Federated Heterogenous Compute and Storage Infrastructure for the PUNCH4NFDI Consortium

CHEP 2023 - Norfolk, VA, USA

Manuel Giffels (KIT), Alexander Drabent (TLS), Matthias Hoeft (TLS), Benoit Roland (KIT), Dominik Schwarz (Uni Bielefeld), Christoph Wissing (DESY)

What is PUNCH4NFDI

- **P**articles, **U**niverse, **N**u**C**lei and **H**adrons for the NFDI
- Consortium in the National Research Data Infrastructure (Germany) (in german: **N**ationale **F**orschungs**d**aten**i**nfrastruktur - NFDI)
- Related scientific fields facing similar data analysis challenges
  - **increasing amount of data** generated by research infrastructures
  - complex algorithms yield to a **high demand of compute resource**
- Benefit from experiences, concepts and tools available in diverse communities

Prime goal is the setup of a federated and "FAIR" science data platform, offering the infrastructures and interfaces necessary for the access to and use of data and computing resources of the involved communities and beyond.

In this contribution:

Federate the considerable amount of available heterogenous compute and storage infrastructures in Germany and provide unified and seamless access to it

What is PUNCH4NFDI

- **P**articles, **U**niverse, **N**u**C**lei and **H**adrons for the NFDI
- Consortium in the National Research Data Infrastructure (Germany) (in german: **N**ationale **F**orschungs**d**aten**i**nfrastruktur - NFDI)
- Related scientific fields facing similar data analysis challenges
  - **increasing amount of data** generated by research infrastructures
  - complex algorithms yield to a **high demand of compute resource**
- Benefit from experiences, concepts and tools available in diverse communities

Prime goal is the setup of a federated and "FAIR platform, offering the infrastructures and interfaces necessary for t and computing resources of the involved communitie

Interested to hear more about PUNCH4NFDI
Tomorrow: Christiane is talking about
„First results from the PUNCH4NFDI Consortium"

In this contribution:

Federate the considerable amount of available heterogenous compute and storage infrastructures in Germany and provide unified and seamless access to it

KIT Karlsruhe Institute of Technology

Who Are We?

DESY.

AIP

TECHNISCHE UNIVERSITÄT DRESDEN

UHH

LMU

JG|U

Max-Planck-Institut für Radioastronomie

TLS

MAX-PLANCK-INSTITUT FÜR KERNPHYSIK

RUB

CERN

DLR

Universität Bielefeld

RHEINISCHE FRIEDRICH-WILHELMS-UNIVERSITÄT BONN

tu technische universität dortmund

FIAS Frankfurt Institute for Advanced Studies

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

JÜLICH Forschungszentrum

UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

htw Hochschule für Technik und Wirtschaft Berlin University of Applied Sciences

KIT Karlsruher Institut für Technologie

Φ DPG

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

UR

Universität Potsdam

GSI

RWTH AACHEN UNIVERSITY

h_da HOCHSCHULE DARMSTADT UNIVERSITY OF APPLIED SCIENCES

HZDR HELMHOLTZ ZENTRUM DRESDEN ROSSENDORF

MPCDF MAX PLANCK COMPUTING & DATA FACILITY

PTB Messen ▪ Forschen ▪ Wissen

TUM

UNI FREIBURG

HUMBOLDT-UNIVERSITÄT ZU BERLIN

Leibniz-Institut für Sonnenphysik (KIS)

lrz

Westfälische Wilhelms-Universität Münster

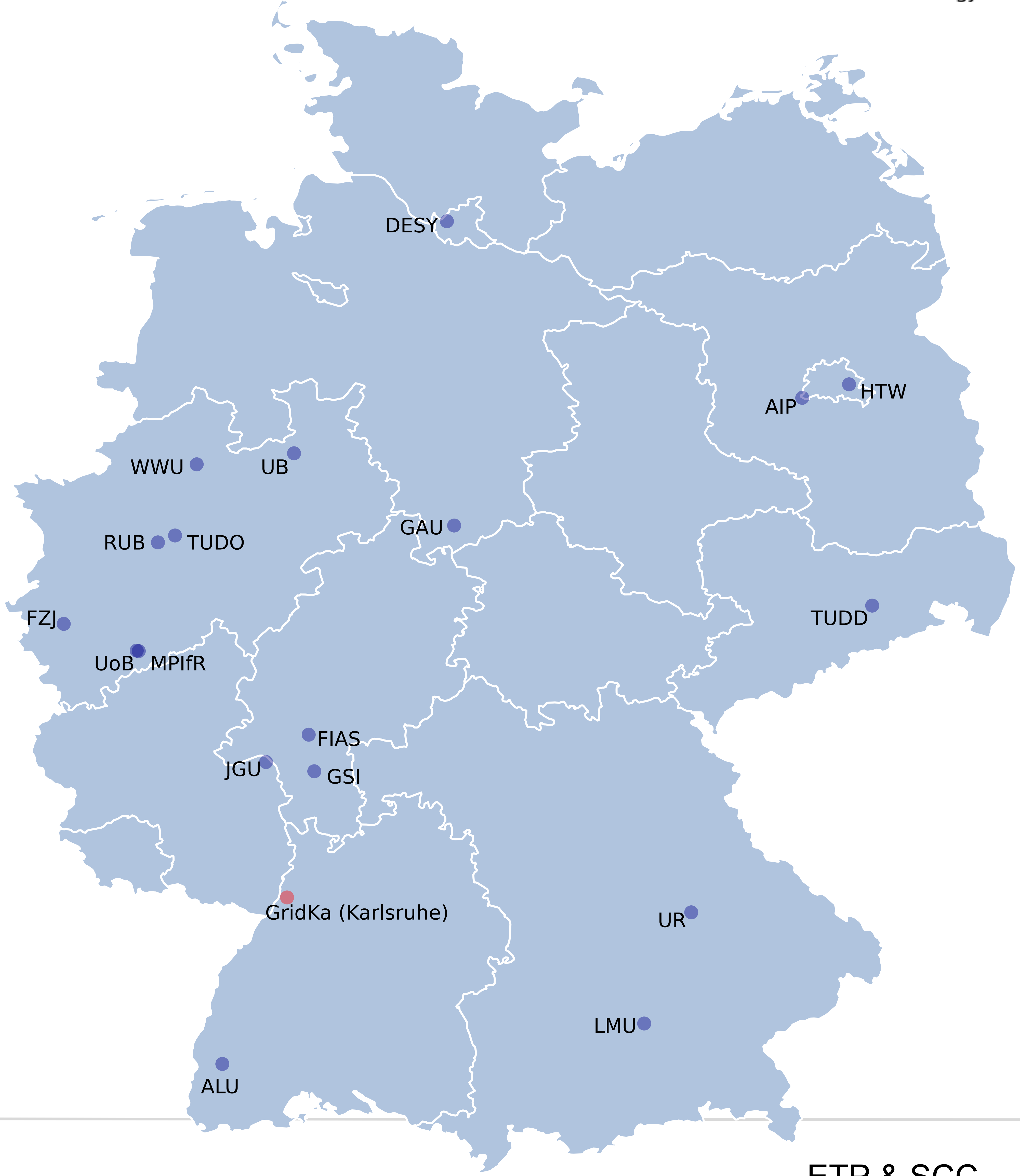
TIB

BERGISCHE UNIVERSITÄT WUPPERTAL

TECHNISCHE UNIVERSITÄT DARMSTADT

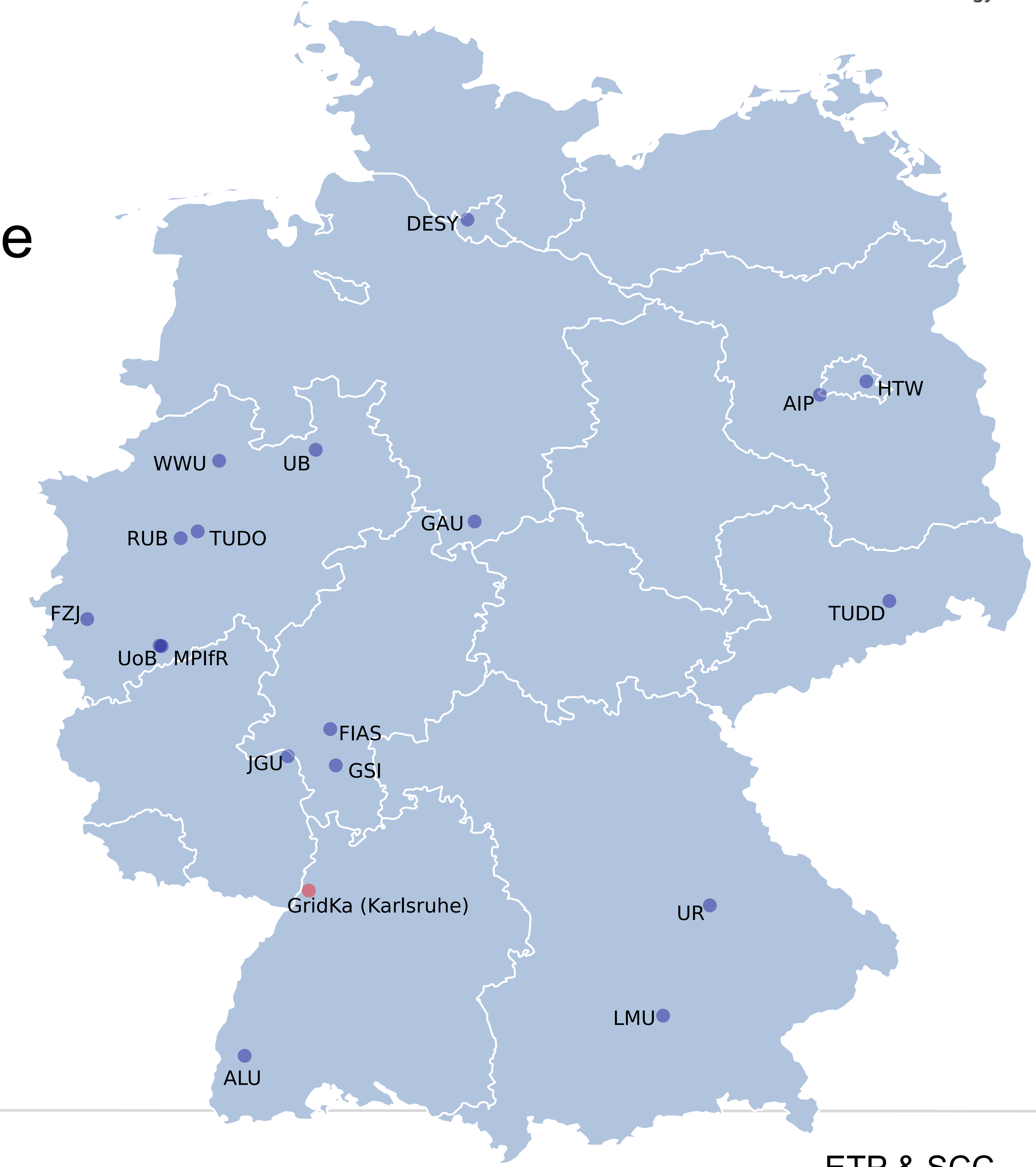
Available Resources of PUNCH4NFDI Institutions

- Substantial amount of HTC, HPC, Cloud compute resources are provided to PUNCH4NFDI



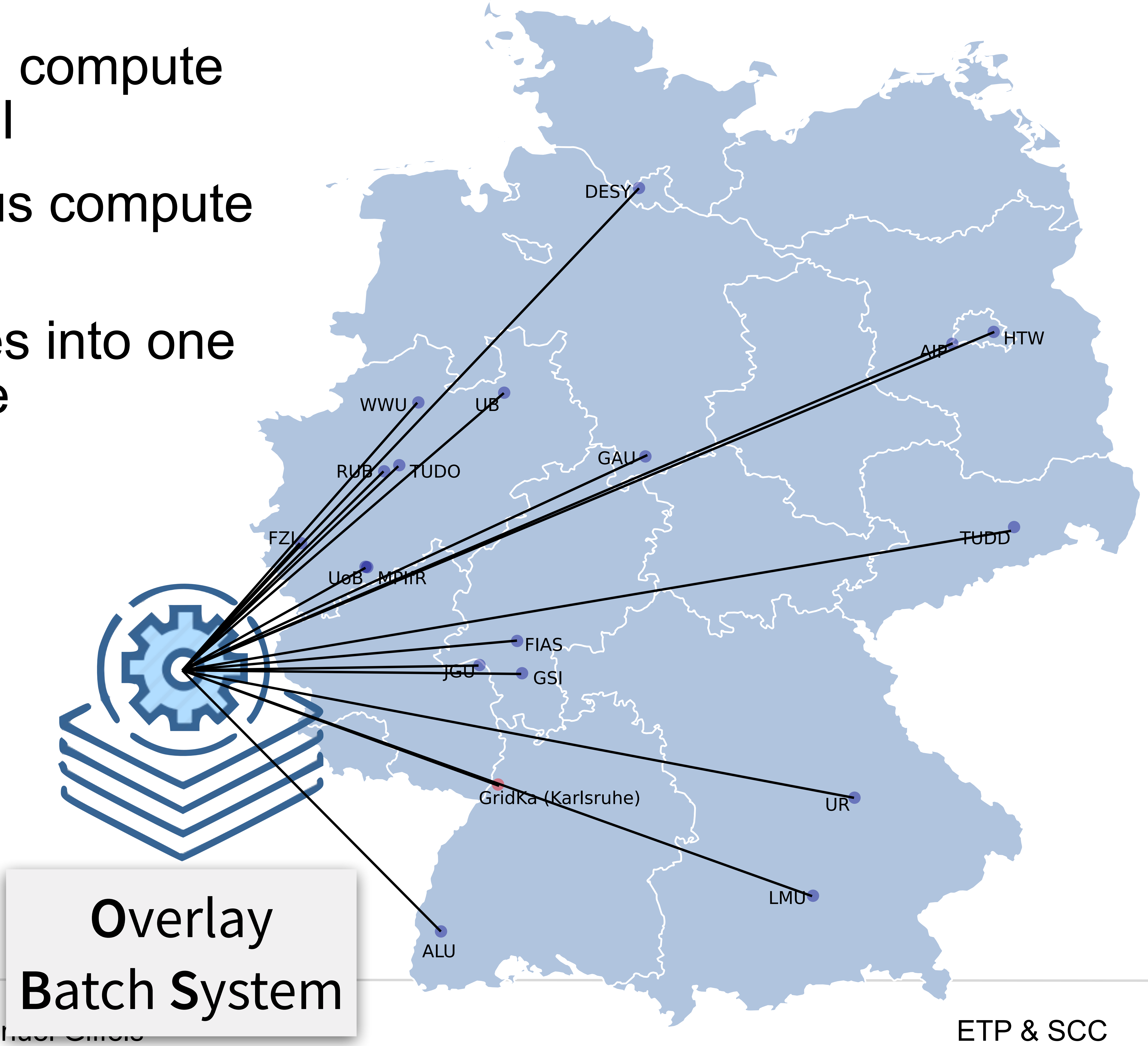
Available Resources of PUNCH4NFDI Institutions

- Substantial amount of HTC, HPC, Cloud compute resources are provided to PUNCH4NFDI
- **Idea:** Establish a federated heterogenous compute infrastructure for PUNCH4NFDI



