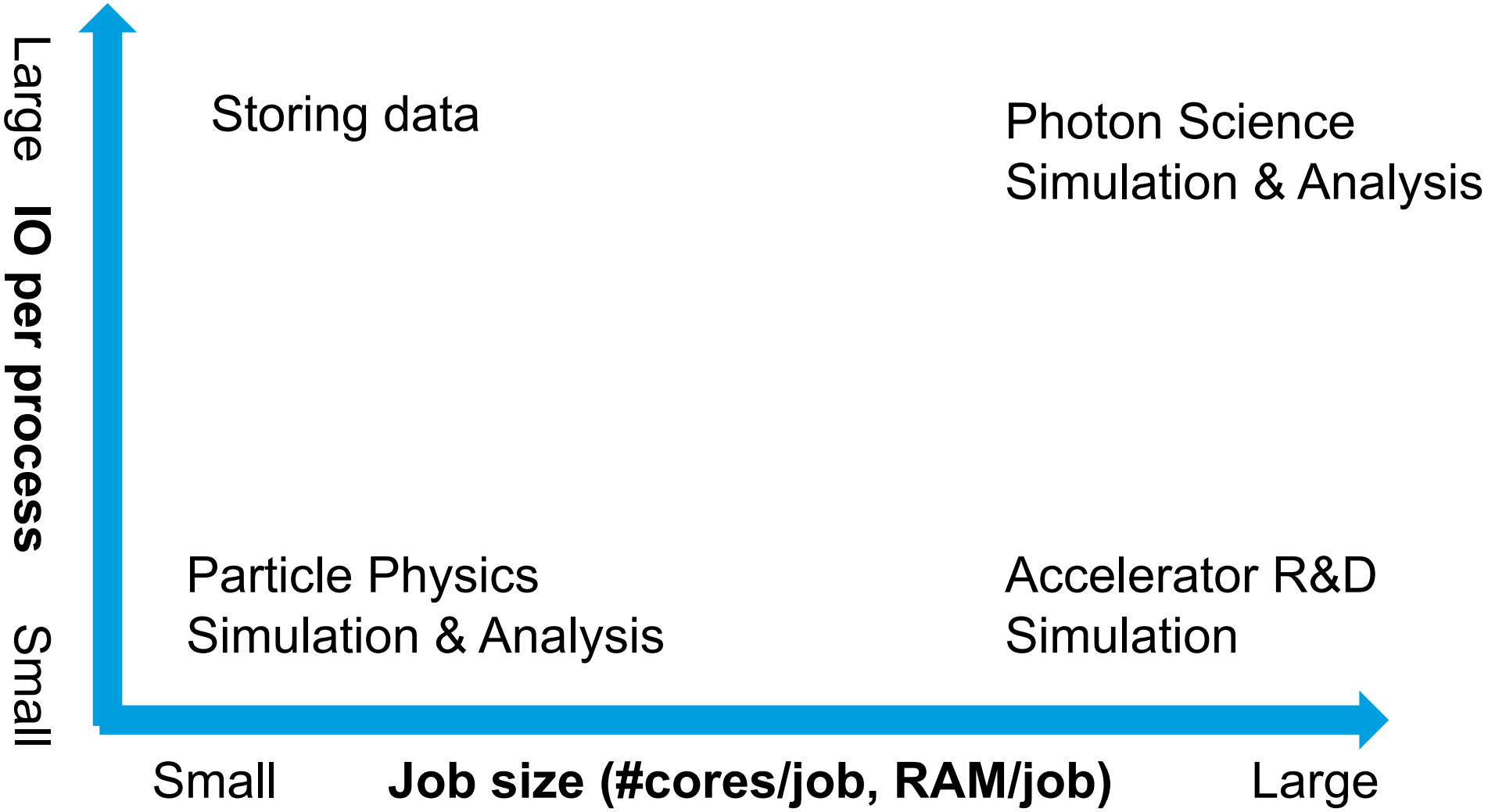
Particle physics »

- Global and national tasks within LHC and BELLE II
- Simulation & computiational Infrastructure for ILC and detector R&D
- Support

Now, designing the compute and storage infrastructures

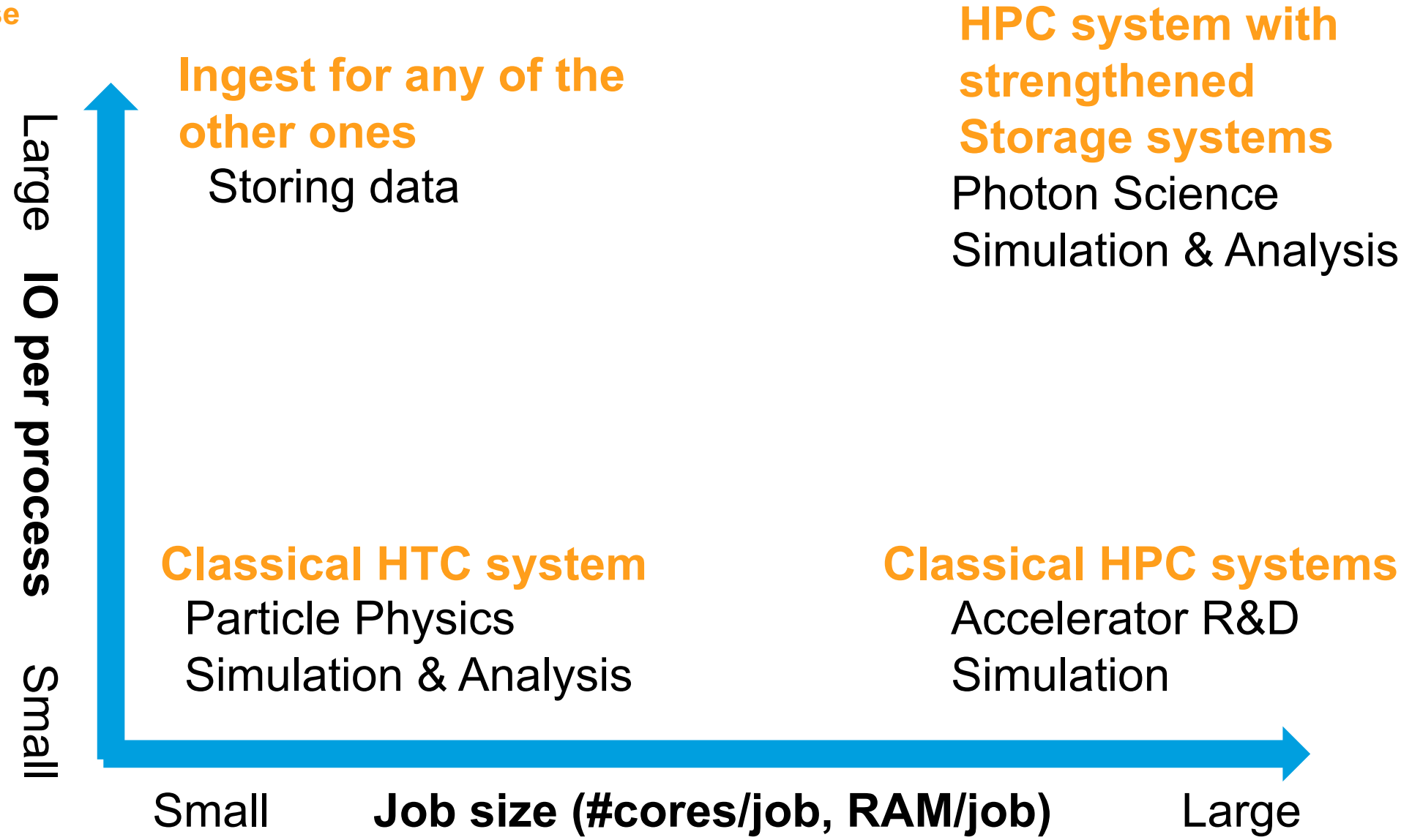
Computational requirements: Job size vs IO needs