Available Resources of PUNCH4NFDI Institutions

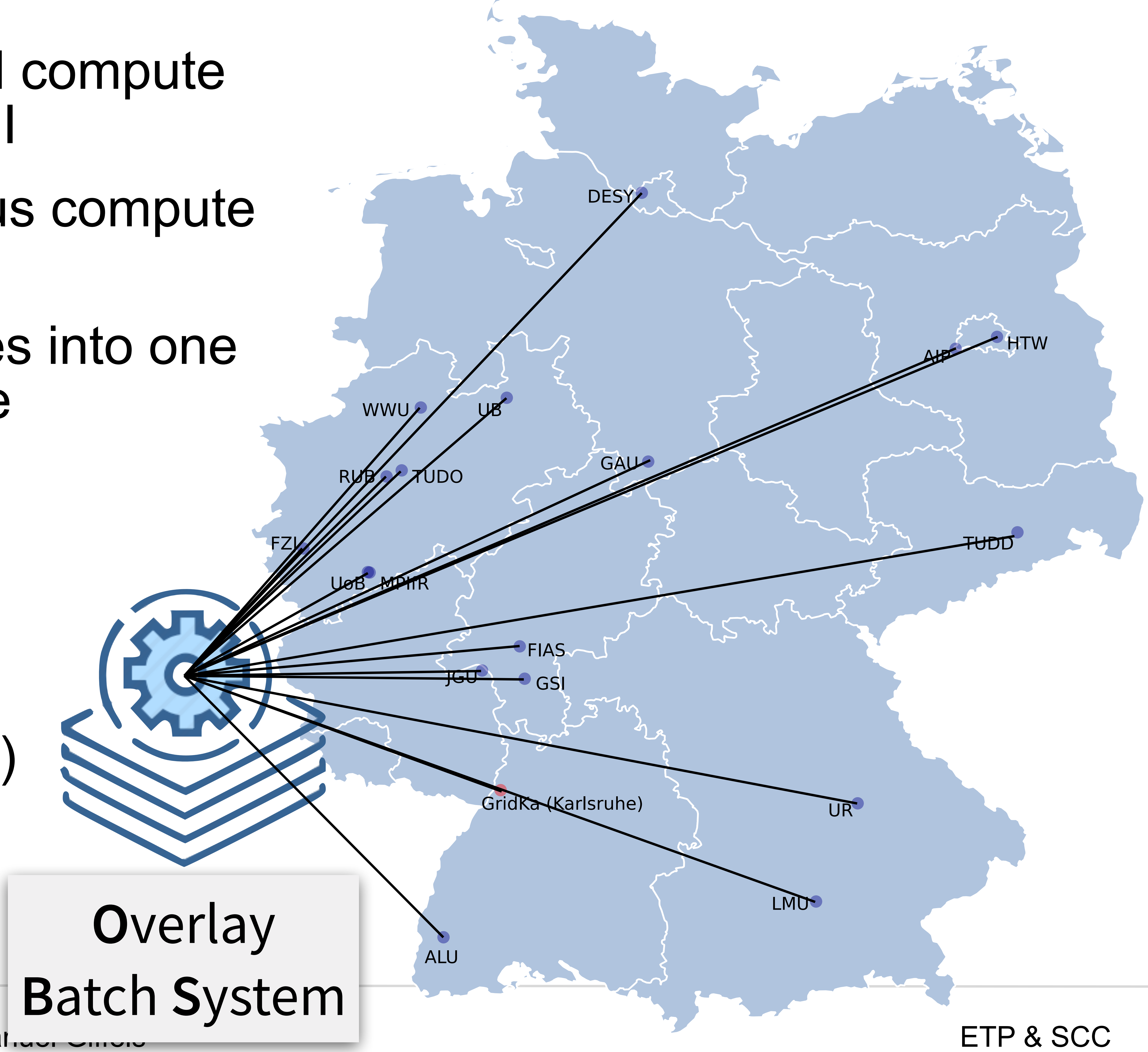
- Substantial amount of HTC, HPC, Cloud compute resources are provided to PUNCH4NFDI
- **Idea:** Establish a federated heterogenous compute infrastructure for PUNCH4NFDI
- Dynamically integrate compute resources into one so called overlay batch system using the `COBalD/TARDIS` meta scheduler [1,2]



Overlay Batch System

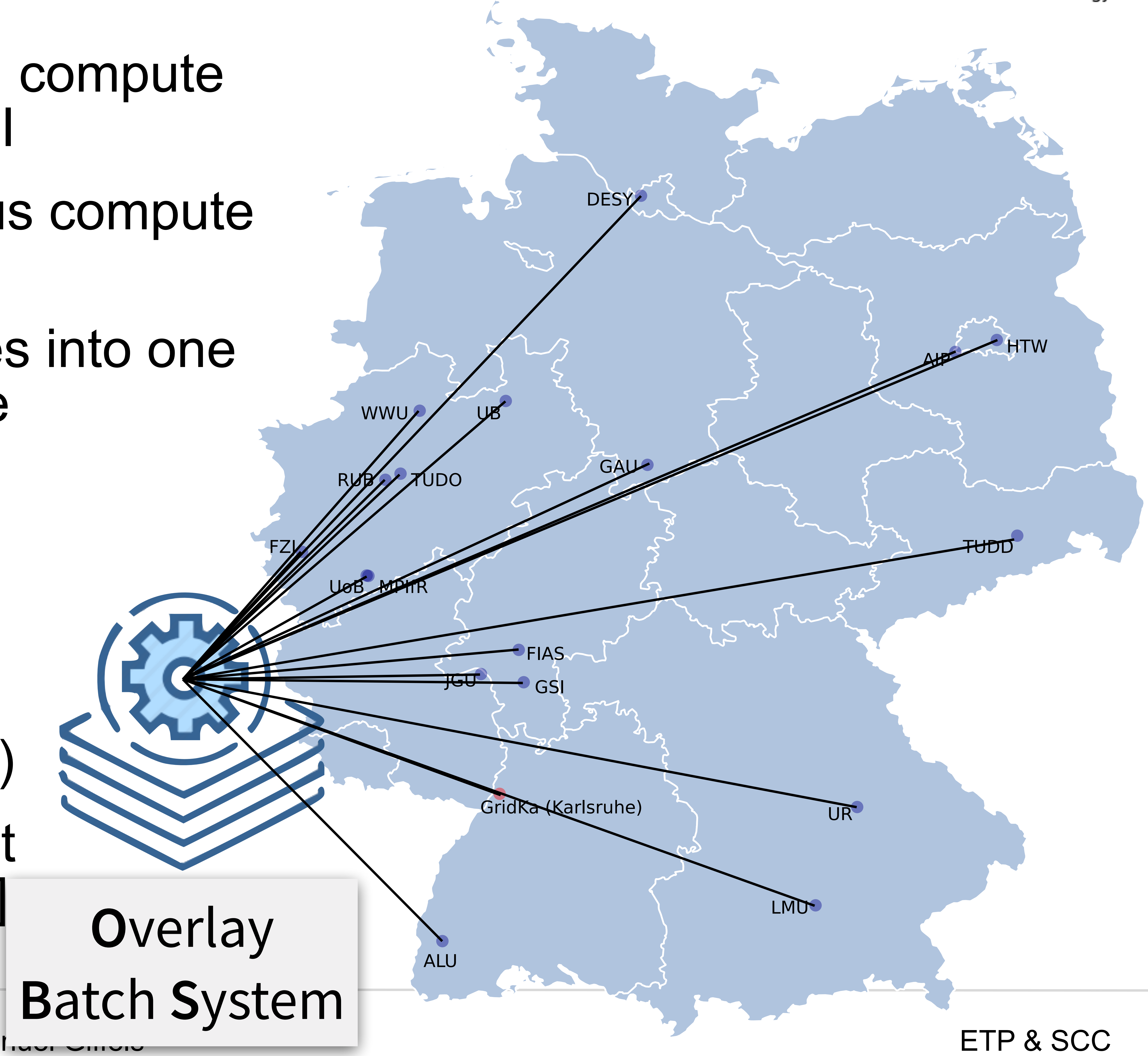
Available Resources of PUNCH4NFDI Institutions

- Substantial amount of HTC, HPC, Cloud compute resources are provided to PUNCH4NFDI
- **Idea:** Establish a federated heterogenous compute infrastructure for PUNCH4NFDI
- Dynamically integrate compute resources into one so called overlay batch system using the `COBalD/TARDIS` meta scheduler [1,2]
- Provide single point(s) of entry to users:
  - Traditional login nodes (available)
  - JupyterHubs (in development)
  - Grid Compute Elements (if necessary)



Available Resources of PUNCH4NFDI Institutions

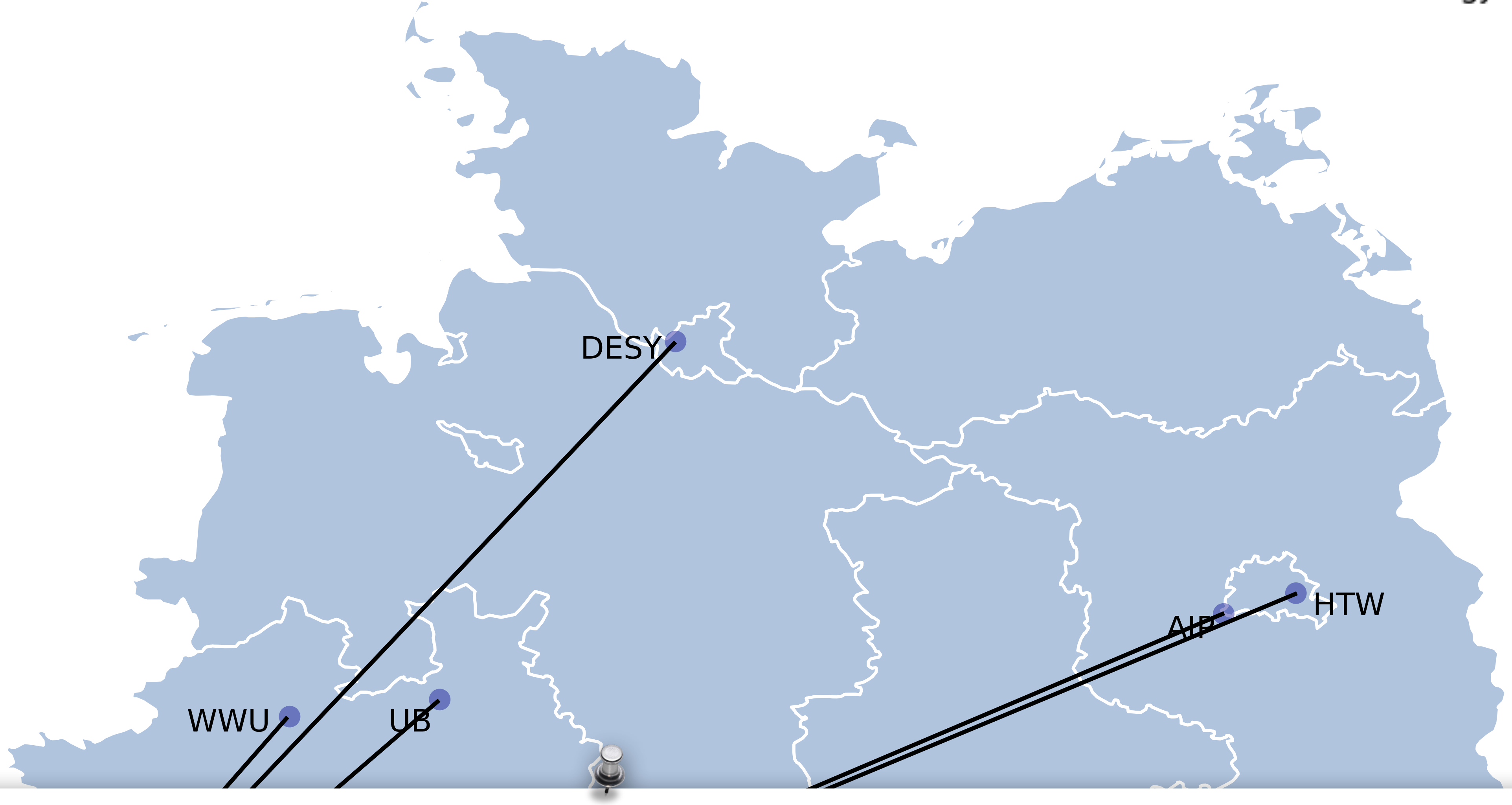
- Substantial amount of HTC, HPC, Cloud compute resources are provided to PUNCH4NFDI
- **Idea:** Establish a federated heterogenous compute infrastructure for PUNCH4NFDI
- Dynamically integrate compute resources into one so called overlay batch system using the `COBalD/TARDIS` meta scheduler [1,2]
- Provide single point(s) of entry to users:
  - Traditional login nodes (available)
  - JupyterHubs (in development)
  - Grid Compute Elements (if necessary)
- Provide necessary software environment using container technology + CVMFS [3]

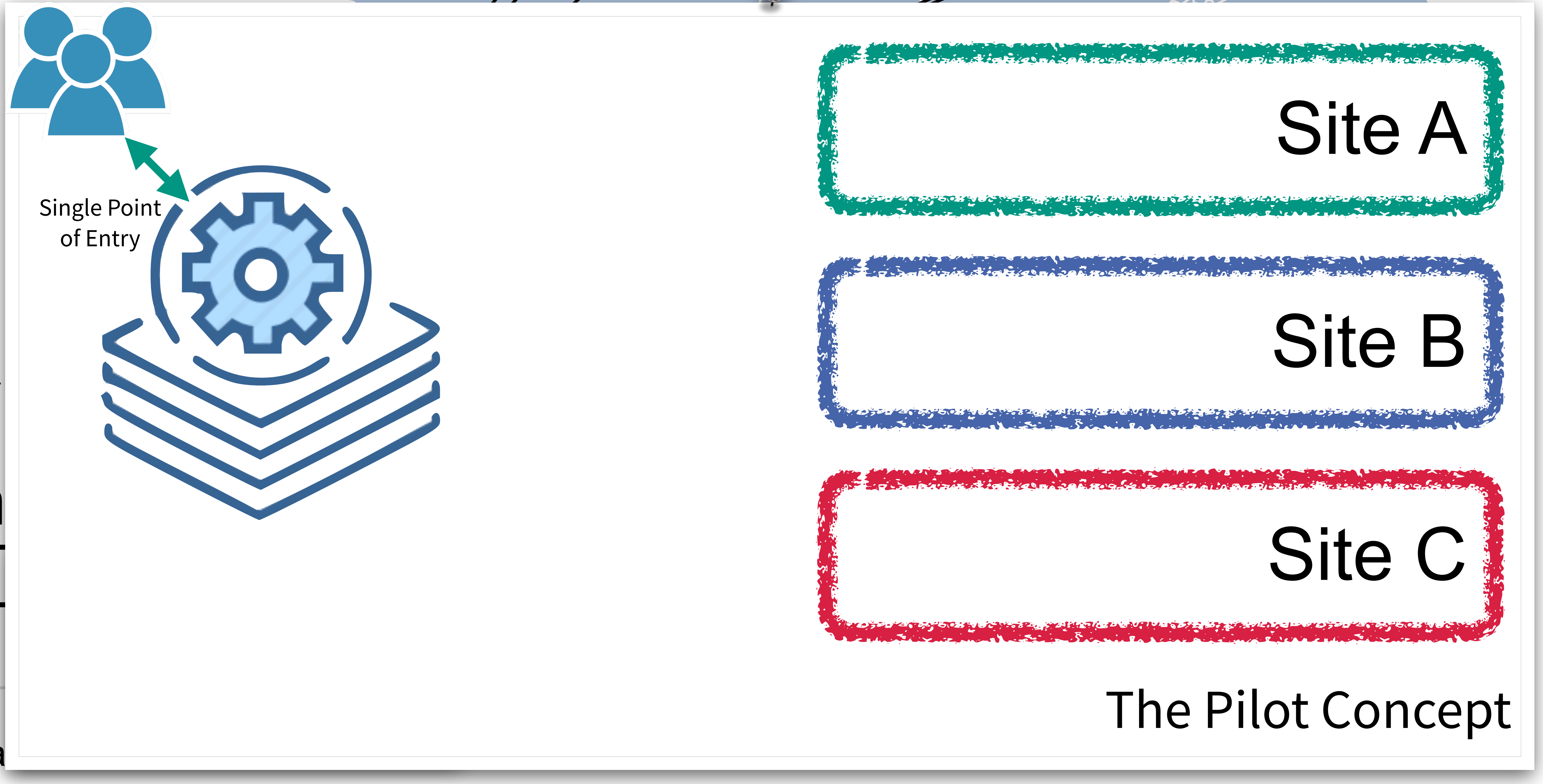


Overlay Batch System

Available Resources of PUNCH4NFDI Institutions

- Substantial amount of HTC, HPC, Cloud compute resources are provided to PUNCH4NFDI
- **Idea:** Establish a federated heterogenous compute infrastructure for PUNCH4NFDI
- Dynamically integrate compute resources into one so called overlay batch system using the `COBalD/TARDIS` meta scheduler [1,2]
- Provide single point(s) of entry to users:
  - Traditional login nodes (available)
  - JupyterHubs (in development)
  - Grid Compute Elements (if necessary
- Provide necessary software environmen using container technology + CVMFS [3

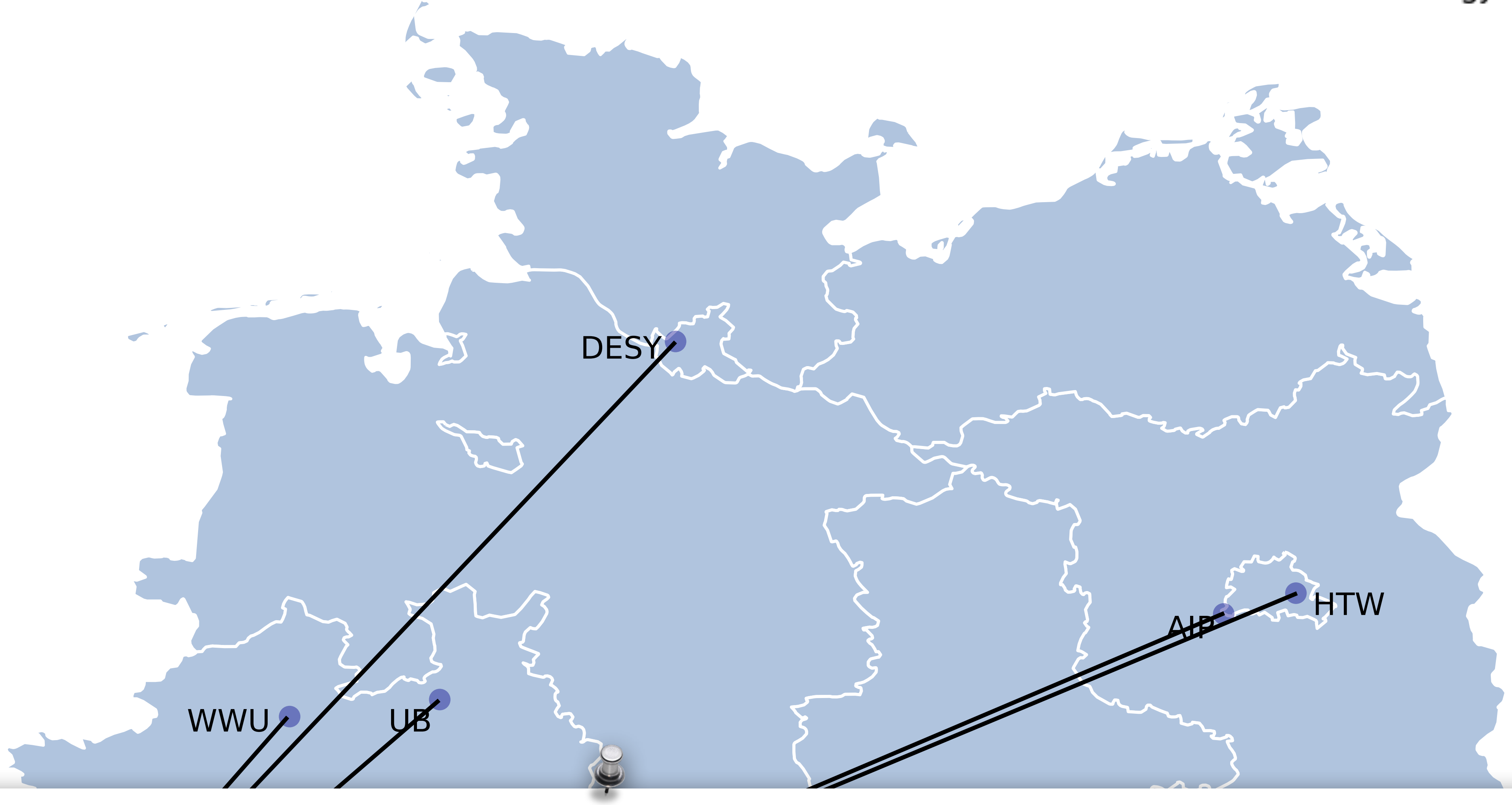


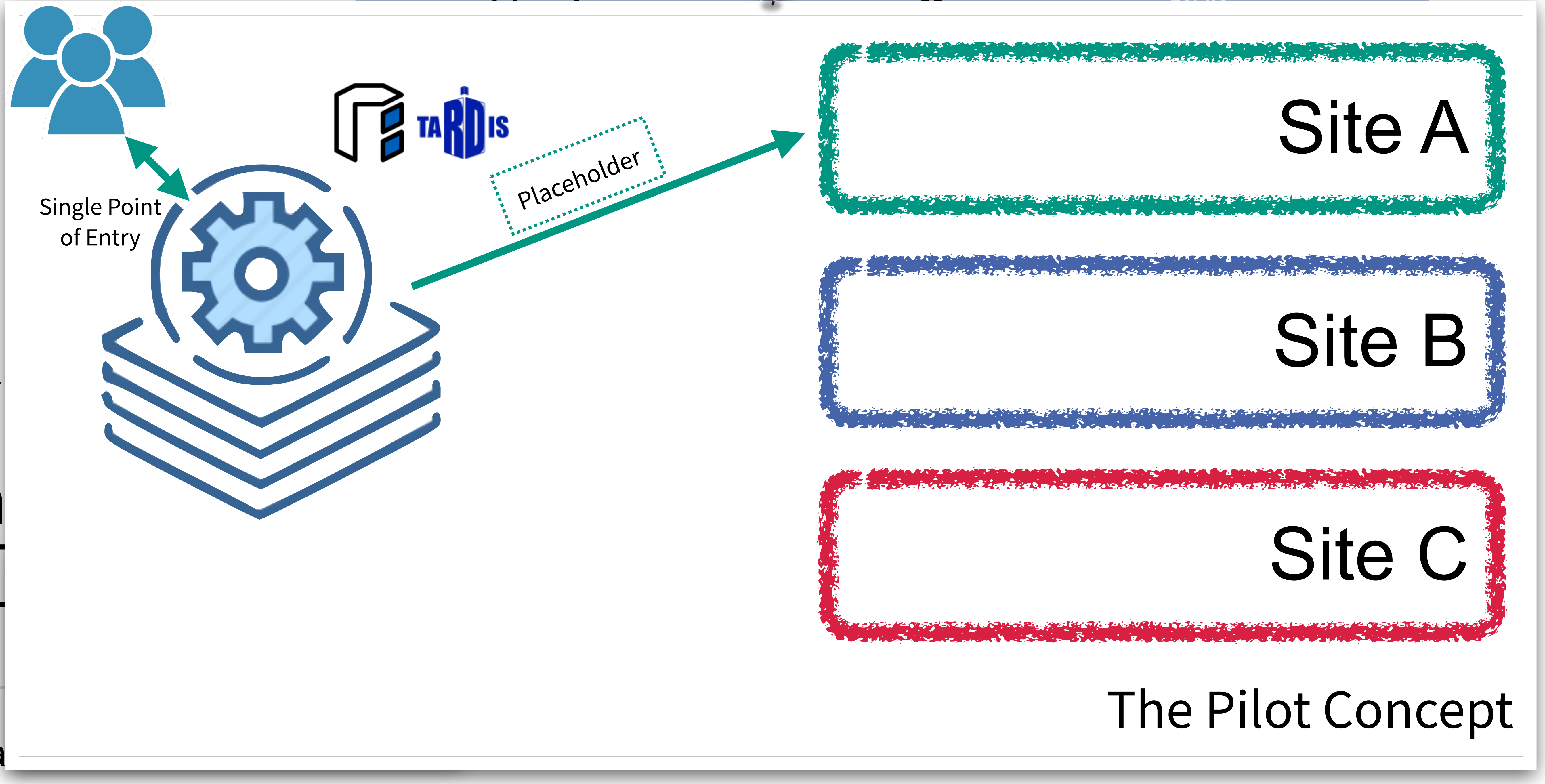


The Pilot Concept

Available Resources of PUNCH4NFDI Institutions

- Substantial amount of HTC, HPC, Cloud compute resources are provided to PUNCH4NFDI
- **<u>Idea:</u>** Establish a federated heterogenous compute infrastructure for PUNCH4NFDI
- Dynamically integrate compute resources into one so called overlay batch system using the `COBalD/TARDIS` meta scheduler [1,2]
- Provide single point(s) of entry to users:
  - Traditional login nodes (available)
  - JupyterHubs (in development)
  - Grid Compute Elements (if necessary
- Provide necessary software environmen using container technology + CVMFS [3