Very very coarse



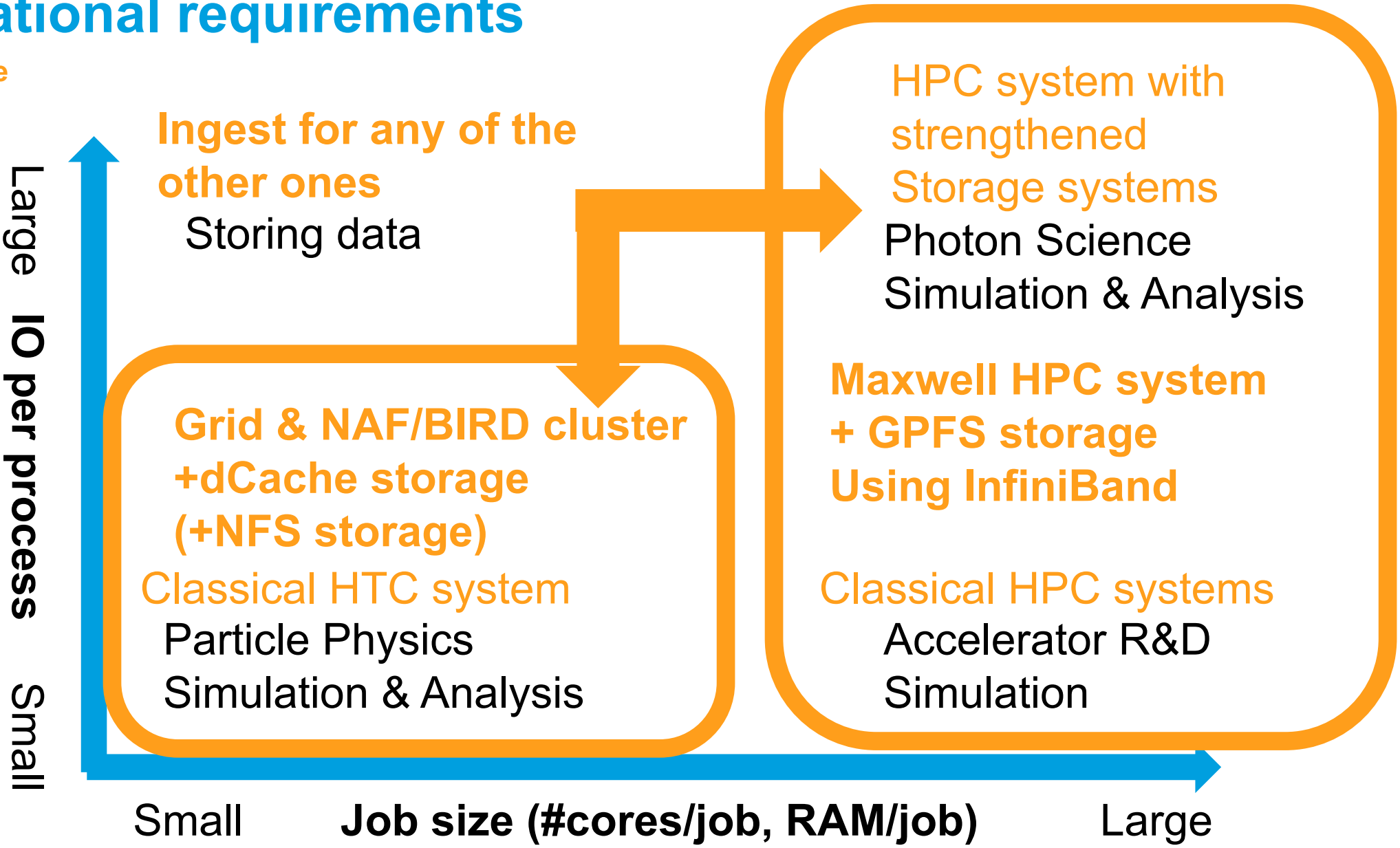
Computational requirements

Very very coarse



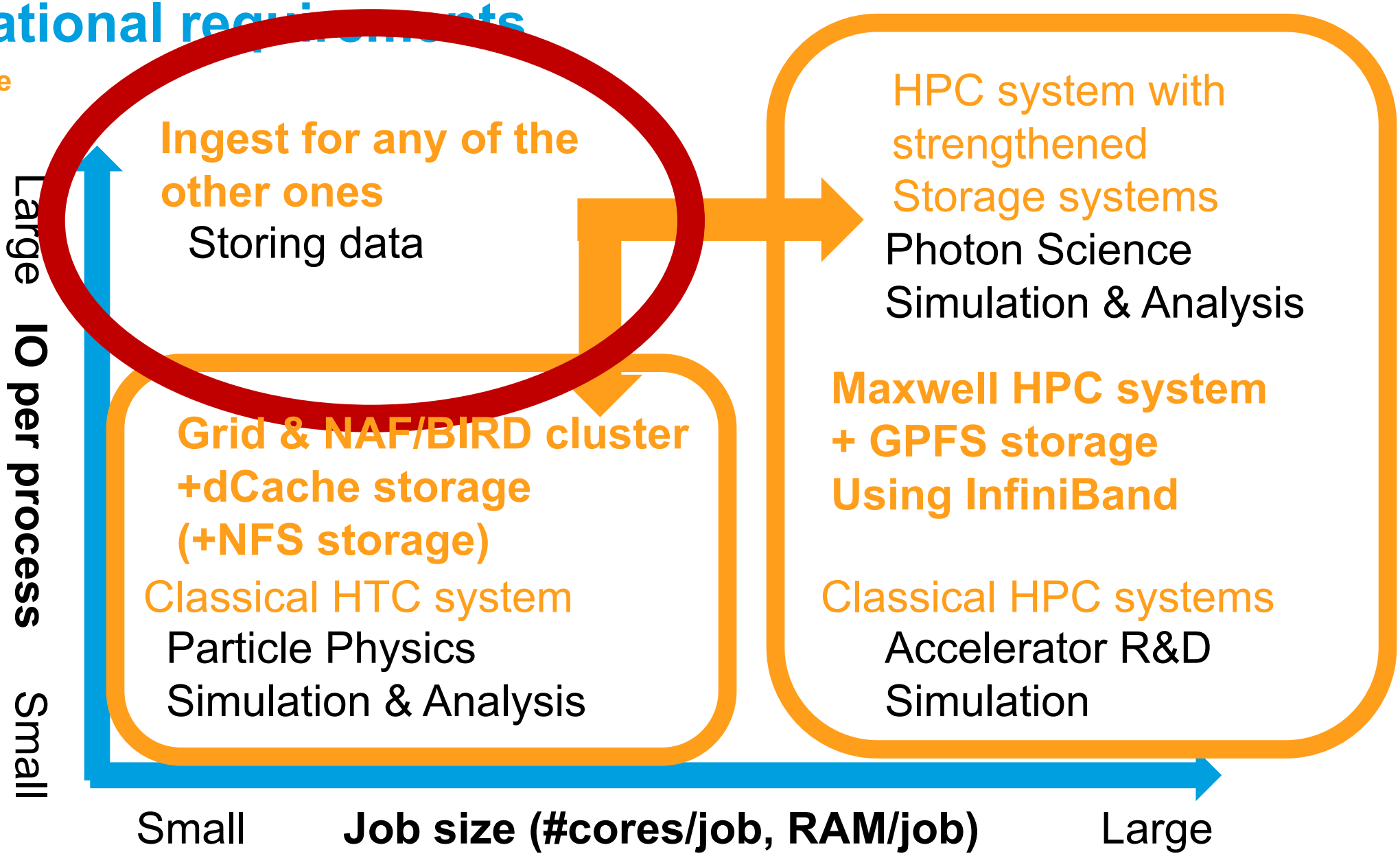
Computational requirements

Very very coarse



Computational requirements

Very very coarse



Data Storage: Essential for Science@DESY

Data Management

Today: Most Scientific Endeavours Produce Large Amounts of Data

- Computing@DESY: Storage of data for all departments and communities

(Astro-)Particle Physics

- Store and archive raw data
- Store and archive simulated data
- Store pre-processed data for
  - Experiment specific workflows
  - Dedicated user analyses

Accelerators and Detectors

- Store and archive simulated data
- Store and archive test-beam data
- Store and archive telemetry data

Photon Science

- Store simulated data
- Store and archive raw data
- Store pre-processed data for analyses

- Data as central element for most research
- Make data the central hub and trigger for scientific workflows

Example For Trigger on Data

AMPEL System Developed and Deployed in Zeuthen

- AMPEL: Real-Time astronomy data analysis framework to find transient objects
- Transient objects disappear to quickly for classic data analysis methods