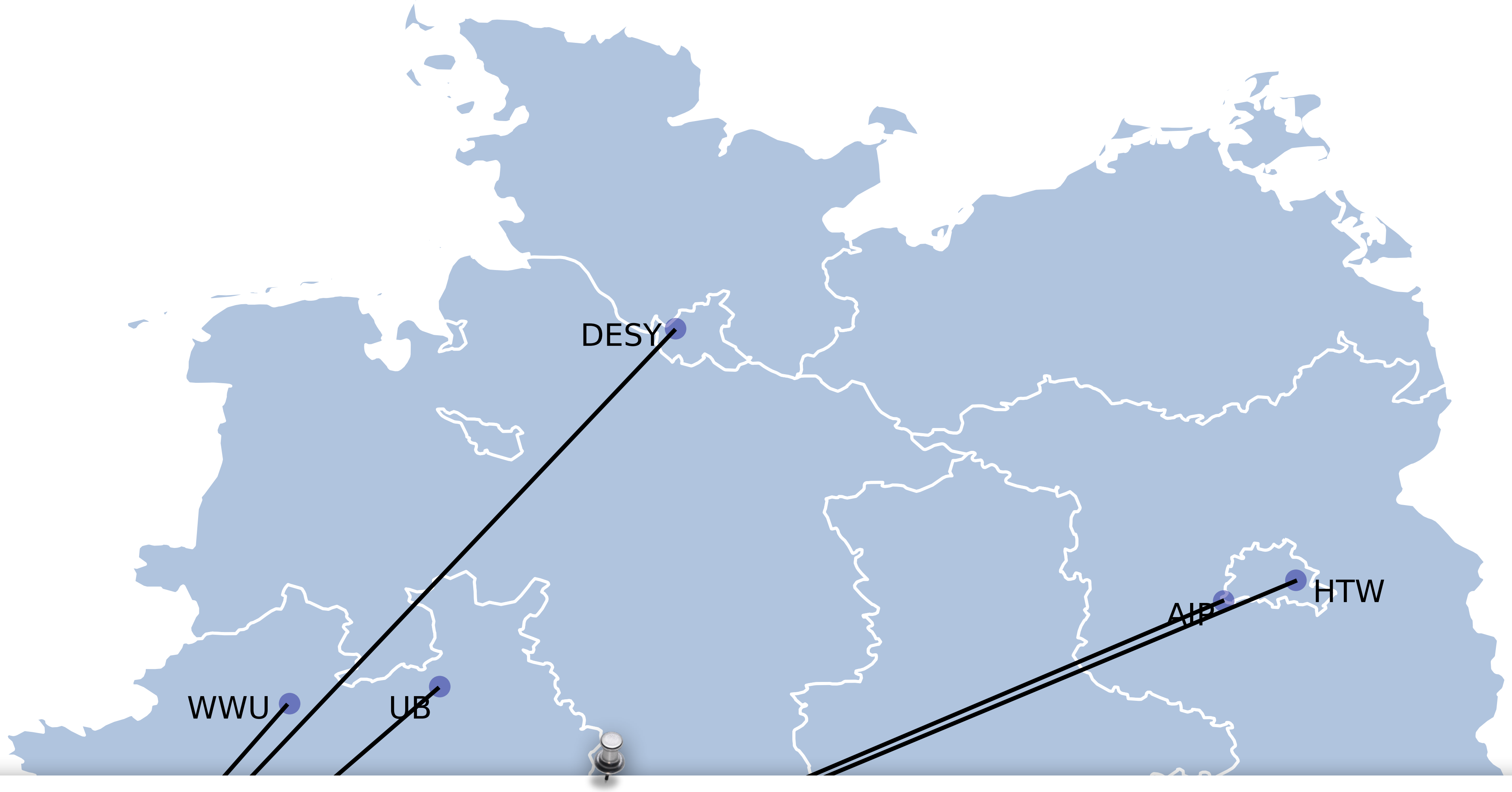


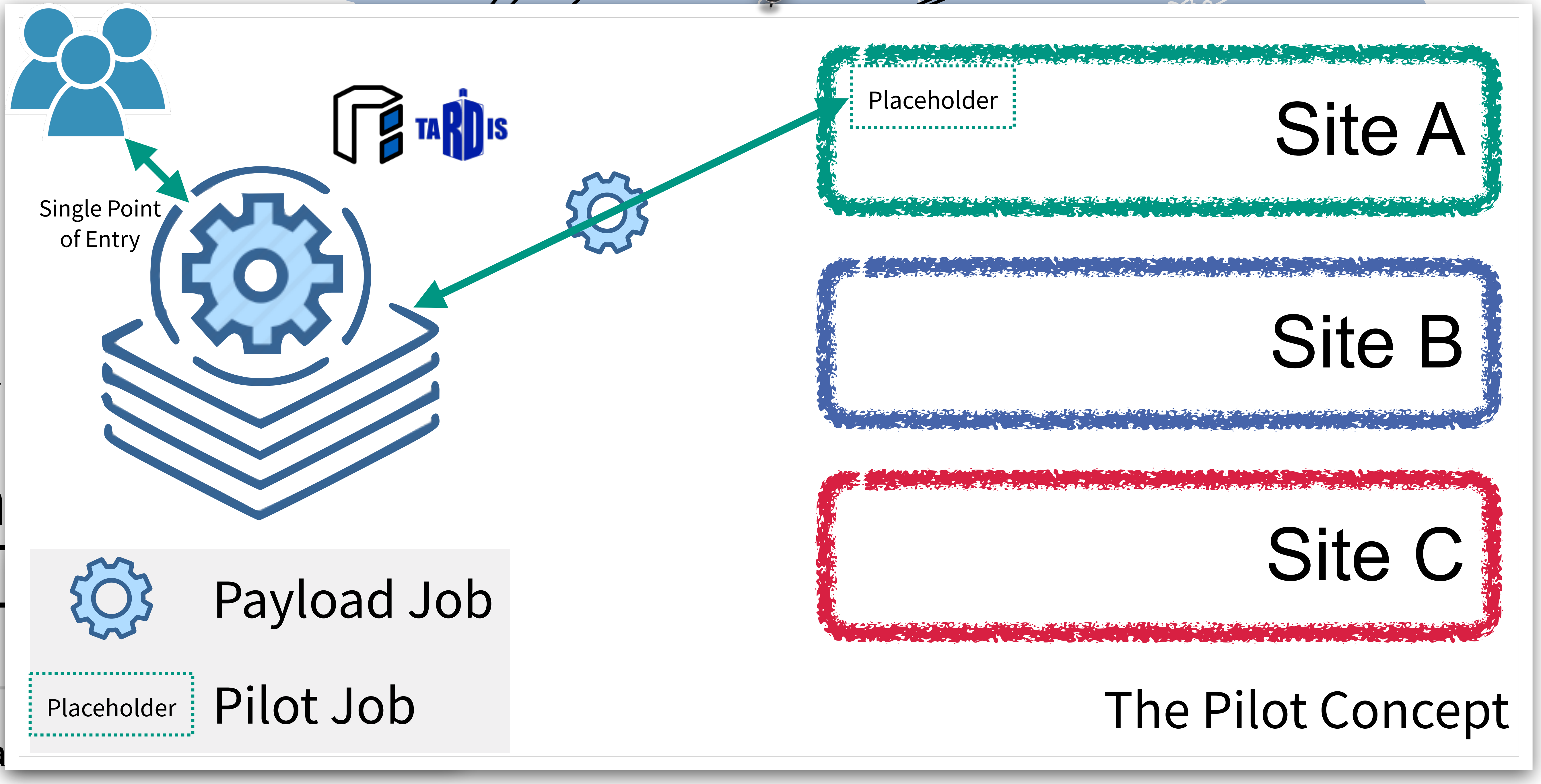


The Pilot Concept

Available Resources of PUNCH4NFDI Institutions

- Substantial amount of HTC, HPC, Cloud compute resources are provided to PUNCH4NFDI
- **Idea:** Establish a federated heterogenous compute infrastructure for PUNCH4NFDI
- Dynamically integrate compute resources into one so called overlay batch system using the `COBalD/TARDIS` meta scheduler [1,2]
- Provide single point(s) of entry to users:
  - Traditional login nodes (available)
  - JupyterHubs (in development)
  - Grid Compute Elements (if necessary
- Provide necessary software environmen using container technology + CVMFS [3

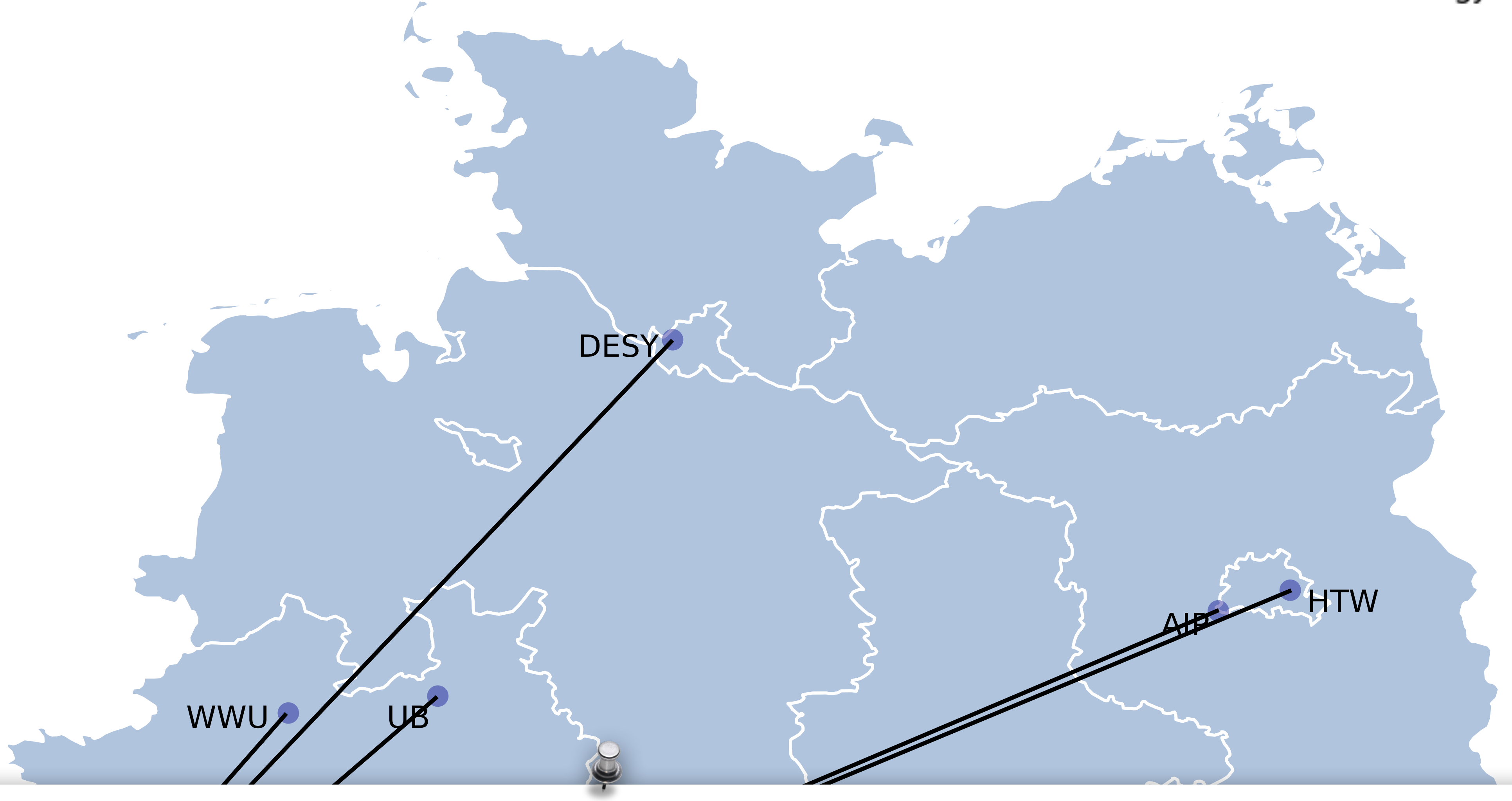


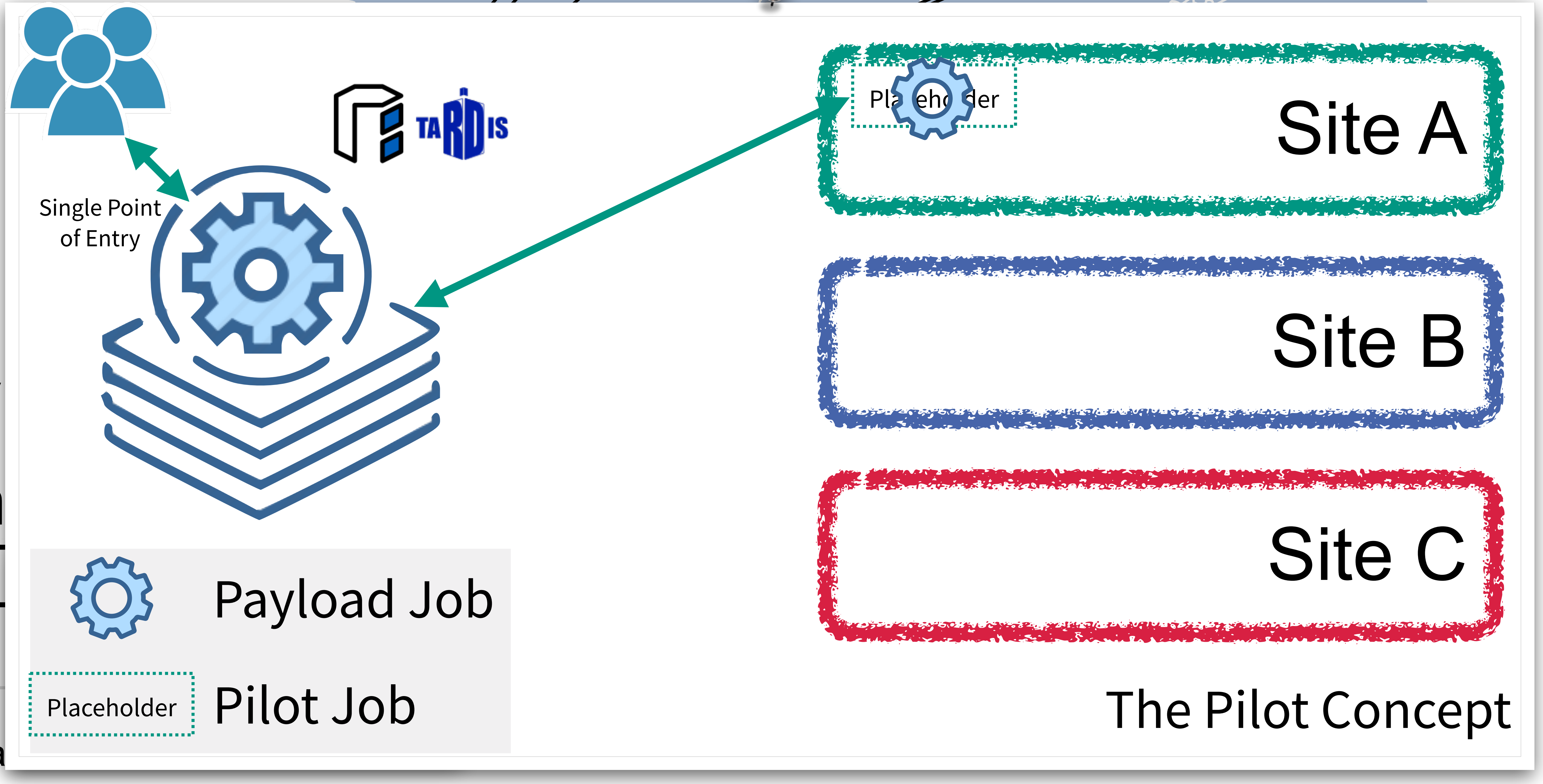


The Pilot Concept

Available Resources of PUNCH4NFDI Institutions

- Substantial amount of HTC, HPC, Cloud compute resources are provided to PUNCH4NFDI
- **Idea:** Establish a federated heterogenous compute infrastructure for PUNCH4NFDI
- Dynamically integrate compute resources into one so called overlay batch system using the `COBalD/TARDIS` meta scheduler [1,2]
- Provide single point(s) of entry to users:
  - Traditional login nodes (available)
  - JupyterHubs (in development)
  - Grid Compute Elements (if necessary
- Provide necessary software environmen using container technology + CVMFS [3