Palomar Observatory

>80 Exposures/h

Message with data

Message-Broker

Collect Message

Publication Message

- Analyse message
- Reconstruct data
- Accept/discard Message

Collect Publication Message

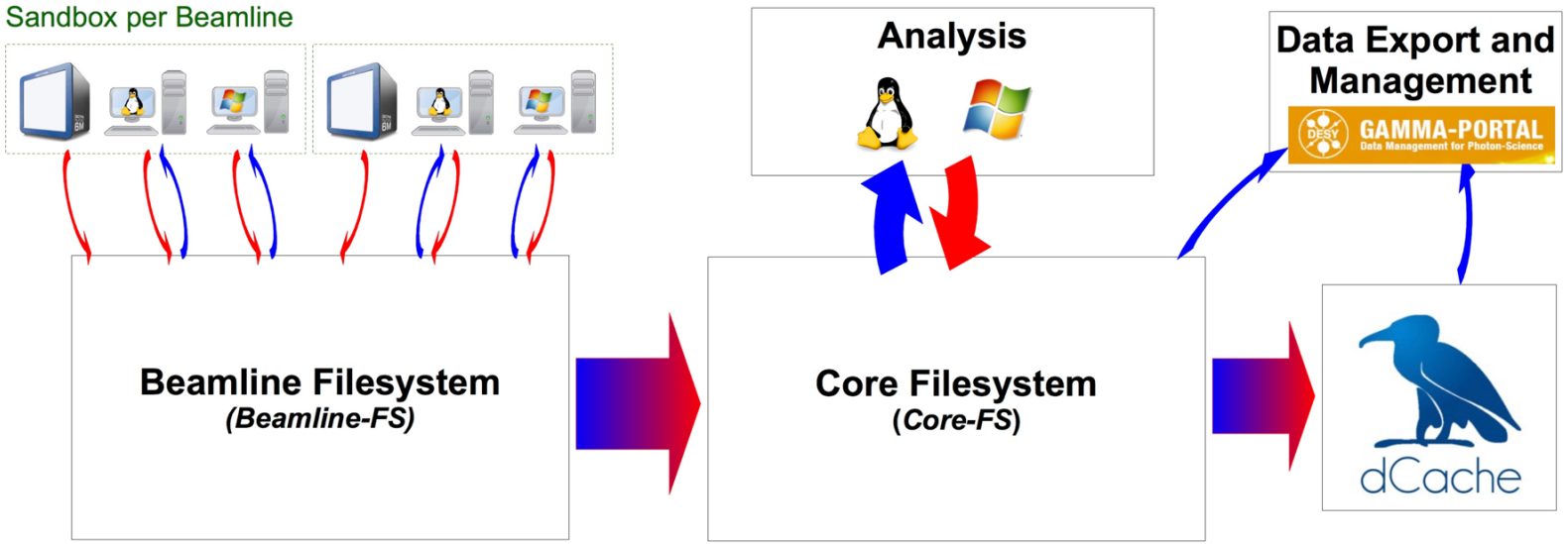
- **Message Producer – Broker – Consumer Model**
  - Any shot generates notification in message system, triggering a detailed analysis, triggering publication message
- Similar application: **Storage Events**
  - Incoming data stream or file triggers analysis → Jobs on Compute Cluster
    → Data flush to persistent storage
  - Data storage as Workflow engine

Trigger on Data: Photon Science Example

Start from the Existing Model

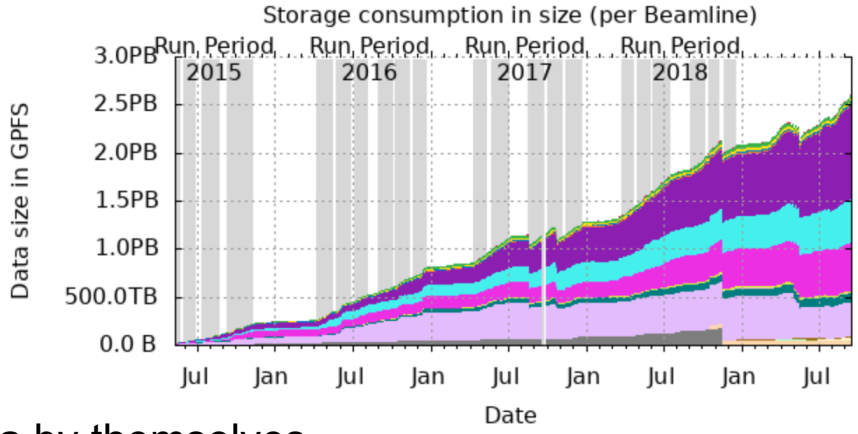
ASAP³

- Coverage of full data life cycle
- Multi-tier storage from:
  - Fast and small
  - Slow and large
- Time delay based synchronisation
- System in place at PETRA III/FLASH
  - High Data Rate Access – HiDRA



Limitations

- Real-time data analysis – data access in microseconds after generation
- Increased data rates and stored data for new detectors
- Revised system in place at European XFEL (e.g. larger focus on mass-storage similar to HEP) due to Proposals ≈1PB
- PETRA IV: 10^4 – 10^5 increased data rates



Paradigm Change for User Experiments

➔ Users do not have resources or might not be experienced enough to process data by themselves

➔ Data management becomes integral part of experiment

Trigger on Data: Photon Science Example

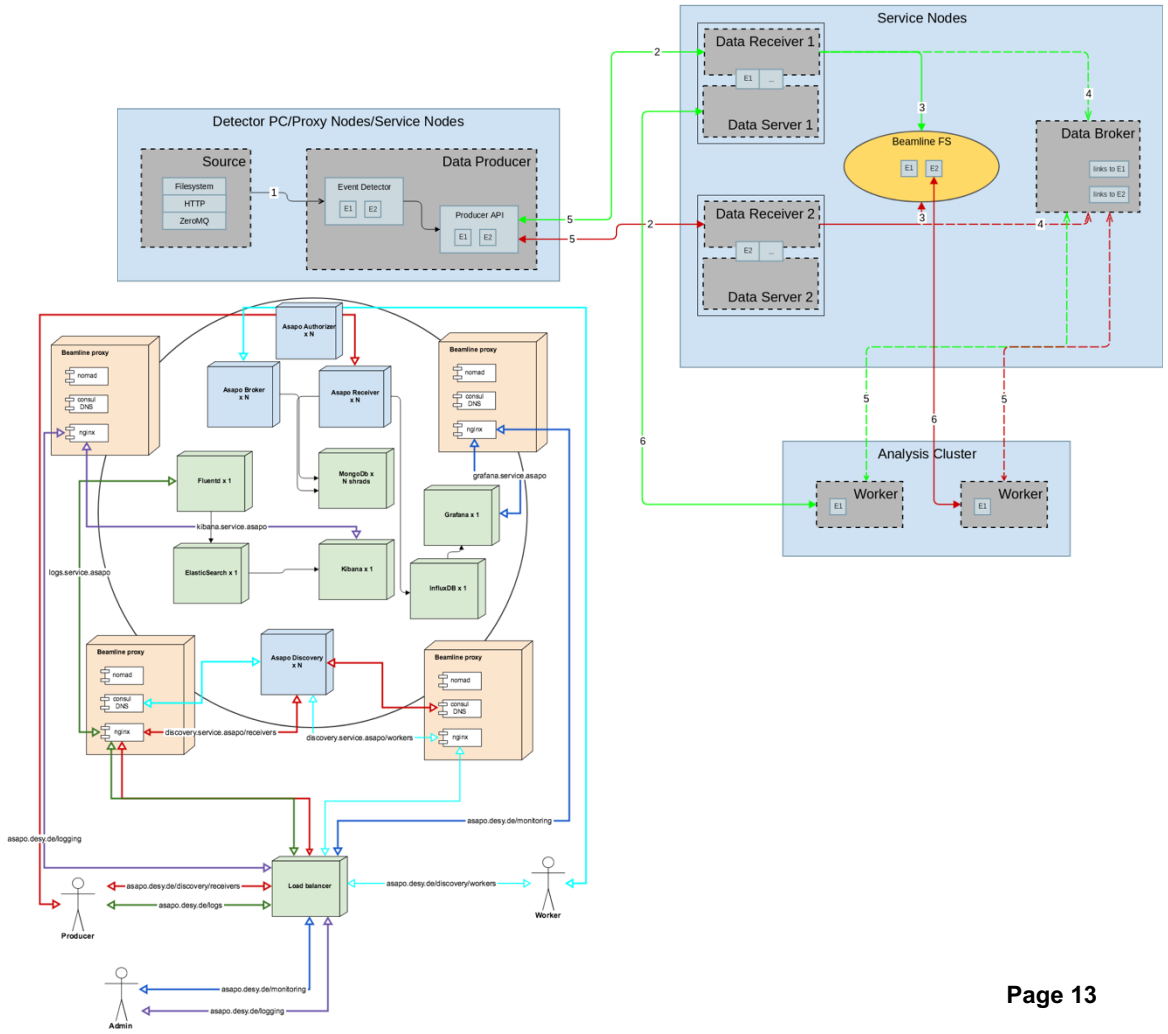
Adapt the Message Model to Photon Science

- **Message Producer – Broker – Consumer Model**

- A distributed streaming framework for high performance scientific data analysis
- File based service in place at PETRA III Beam Line `p02`