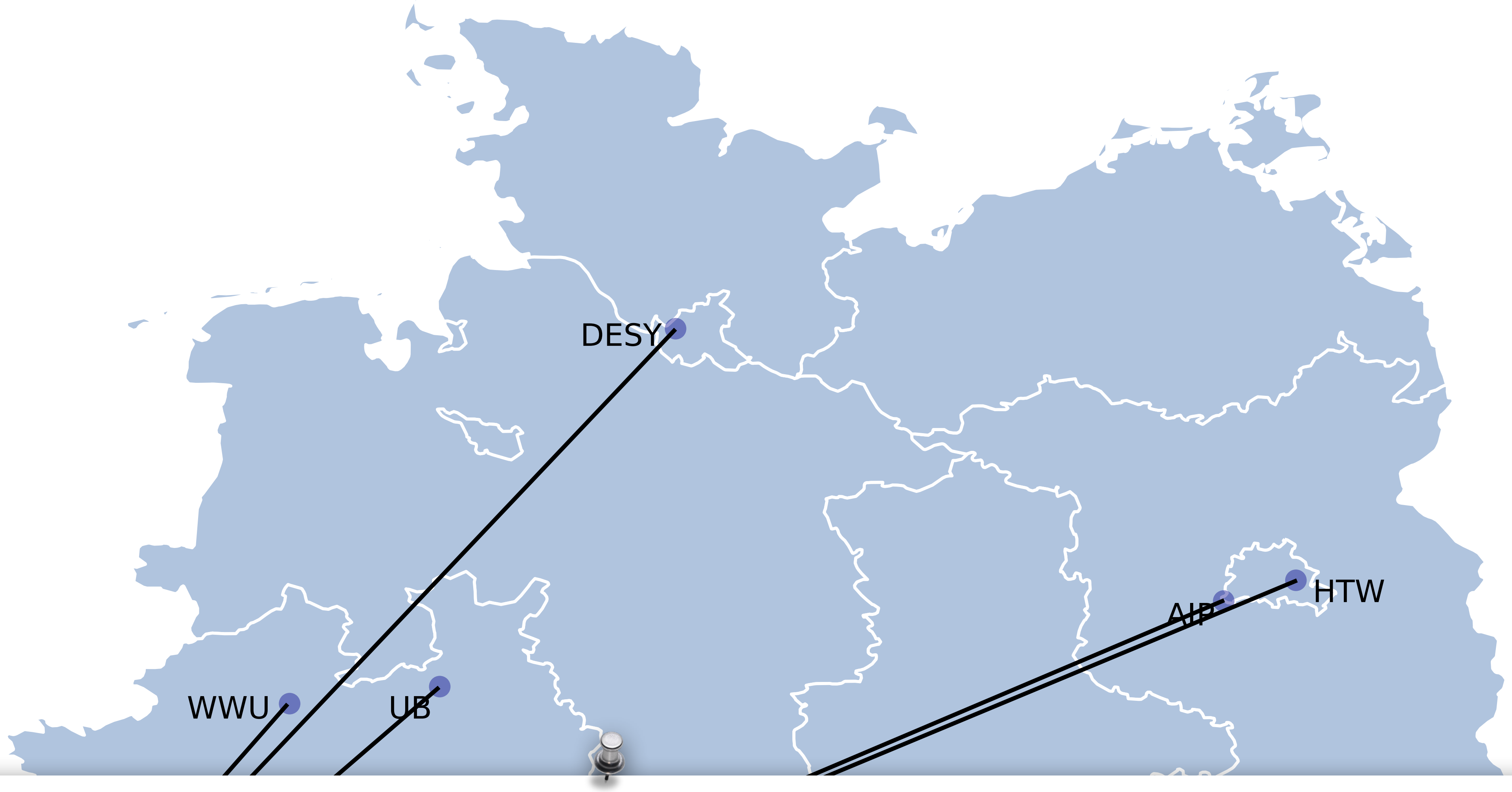


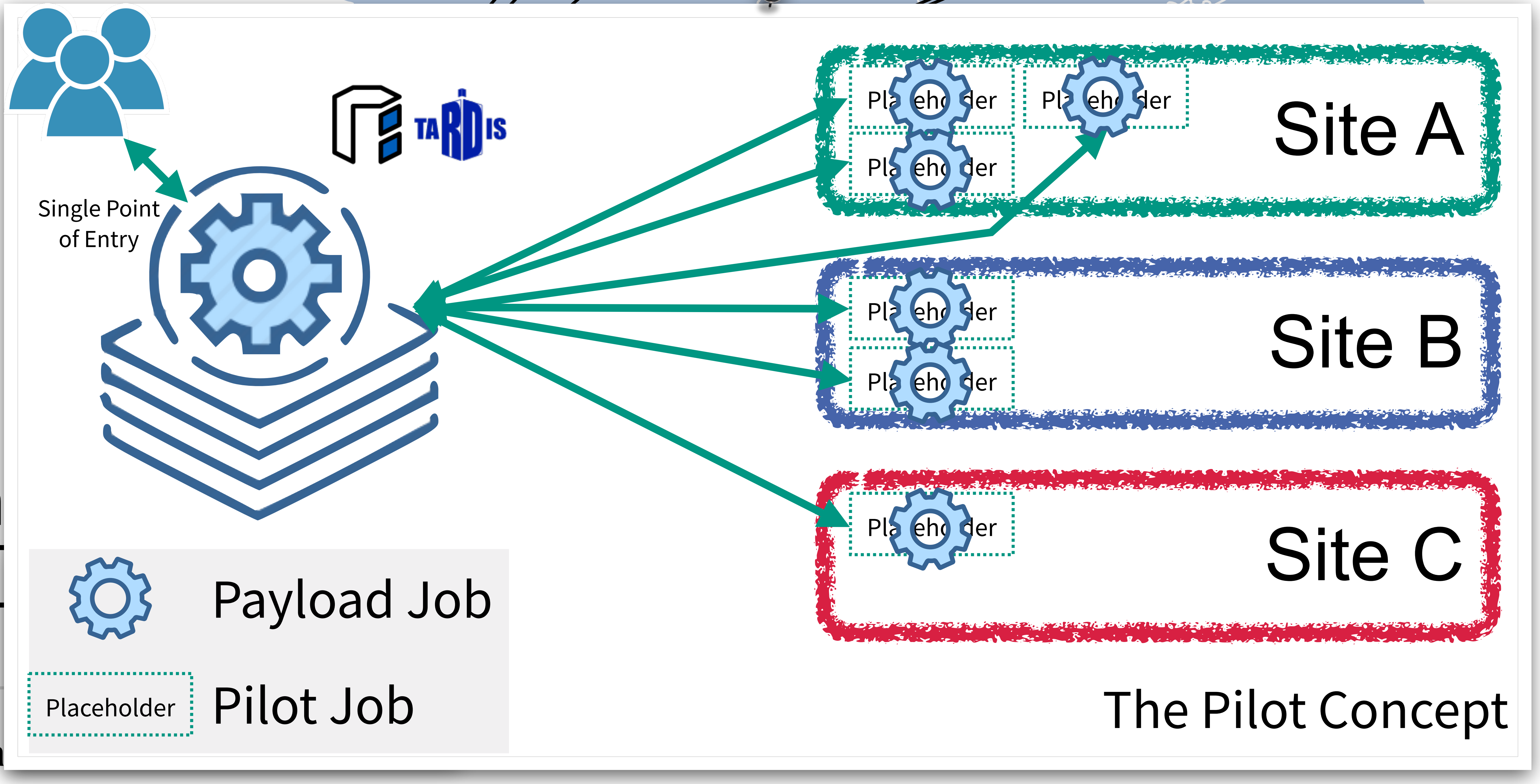


The Pilot Concept

Available Resources of PUNCH4NFDI Institutions

- Substantial amount of HTC, HPC, Cloud compute resources are provided to PUNCH4NFDI
- **Idea:** Establish a federated heterogenous compute infrastructure for PUNCH4NFDI
- Dynamically integrate compute resources into one so called overlay batch system using the `COBalD/TARDIS` meta scheduler [1,2]
- Provide single point(s) of entry to users:
  - Traditional login nodes (available)
  - JupyterHubs (in development)
  - Grid Compute Elements (if necessary
- Provide necessary software environmen using container technology + CVMFS [3

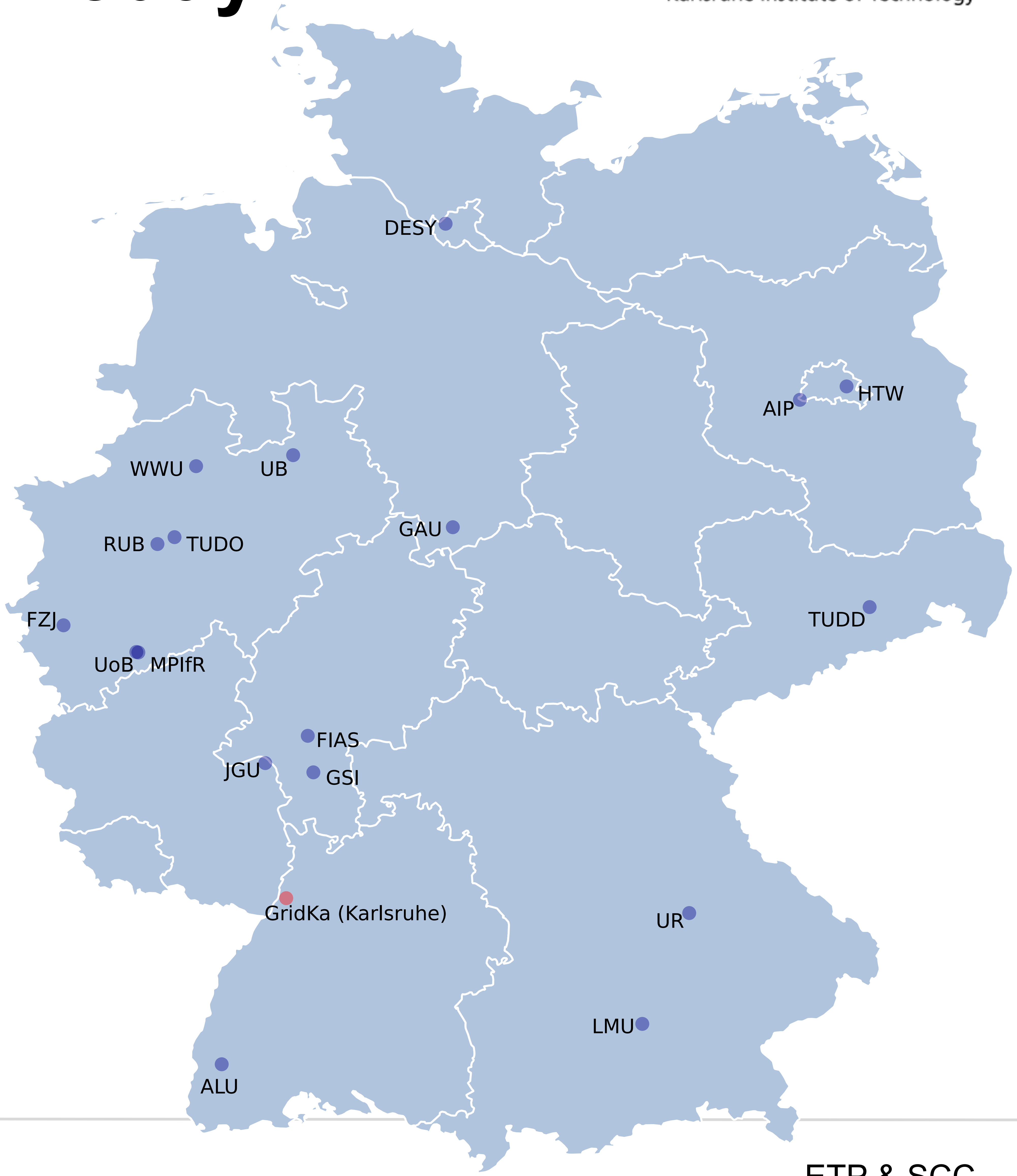




The Pilot Concept

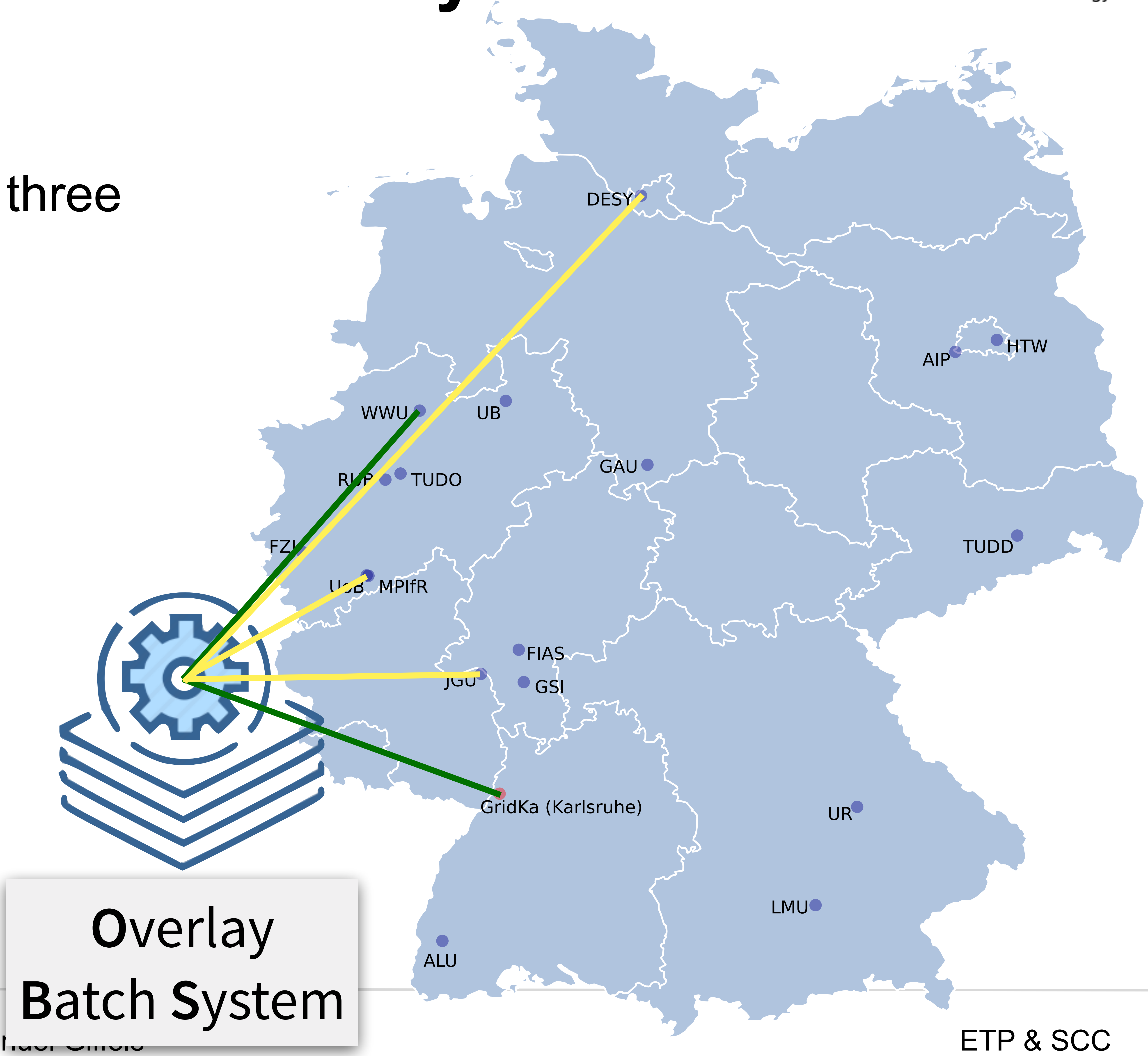
The Compute4PUNCH Infrastructure Today

- Prototype of federated Compute4PUNCH infrastructure is available



The Compute4PUNCH Infrastructure Today

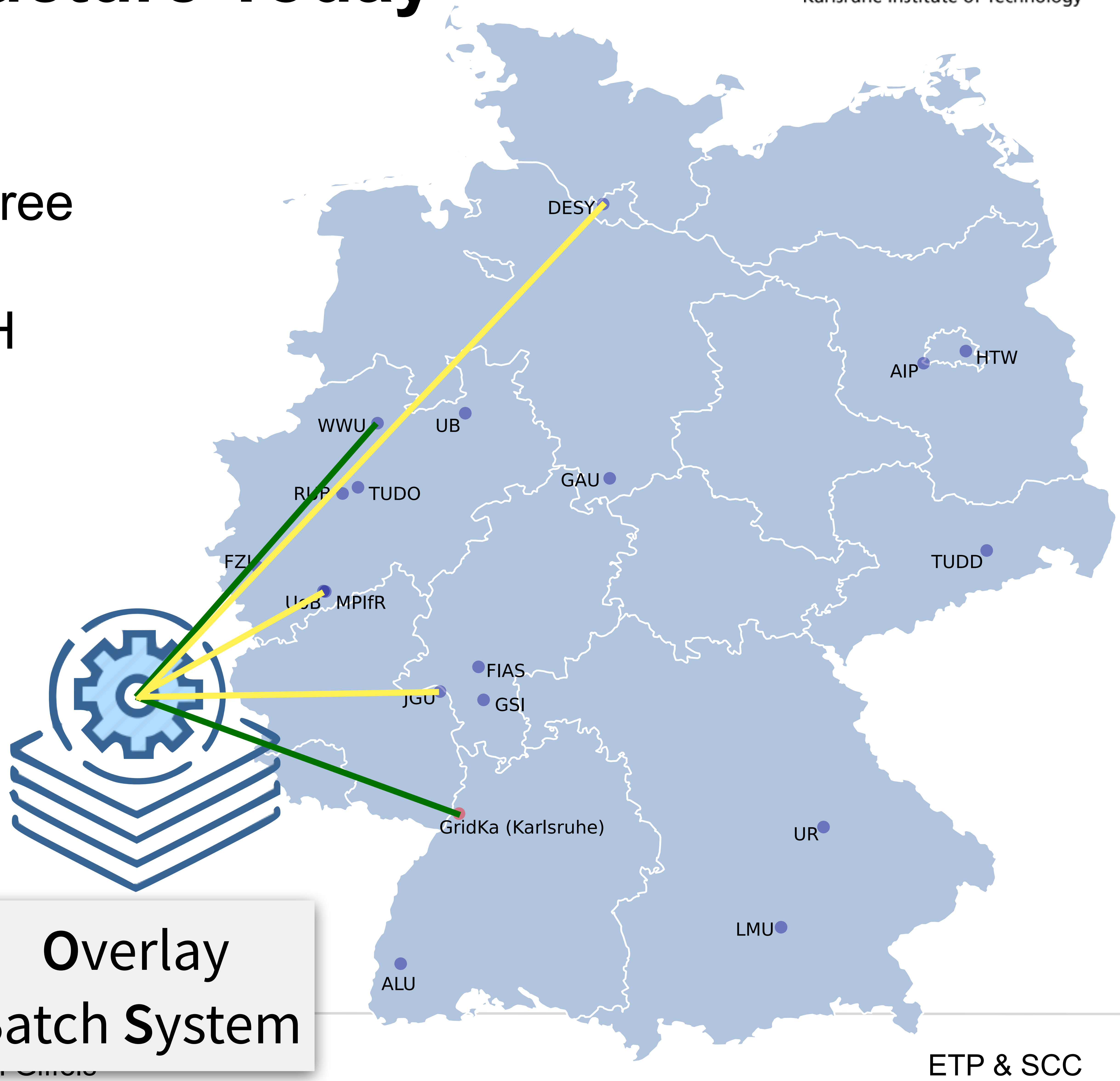
- Prototype of federated Compute4PUNCH infrastructure is available
- Dynamic integration of two compute sites, three more will follow soon



The Compute4PUNCH Infrastructure Today

- Prototype of federated Compute4PUNCH infrastructure is available
- Dynamic integration of two compute sites, three more will follow soon
- AAI based login node available to all PUNCH users

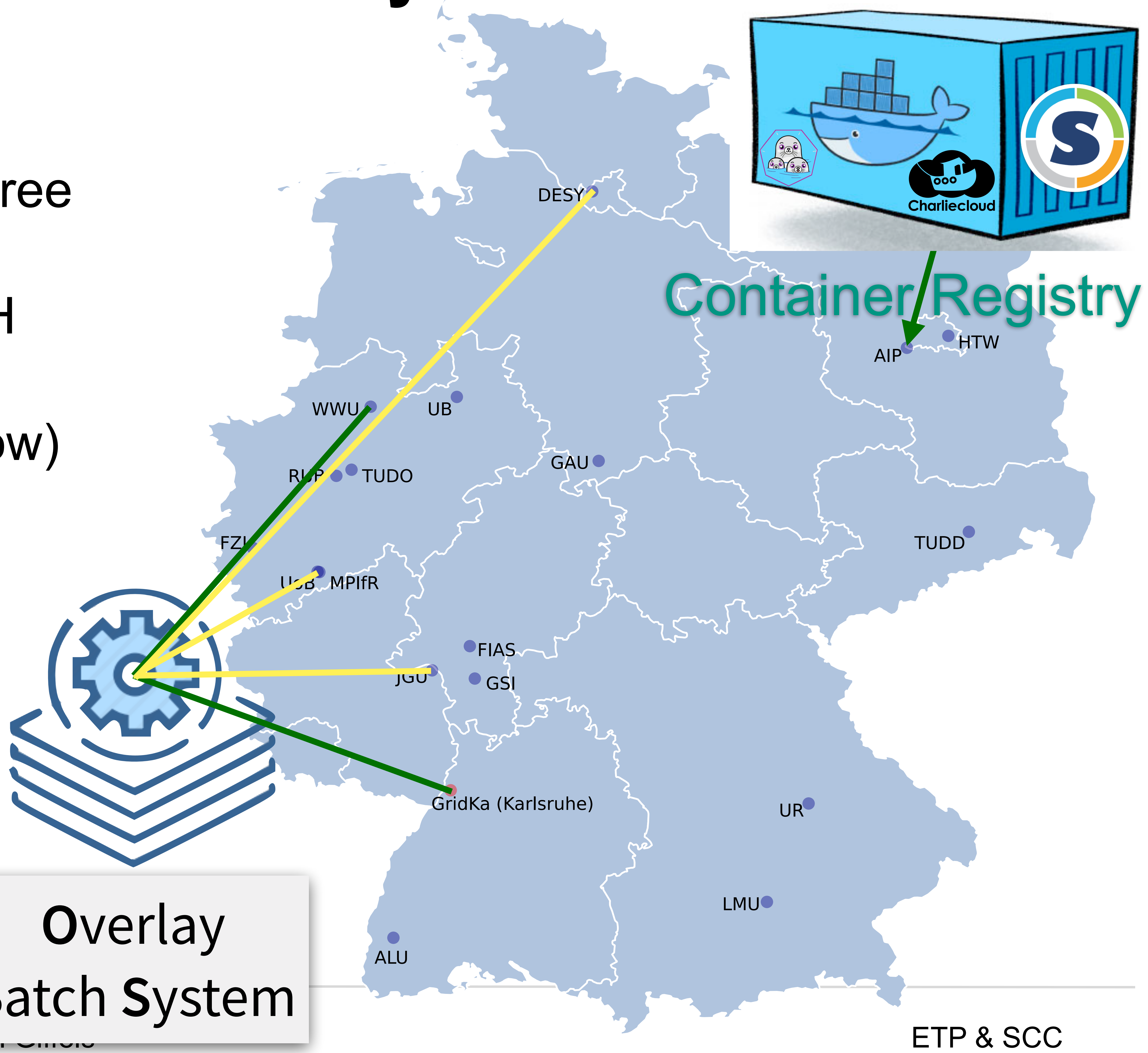
c4p-login.gridka.de



The Compute4PUNCH Infrastructure Today

- Prototype of federated Compute4PUNCH infrastructure is available
- Dynamic integration of two compute sites, three more will follow soon
- AAI based login node available to all PUNCH users
- Container registry available (+ CI/CD workflow)

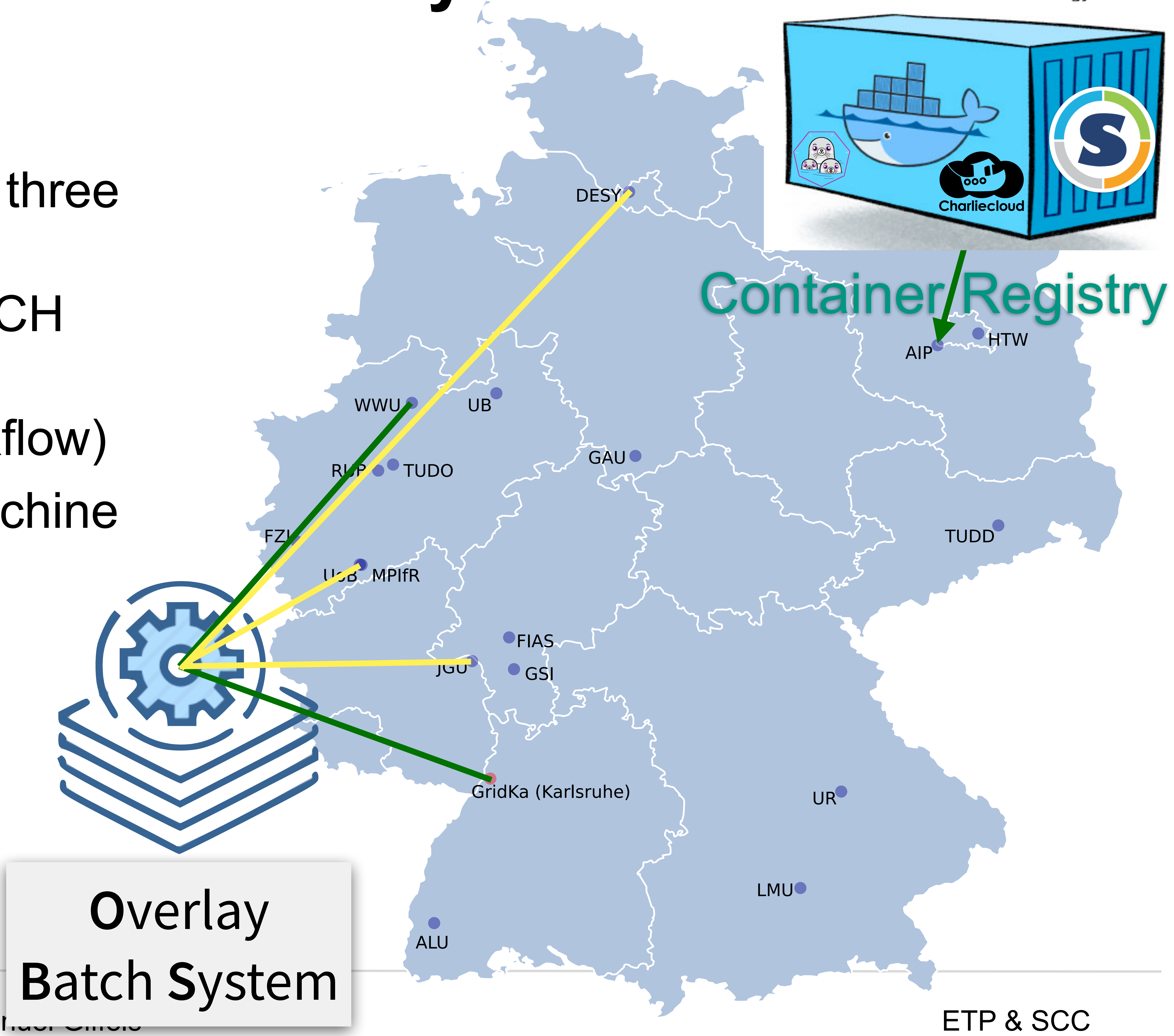
c4p-login.gridka.de



The Compute4PUNCH Infrastructure Today

- Prototype of federated Compute4PUNCH infrastructure is available
- Dynamic integration of two compute sites, three more will follow soon
- AAI based login node available to all PUNCH users
- Container registry available (+ CI/CD workflow)
- Container distributed via CERN Virtual Machine File System (CVMFS) for scaling

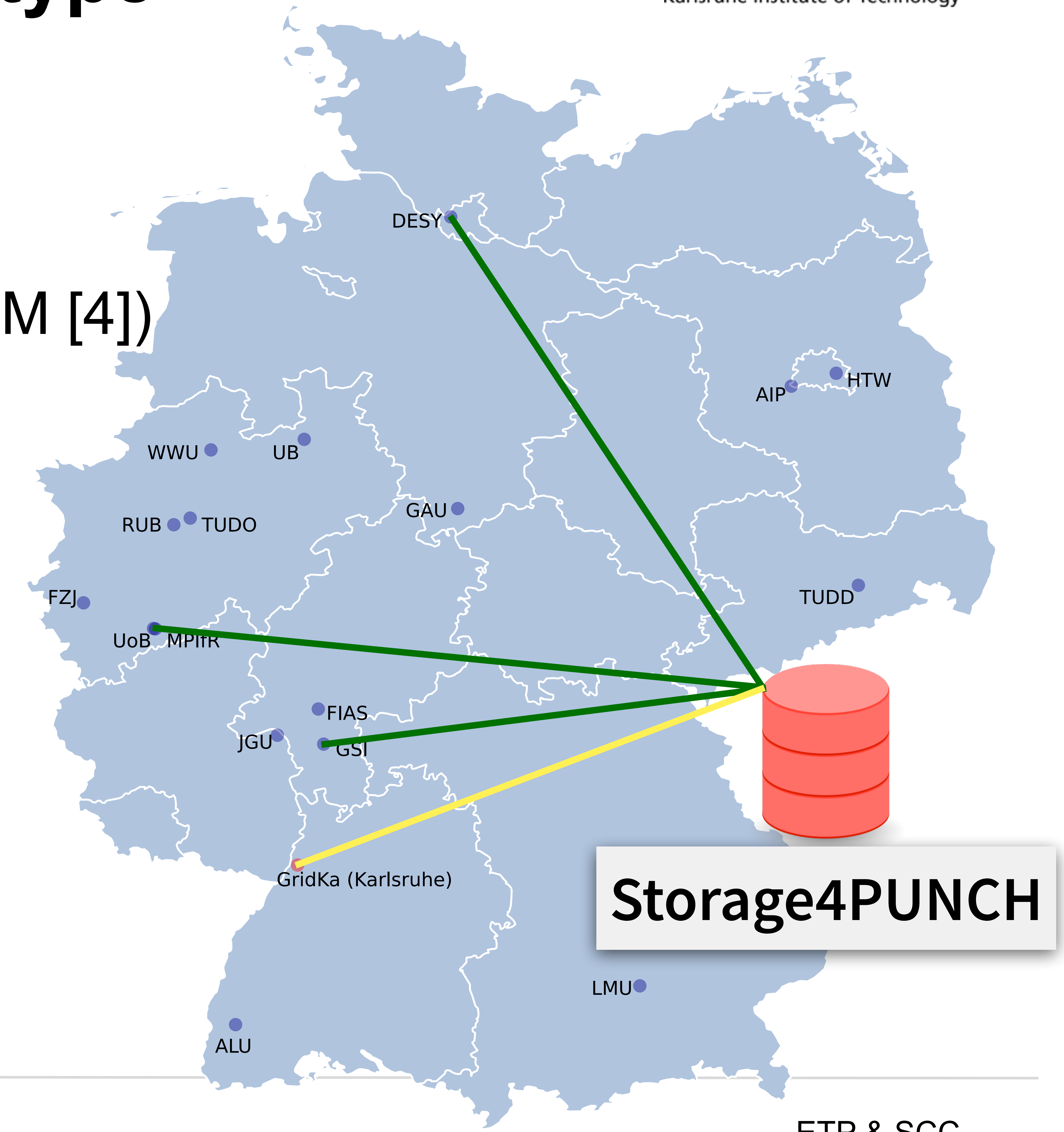
c4p-login.gridka.de



Overlay Batch System

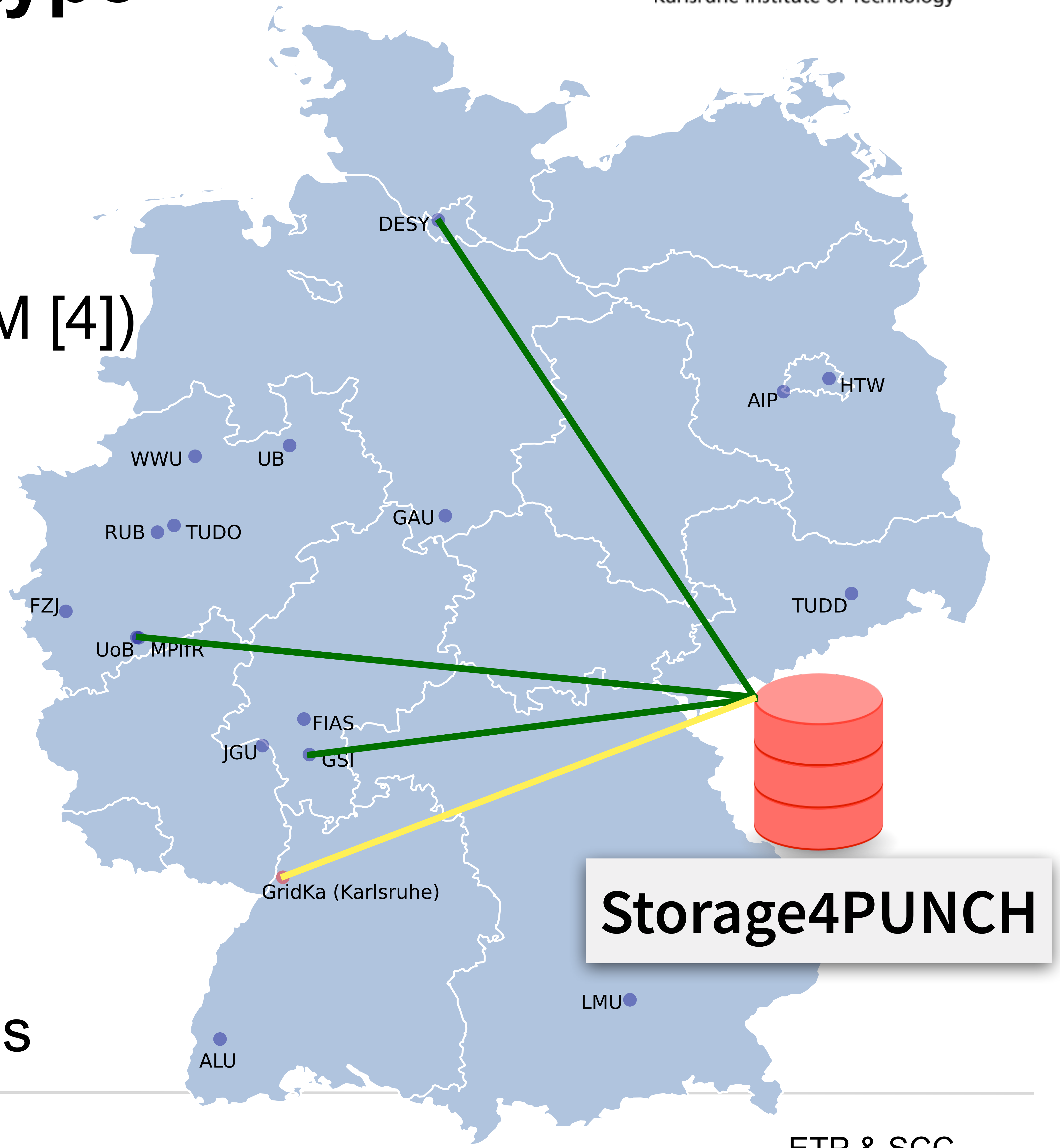
Storage4PUNCH - Distributed Prototype

- Based upon different technologies
  - dCache (Test Endpoint at DESY)
  - XRootD (Test Endpoints Bonn & GSI)
- Token based access using PUNCH AAI (Unity IdM [4])
- Using WebDav/XRootD as transfer protocol