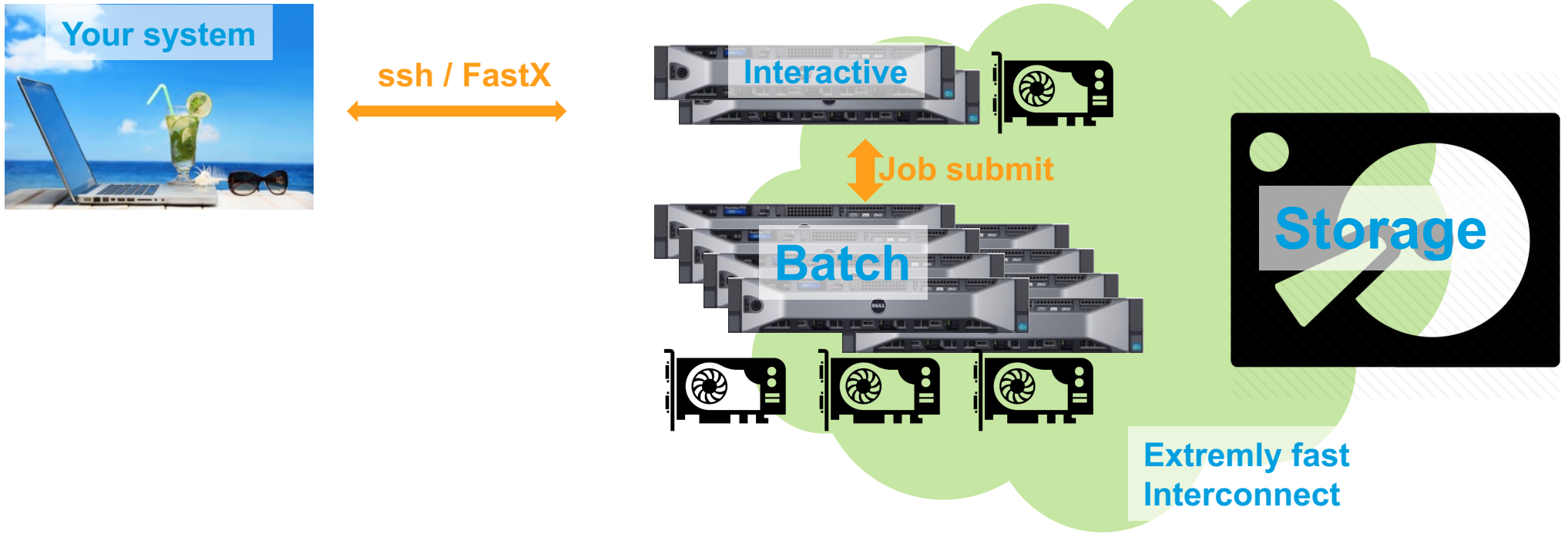
- Future: Keep data in transient memory
- Prototype code done – including metadata DB (mongo)
- Builds for Linux & Windows
  - Performance critical code in C++
  - Python API (HiDRA compatible)
  - Deployment on Kubernetes
- Store of data after online analysis

- Deployment on HPC Resources on Maxwell



Analysis & simulation infrastructure at DESY

Basic setup at DESY



The Setup for particle physics

ssh / FastX

NAF (+Grid)

Interactive

Job submit

Batch

Storage

~~Extremly~~ fast Interconnect

The Setup for photon science & accelerator R&D

ssh / FastX

Maxwell

Interactive

Job submit

Batch

Extremly fast Interconnect

Storage

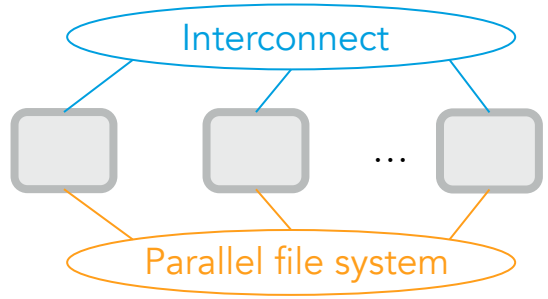
Beamline Filesystem
(Beamline-FS)

Sandbox per Beamline

Slide stolen from Maxence Thévenet

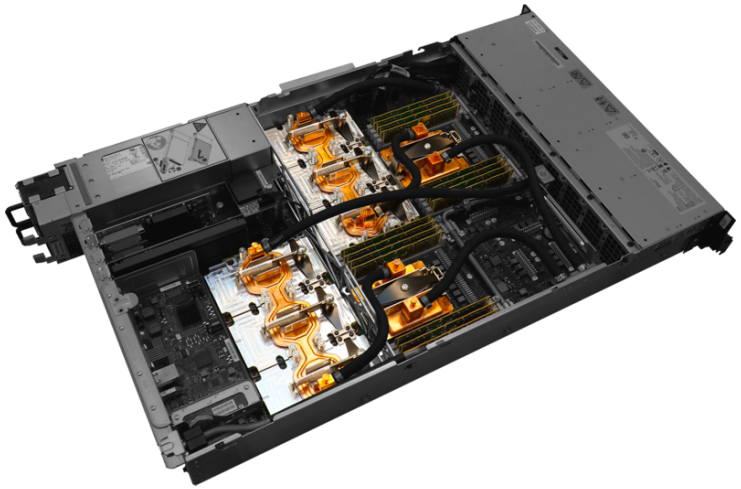
Architecture of a supercomputer

Number-crunching compute nodes + interconnect + file system



- **Compute node**
  - *Homogeneous within a partition of a supercomputer*
  - *Accelerated computing (Graphics Processing Unit)*
  - *More on that later* **Maxwell: - Homogeneous within partition: We try…**
    - GPUs: Yes!
- **Interconnect**
  - *Invisible to the user (send a message)*
  - *No all-to-all connections* **Maxwell: Using InfiniBand in a blocking fat tree topology**
  - *Multiple topologies (Fat Tree, Torus, Dragonfly)*
  - *InfiniBand is a widespread communication standard*
- **Parallel file system (I/O)** **Maxwell: InfiniBand based storage:**
  - *GPFS, Lustre*
    - **GPFS for $HOME , P-3 and XFEL**
    - **BeeGFS as ”project space”**
- **Software**
  - *Open-source, Linux-based* **Maxwell: Slurm**
  - *Job scheduler: Slurm, LSF* **Supporting MPI**
  - *Launcher (resource allocation, placement): mpirun, srun*

Many other applications available, incl. commercial ones



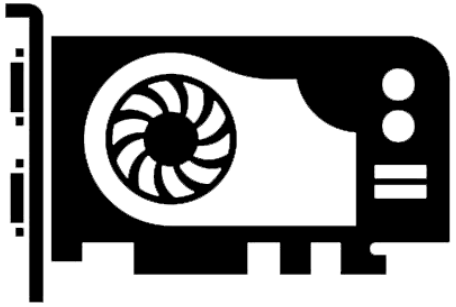
https://www.olcf.ornl.gov/summit/

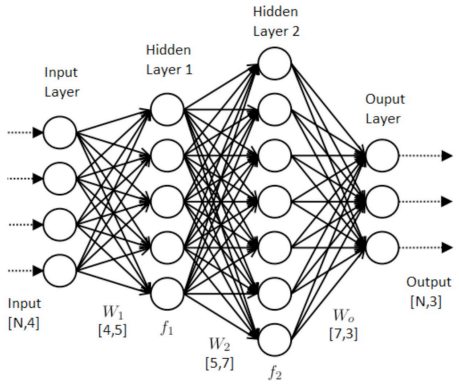
Comparing Maxwell HPC & GRID/NAF HTC systems

Feature	Maxwell	GRID/NAF
Size	~550 nodes / 35k cores / 280 TB RAM ~100 nodes with GPU	~ 900 nodes / ~30k cores / ~100 TB RAM
Network	InfiniBand for fast data & IPC, 10 GE Ethernet	1 GE - 10 GE Ethernet
Storage	Access to GPFS data (IB), dCache (NFS, Ethernet). BeeGFS for projects (IB)	dCache (NFS, Ethernet), GPFS (NFS, Ethernet)
Batch strategy	Whole/Multi-node-scheduling. Integration of private resources possible, with prioritized access.	Per-core-scheduling, no multi-node. Centrally procured resources. Fairshare on group basis.
Product	SLURM	HTCondor

GPU computing & Machine Learning

- **General GPU computing** established in HPC systems
  - … so in Maxwell: ~100 nodes equipped with GPUs
  - Different generations, different setups: From one GPU/Server to four GPU/Server
- Maxwell HPC cluster natural candidate for hosting GPU computing
  - Users have applications profiting from GPUs
  - GPUs benefit from “HPC-like” environment





- **Machine Learning**
  - Boosted by the usage of GPUs for training (and inference)
  - Benefits heavily from fast access to (large amounts of) data, and high-RAM machines
  - Maxwell is natural environment
- **Future** of GPU computing & Machine Learning
  - We see an increase in demand for “multi-GPU nodes” (~4 GPU/node)
    - Expensive, few nodes, challenging from scheduling point of view
  - Look for alternatives to NVIDIA. Have some examples in the lab. Dependency from CUDA challenging

Maxwell: plans for the furutre

- Strengthen Maxwell role as central DESY simulation and analysis hub
- Use the opportunity with the IDAF to optimize resource allocation, based on concrete needs of the job
  - Long-term goal: Transparent Cross-Cluster usage
- Keep Maxwell up-to-date:
  - In terms of hardware, quantity and quality
  - In terms of services offered
- Maxwell is a HPC cluster … and more:
  - Uses HPC technology and concepts (hardware, software, scheduling, …)
  - Platform for online Petra-III & XFEL datataking and analysis
  - Flexible access to resources. Do things one cannot do at large HPC facilities

Making batch more user-friendly – and maybe overcome it?