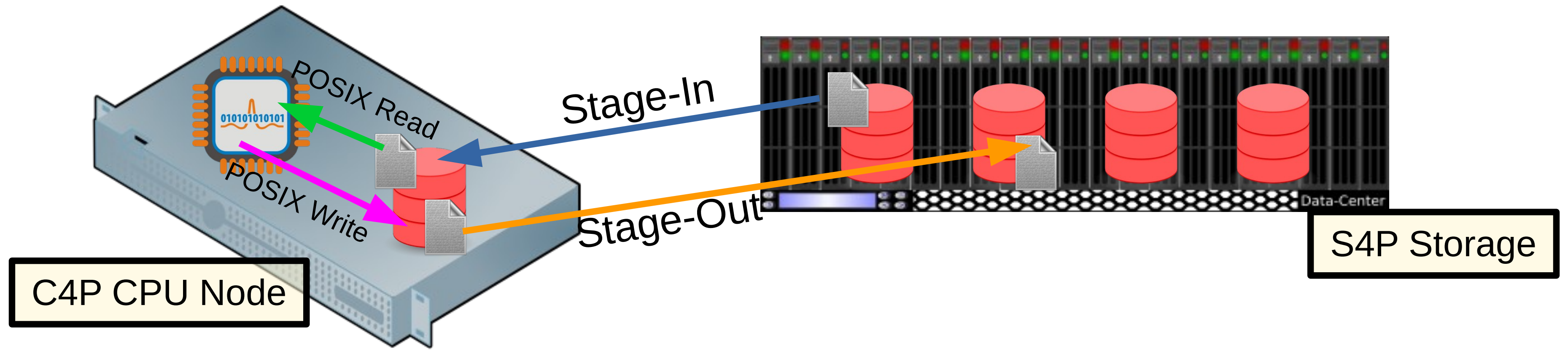
Storage4PUNCH - Distributed Prototype

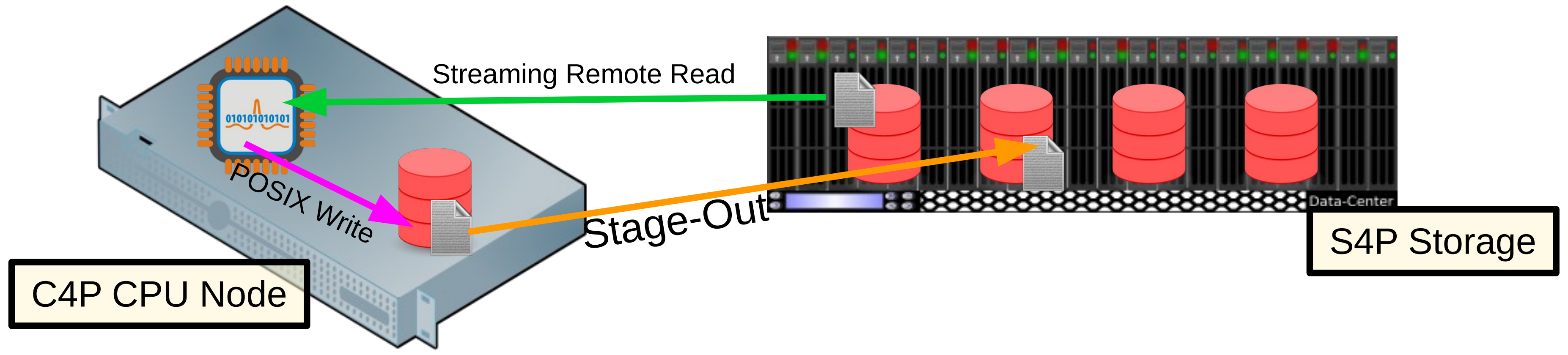
- Based upon different technologies
  - dCache (Test Endpoint at DESY)
  - XRootD (Test Endpoints Bonn & GSI)
- Token based access using PUNCH AAI (Unity IdM [4])
- Using WebDav/XRootD as transfer protocol
- **<u>To be integrated and evaluated:</u>**
  - File/Replica catalog candidates
    - RUCIO - Data management tool (HEP) [5]
    - LQCD - Meta data catalog (Lattice QCD) [6]
  - Federation options
    - XRootD federation
    - Hash table based data placement and access



Integration of Compute4PUNCH & Storage4PUNCH

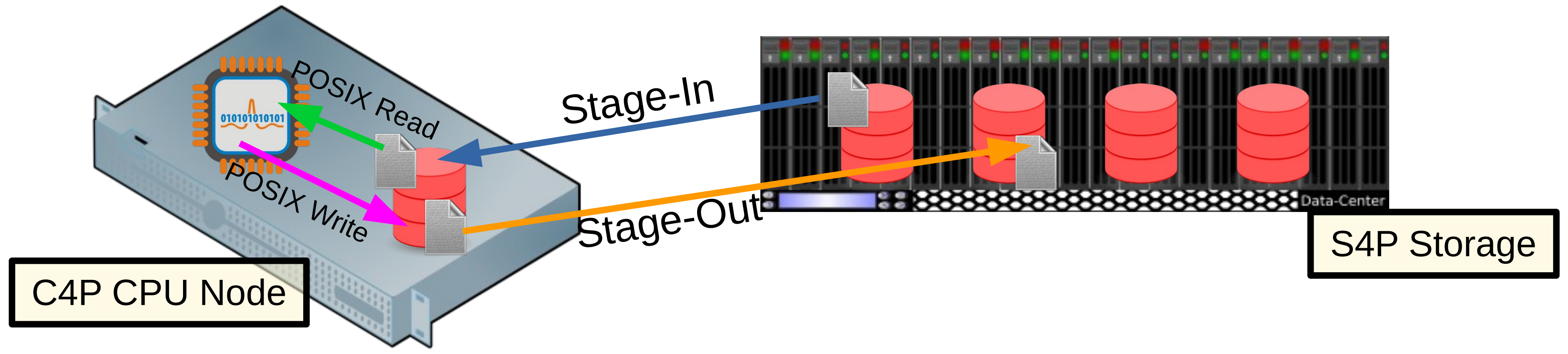
- Storage4PUNCH is not POSIX accessible
  - Files need to be staged to local POSIX compliant storage (usually inefficient)
  - Application needs to support streaming (preferred method, not always supported)

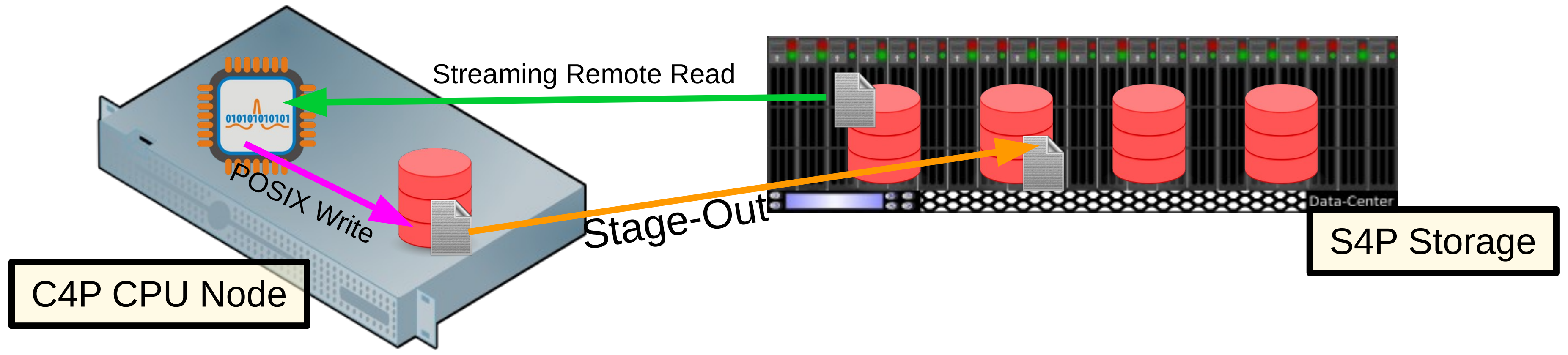




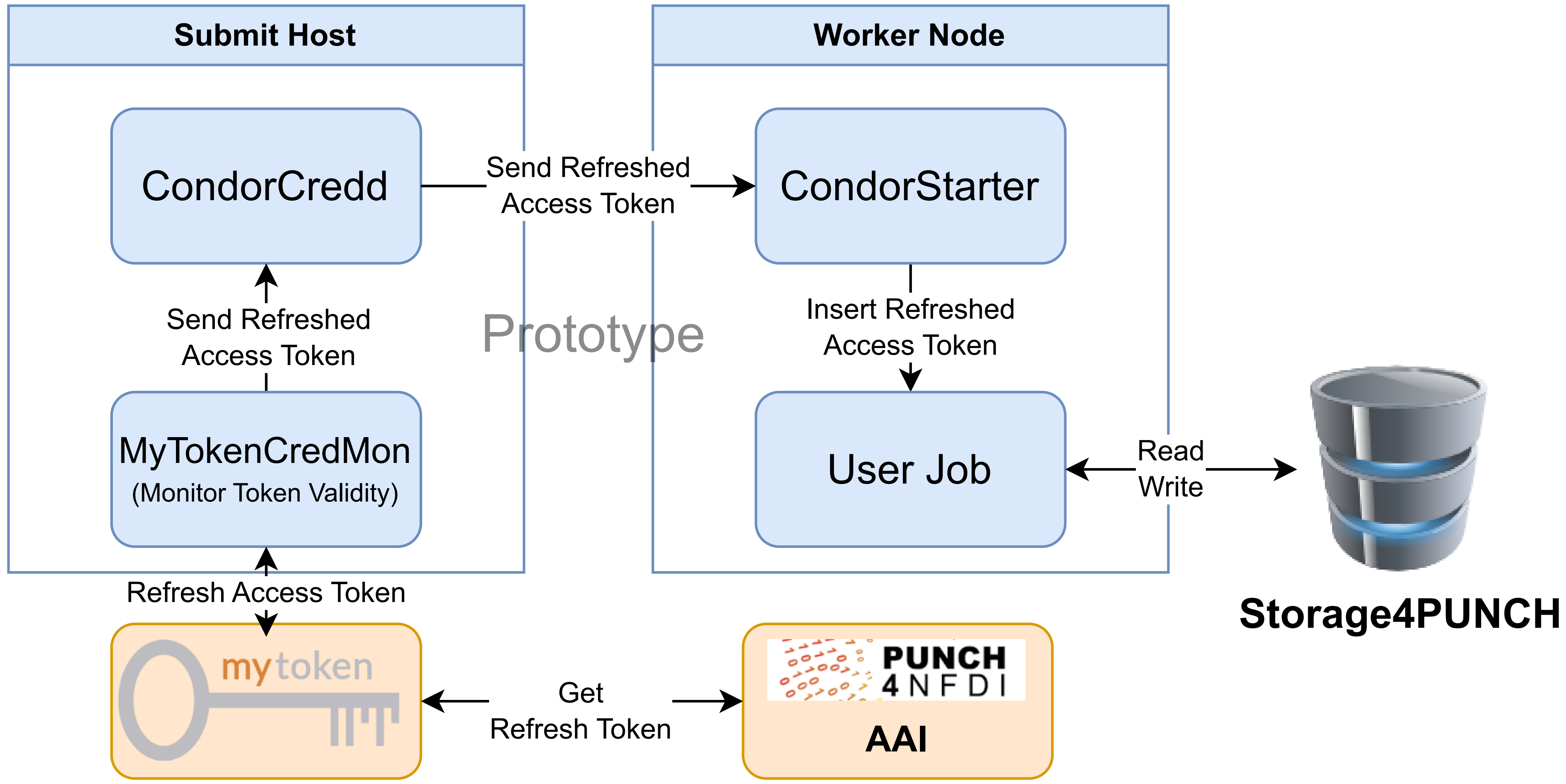
Integration of Compute4PUNCH & Storage4PUNCH

- Storage4PUNCH is not POSIX accessible
  - Files need to be staged to local POSIX compliant storage (usually inefficient)
  - Application needs to support streaming (preferred method, not always supported)





- Access token has a limited lifetime, usually shorter than the job runtime
  - Add refresh token to MyToken service [7]
  - Use HTCondor CredMon to create, monitor and refresh access token of the user
  - Use HTCondor Credd to synchronize access token to user jobs on worker nodes



Workflows on Compute4PUNCH & Storage4PUNCH

LOFAR Radio imaging workflow

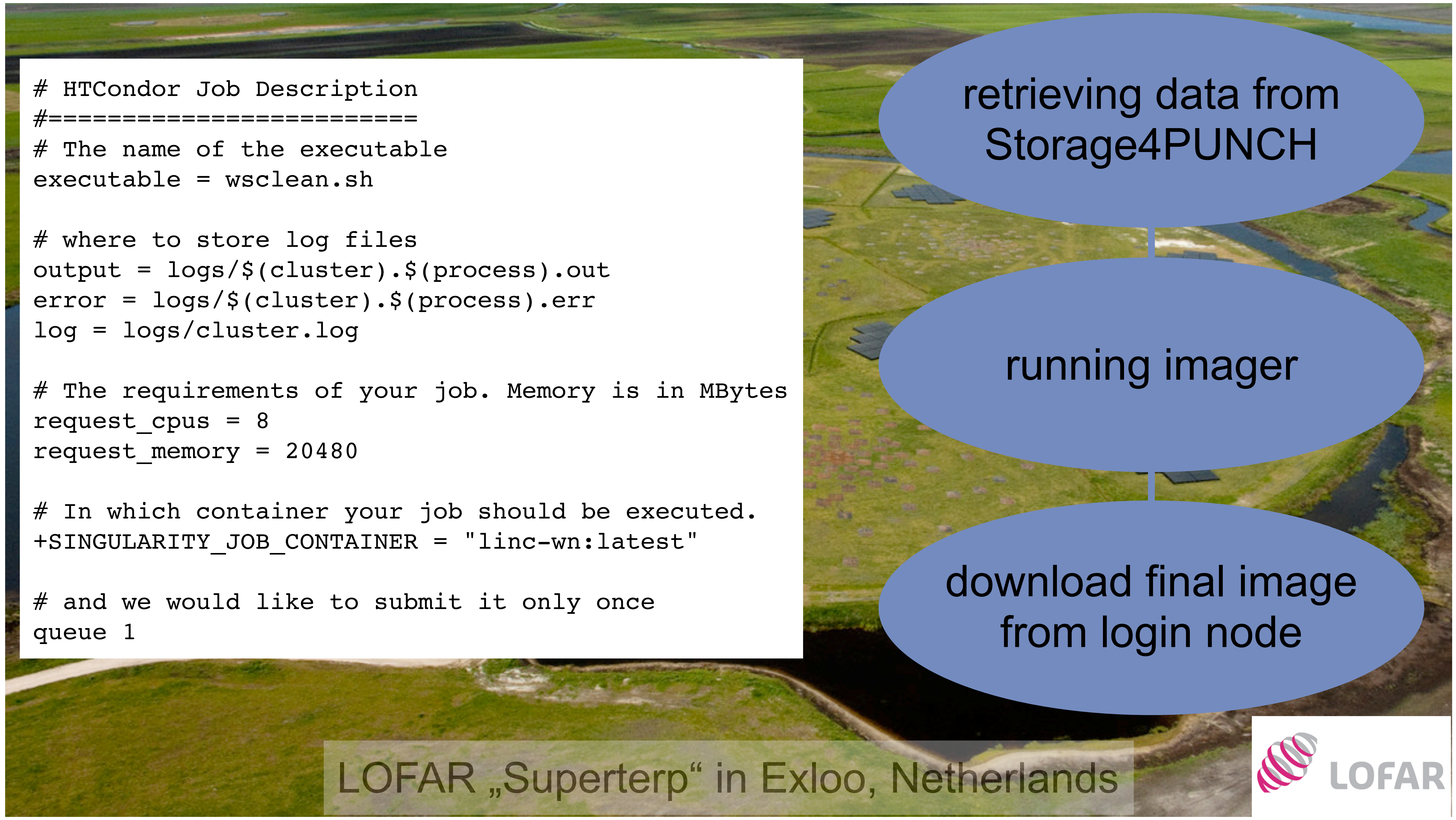
- **L**ow **F**requency **Ar**ray (LOFAR)