- Select a good scheduler ... With active developer
- Containers healing the OS & software incompatibilities
  - Started on Maxwell in 2016, using Docker technology
- Interactivity & access: Jupyter
  - Integrate interactivity into batch
  - „Tragedy of the commons“
- git based workflows & CI/CD
  - gitlab pilot phase, launch probably early 2021
- And ever and ever again, do training, taking by the hand, ...

slurm workload manager

+$CONTAINER

docker

jupyterhub

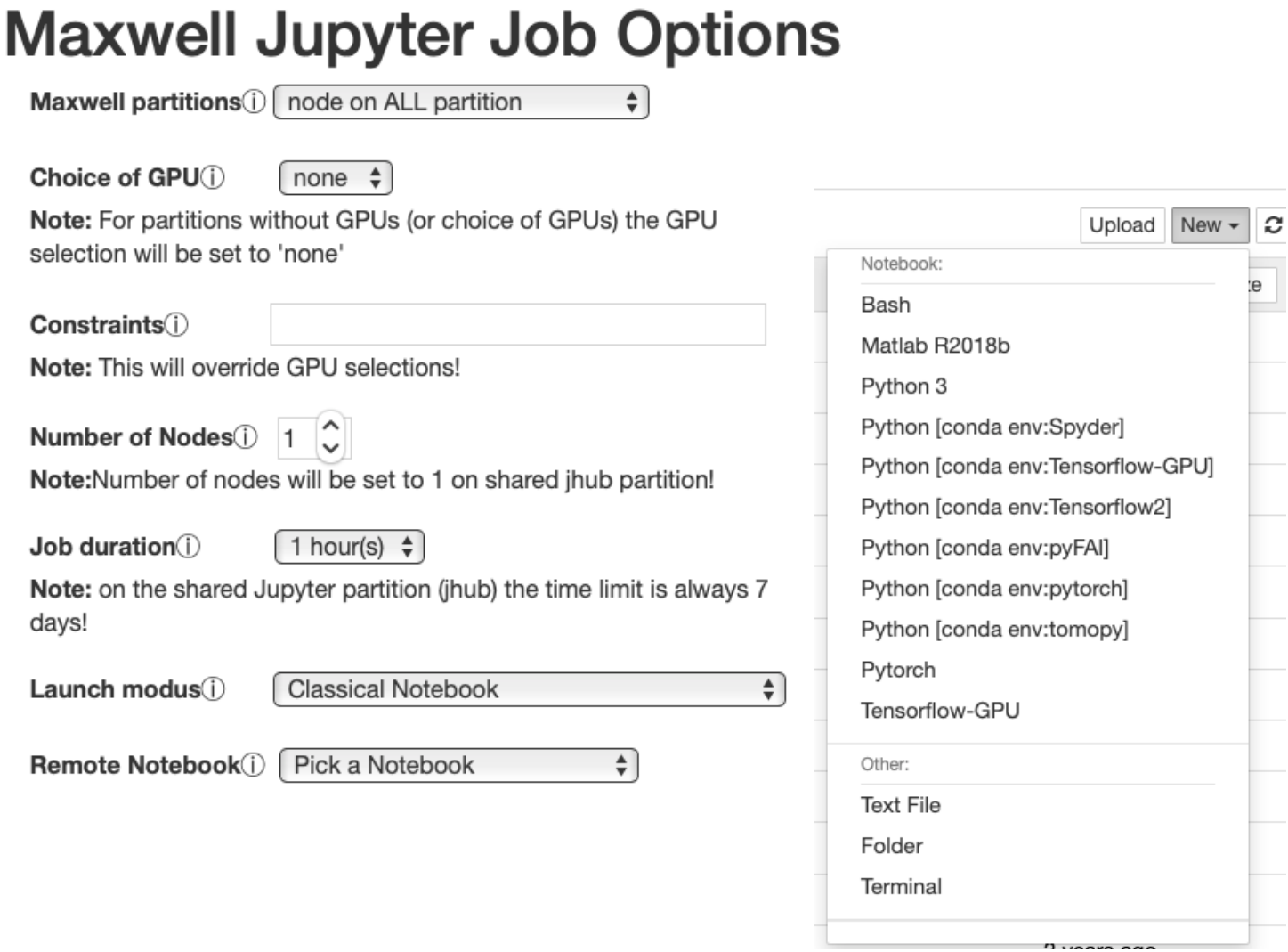
GitLab

Jupyter: Interactive & easy remote access

Jupyter notebooks and Maxwell

What are Jupyter Notebooks? Data analysis and simulation in your browser

- Python based interpreter for Python, Matlab, …
- Access via web-browser through portal
- Computation itself happens on Maxwell: Integration with SLURM scheduler
- https://max-jhub.desy.de/



dCache Storage System
- Mass storage for HEP, photon science & accelerator

The dCache Storage System

Distributed Scalable Mass Storage System

- Central element in overall storage strategy
- Collaborative development under open source licence by
  - DESY (leading laboratory)
  - Fermilab
  - Nordic E-Infrastructure Collaboration (ex. NDGF)

Particle Physics

- In production at 9 of 13 WLCG Tier-1 centres
- In use at over 60 Tier-2 sites world wide
- 75% of all remote LHC data stored on dCache
- In addition: Tevatron and HERA data

Astronomy & Radio-Astronomy

- LOFAR Long Term Archive (~40 PB) & CTA
- SkySurvey

Photon Science

- European XFEL, CFEL and others for archival

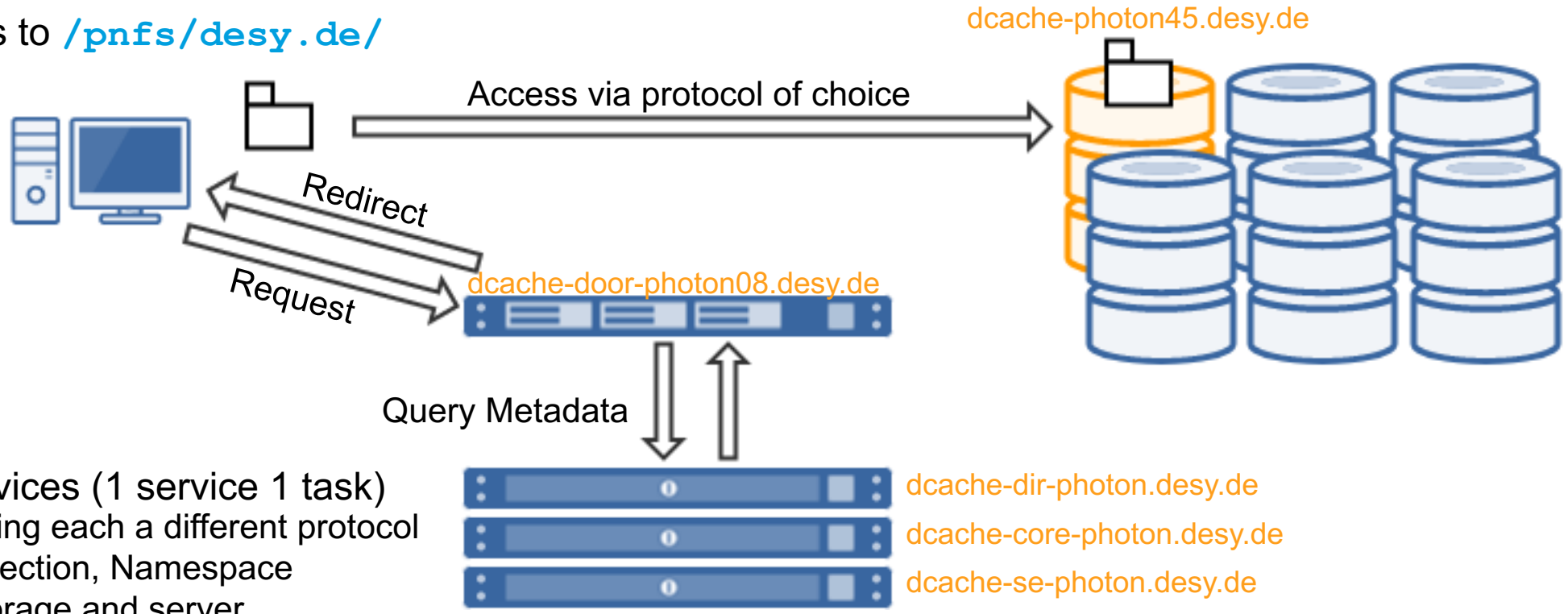
Accelerator and Detectors

- FLASH, LINAC telemetry

dCache: Architecture

User Access to dCache Responsible to Store Machine Data

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



- Based on Micro-Services (1 service 1 task)
  - `Doors` – supporting each a different protocol
  - `Heads` – pool selection, Namespace
  - `Pools` – data storage and server

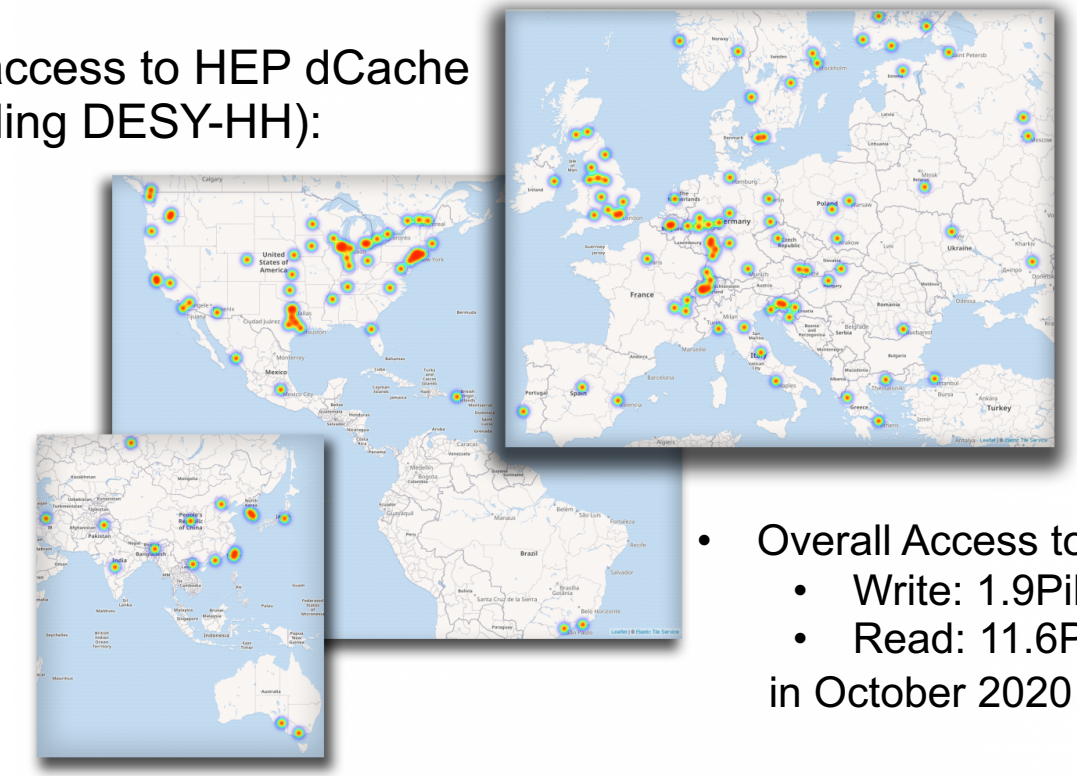
- dCache instances for Photon Science/Machine, European XFEL, ATLAS, CMS, Belle/ILC/DPHEP, Sync&Share
- Similar layout: three head-nodes, doors for requested protocols and pools nodes
- Scale-out horizontally: 10 pool nodes for Sync&share and 200 for European XFEL with 100 more ordered
- Scale-out horizontally: client always to connect to pools for transfer, no data access through doors

dCache: Capacity of Local DESY Instances

Available and Used dCache Storage

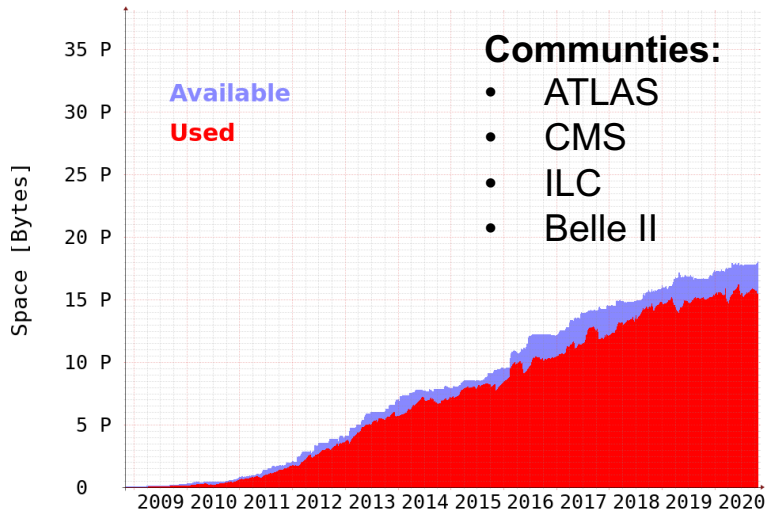
- Steady increase for HEP since inception of dCache
- Exponential increase for Photon science since start of European XFEL
- HEP dCache is connected to the WLCG
- Transfers all over the world

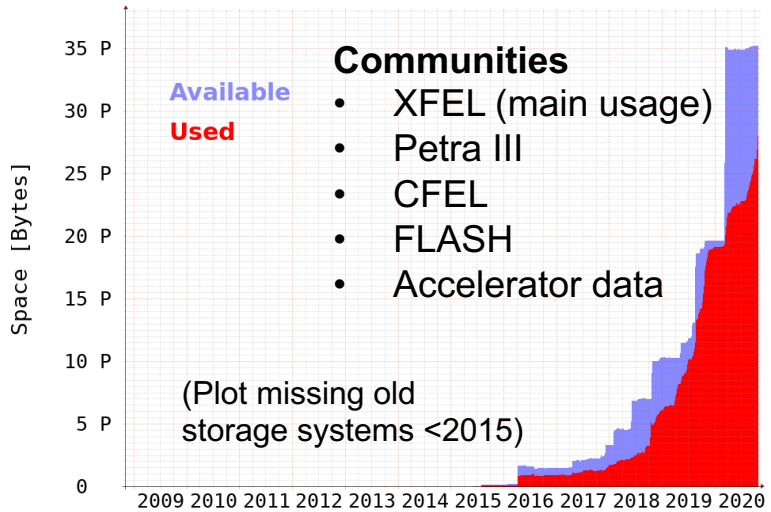
- Read access to HEP dCache (excluding DESY-HH):



- Overall Access to HEP dCache
  - Write: 1.9PiB
  - Read: 11.6PiB

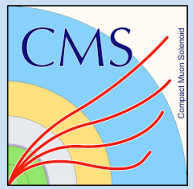
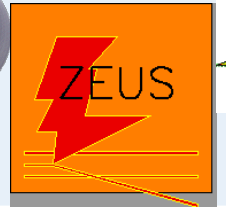
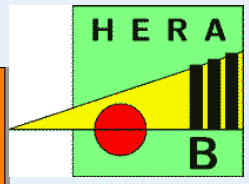
  in October 2020





A view to particle physics analysis

HEP communitites at DESY

Community / Experiments	Compute activities
  	Grid Tier-2, German NAF users
	Compute & Storage, Management services, Collaborative tools, ...
 	Compute & Storage, Management services
   	Compute & Storage, Management services

The Setup for particle physics

ssh / FastX

NAF (+Grid)

Interactive

Job submit

Batch

Storage

Extremly fast
Interconnect

CERN

WLCG
Worldwide LHC Computing Grid

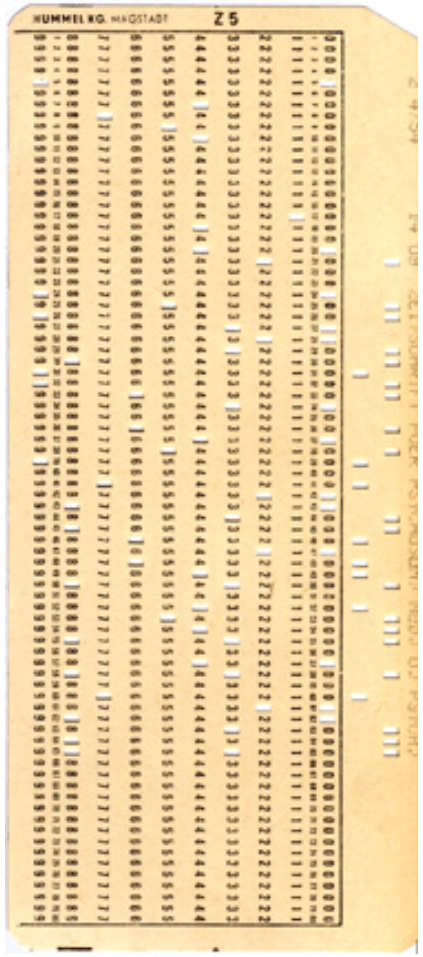
Belle II

dCache.org

HEP and Batch?

- Batch based computing Ansatz long established in HEP
- Nevertheless: Alternatives are being investigated:

- NAF: Augment Grid with interactive resources
- Jupyter as new access method is being rolled out successively
  - Investigation on Jupyter resource scaling
  - „Tragedy of the commons“
- Investigating DASK & Spark as non-batch compute organization
- gitlab / CI/CD workflows … connection to batch?
- Batch ↔ Cloud integration



(Compute) Cloud at DESY

(Compute) Cloud at DESY

- Compute Cloud Infrastructure @DESY:
  - Running OpenStack, with CEPH backend
  - ~1000 CPU cores
- Objectives:
  - Boring: Bring classical server consolidation workflows to cloud VMs
  - Interesting: Adding flexibility to current batch clusters (and IT infrastructures)
  - Thrilling: Compute cloud as enabler of novel Scientific Workflows (and general IT)
- Status:
  - Partially in production for D/EU projects
  - DESY pilot phase planned for 2021
  - Constant development and evolution ahead!

Unified Compute Infrastructure a.k.a. InterDisciplinary Analysis Facility IDAF

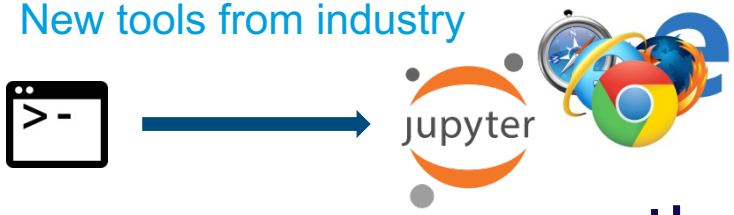
One vision for the future

Following Changes in Communities

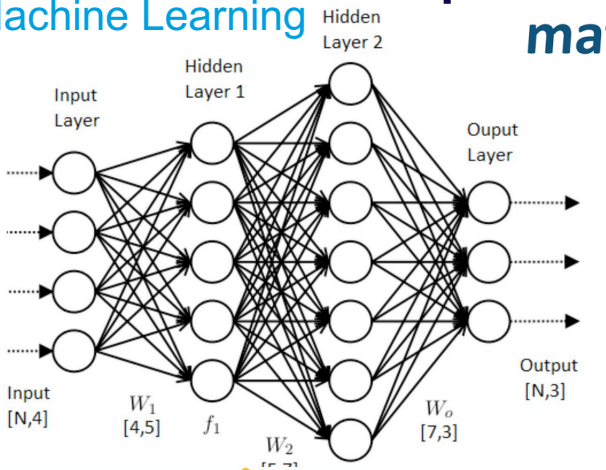
New Tools and Workflows Bring Communities Closer Together

Revisit diagram from beginning

- New tools from industry



- Rise of Machine Learning



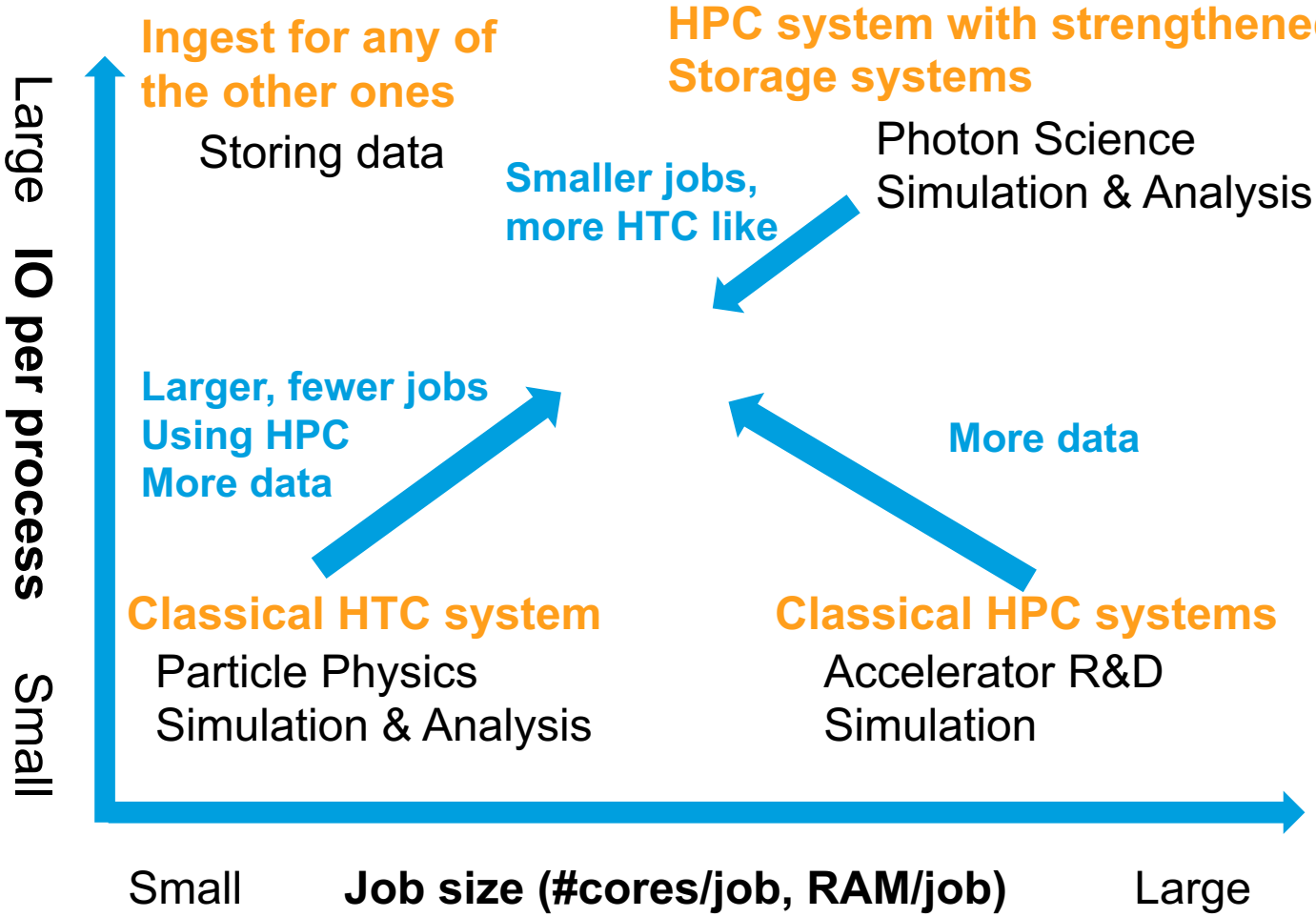
pandas

matplotlib

DASK

K

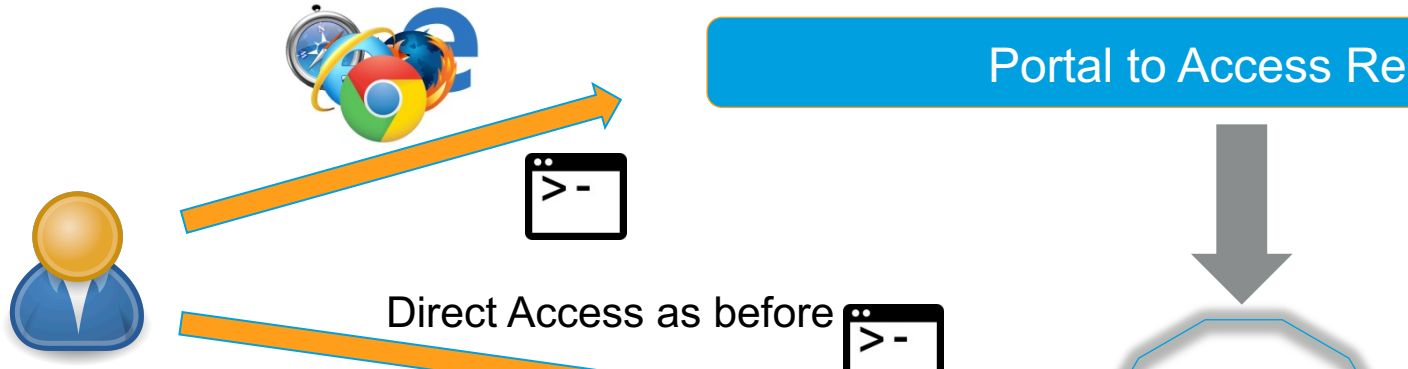
PyTorch



Merging of Infrastructure

Drafting the IDAF

- Unify infrastructure to use one interface



Portal to Access Resources

Direct Access as before

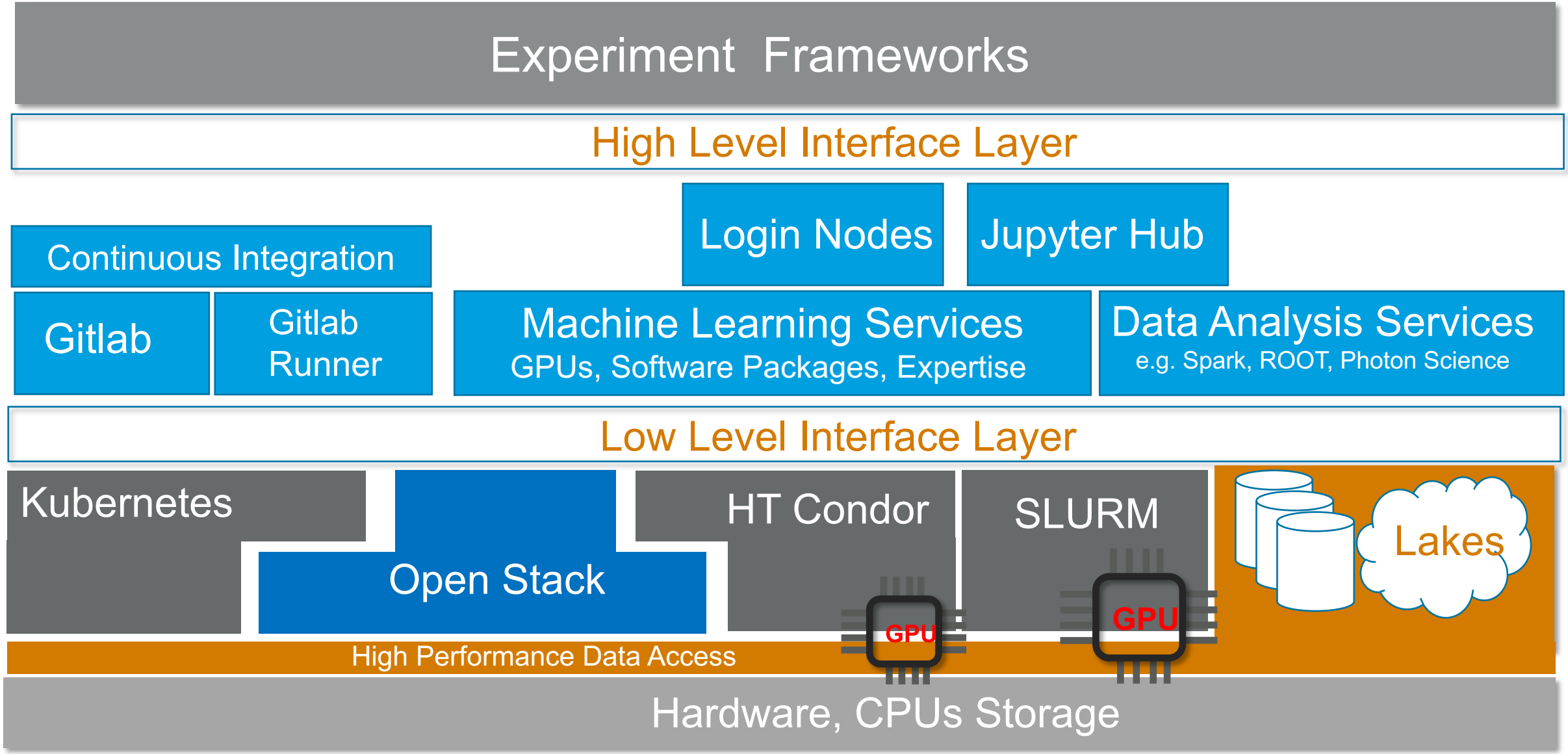
- Cloud/Container technologies for payload
- CI/CD style Workflows
- Automatic Submission to Resource of Choice
- Workflows based on messages

Maxwell Cluster

NAF & Grid Cluster

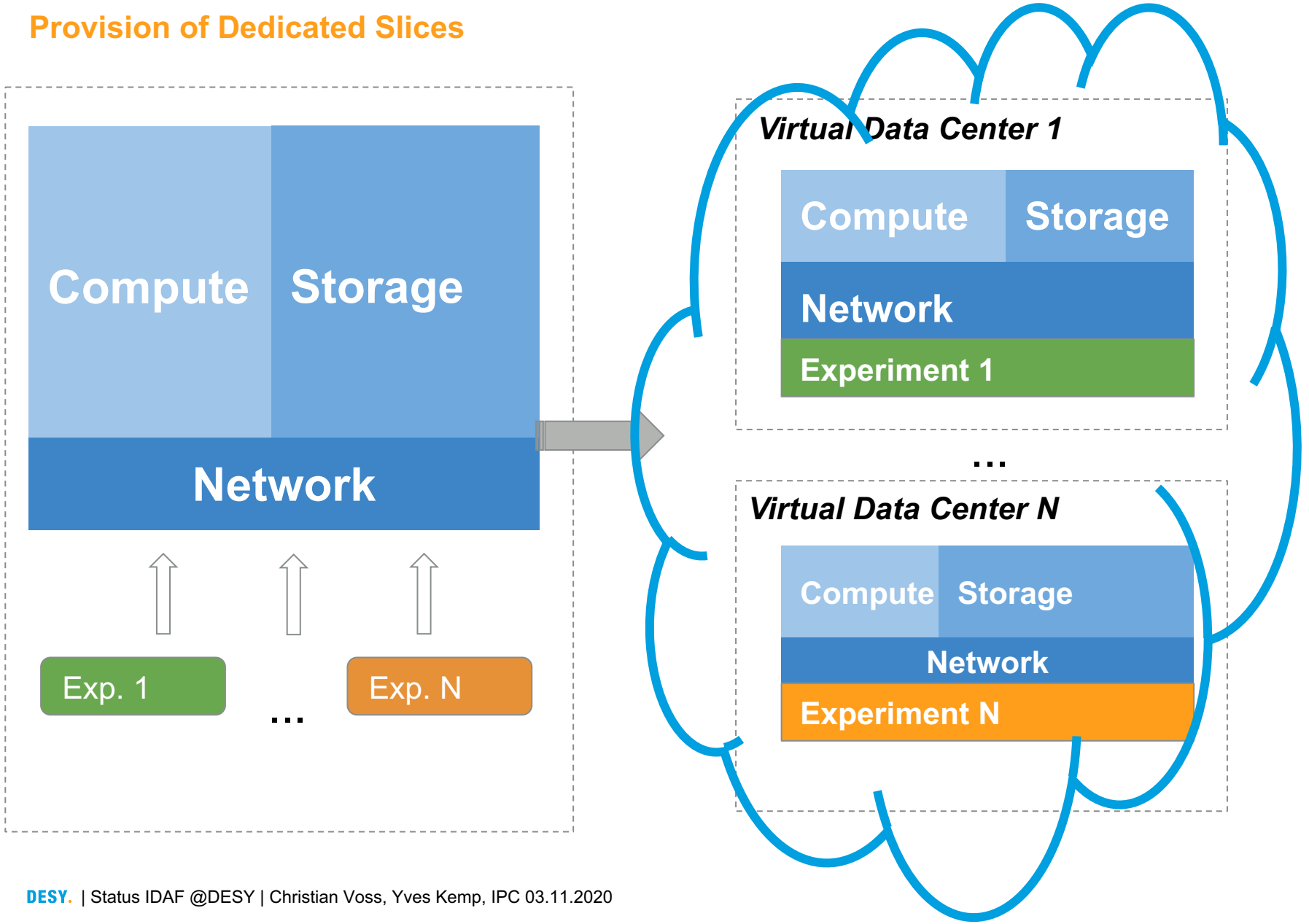
HTCondor
High Throughput Computing

Technical Design Layout



Changes to Interfaces for Beam Times at PETRA III and FLASH

Provision of Dedicated Slices



Dynamic provisioning:

- Cloud mechanisms
- Container orchestration

Infrastructure

- Cloud storage, compute and network
- Local and federated resources

Users

- Local, national, international, ad-hoc, proposal, MoU, ...

Summary & Outlook

Interdisciplinary

DESY IT already serves all branches of Science at DESY
- Infrastructures are there, and working well ...

Data

Science produces large amount of data
- Detectors, Acclerators and Simulation

Analysis

Main goal is to provide best possible analysis infrastructure for all our users.
- Large scale offline, and fast online ... overcome online/offline barrier for analysis

Facility

Not an institute cluster: *Facility* for internal *and* external users
- With state of the art resources and following technology evolution
- Providing more than just „compute access“: CI/CD, Jupyter, Container&Orchestration...