LOFAR „Superterp“ in Exloo, Netherlands

Workflows on Compute4PUNCH & Storage4PUNCH

LOFAR Radio imaging workflow

- **Lo**w **F**requency **Ar**ray (LOFAR)
- Reconstruction of the sky brightness distribution from recorded interferometry data
- Software provided via apptainer container
- Data is available on Storage4PUNCH (~150 GB)



```
# HTCondor Job Description
#=========================
# The name of the executable
executable = wsclean.sh

# where to store log files
output = logs/$(cluster).$(process).out
error = logs/$(cluster).$(process).err
log = logs/cluster.log

# The requirements of your job. Memory is in MBytes
request_cpus = 8
request_memory = 20480

# In which container your job should be executed.
+SINGULARITY_JOB_CONTAINER = "linc-wn:latest"

# and we would like to submit it only once
queue 1
```

retrieving data from Storage4PUNCH

running imager

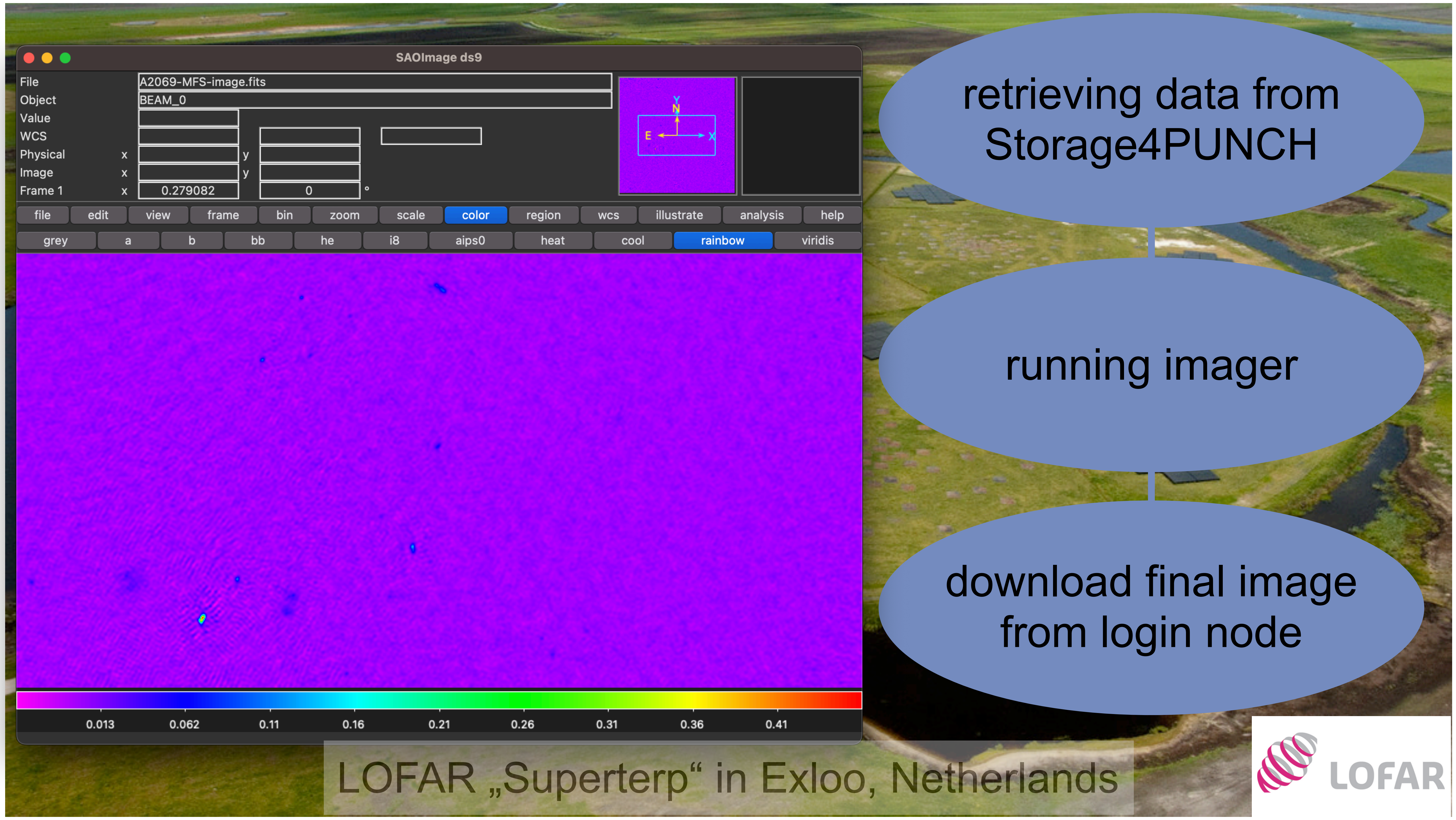
download final image from login node

LOFAR „Superterp“ in Exloo, Netherlands

Workflows on Compute4PUNCH & Storage4PUNCH

LOFAR Radio imaging workflow

- **Lo**w **F**requency **Ar**ray (LOFAR)
- Reconstruction of the sky brightness distribution from recorded interferometry data
- Software provided via apptainer container
- Data is available on Storage4PUNCH (~150 GB)



SAOImage ds9

File: A2069-MFS-image.fits
Object: BEAM_0
Frame 1 x: 0.279082 0 °

Value
WCS
Physical x y
Image x y
file edit view frame bin zoom scale color region wcs illustrate analysis help
grey a b bb he i8 aips0 heat cool rainbow viridis
0.013 0.062 0.11 0.16 0.21 0.26 0.31 0.36 0.41
N
E
X
Y
LOFAR

retrieving data from Storage4PUNCH

running imager

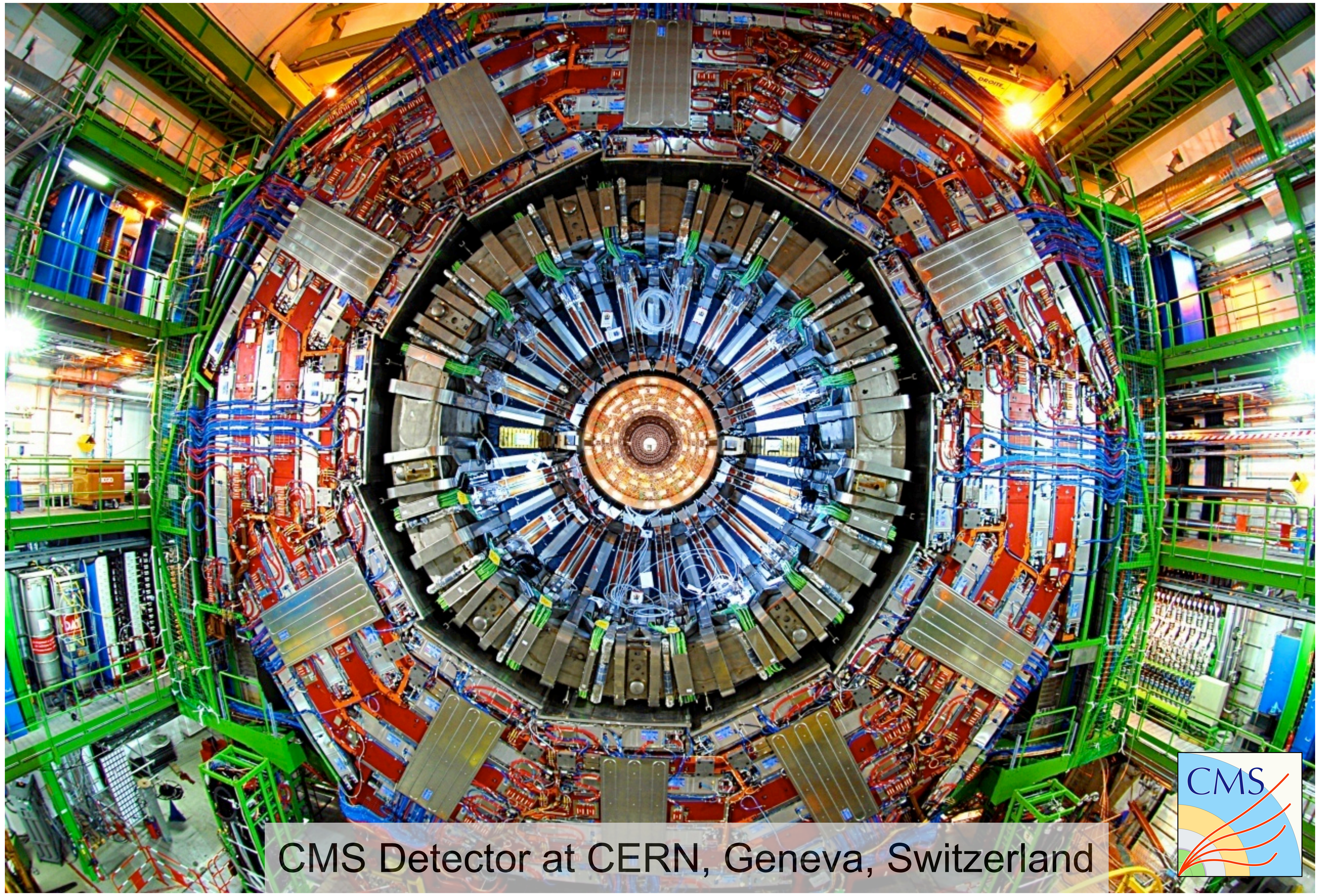
download final image from login node

LOFAR „Superterp“ in Exloo, Netherlands

Workflows on Compute4PUNCH & Storage4PUNCH

CERN Open Data Workflow

- Data taken at the **C**ompact **M**uon **S**olenoid (CMS) detector at LHC, CERN

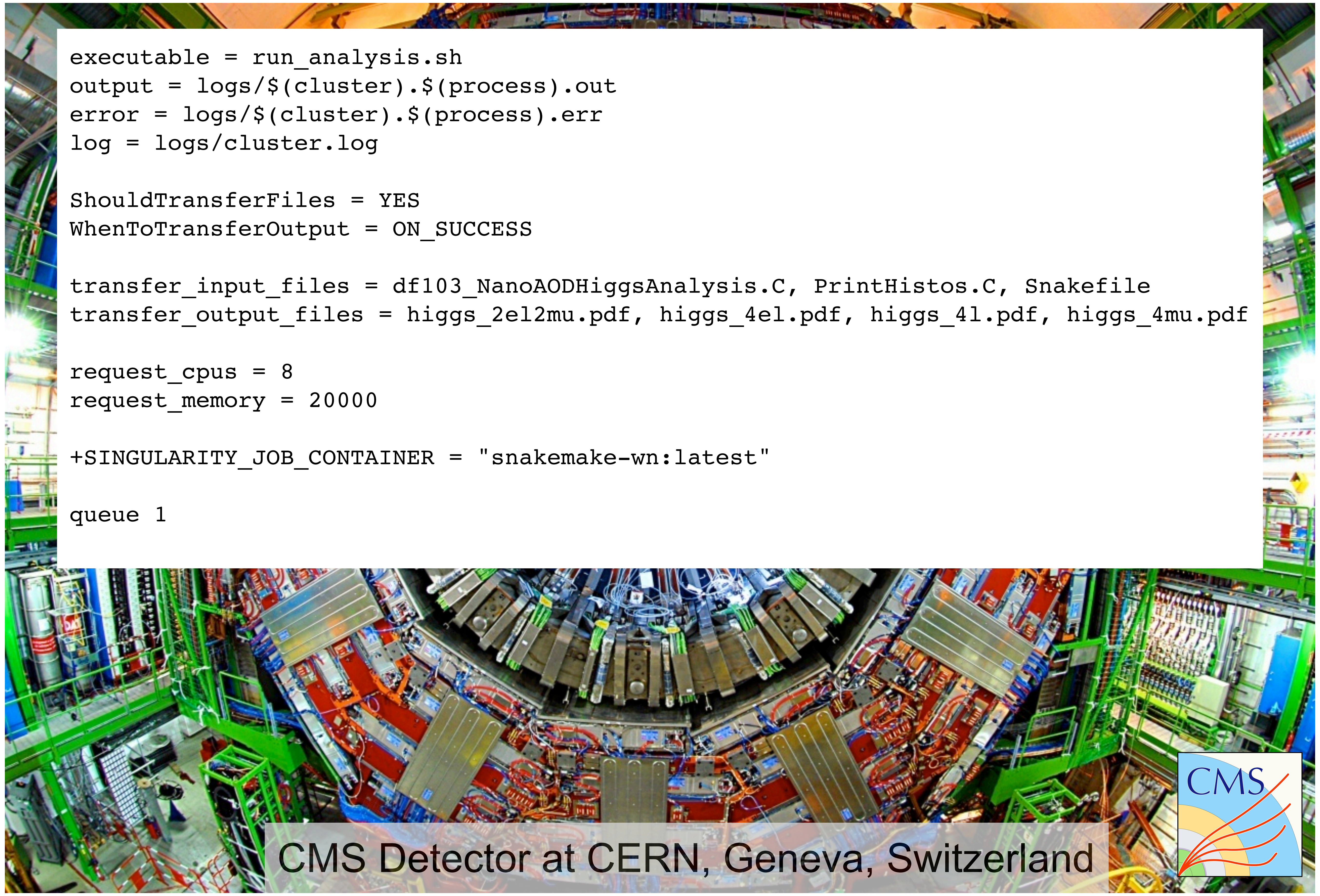


CMS Detector at CERN, Geneva, Switzerland

Workflows on Compute4PUNCH & Storage4PUNCH

CERN Open Data Workflow

- Data taken at the **C**ompact **M**uon **S**olenoid (CMS) detector at LHC, CERN
- Perform simplified Higgs analysis using data taken back in 2012
- Software from CVMFS
- Data directly streamed from EOS Filesystem at CERN (~13 GB)



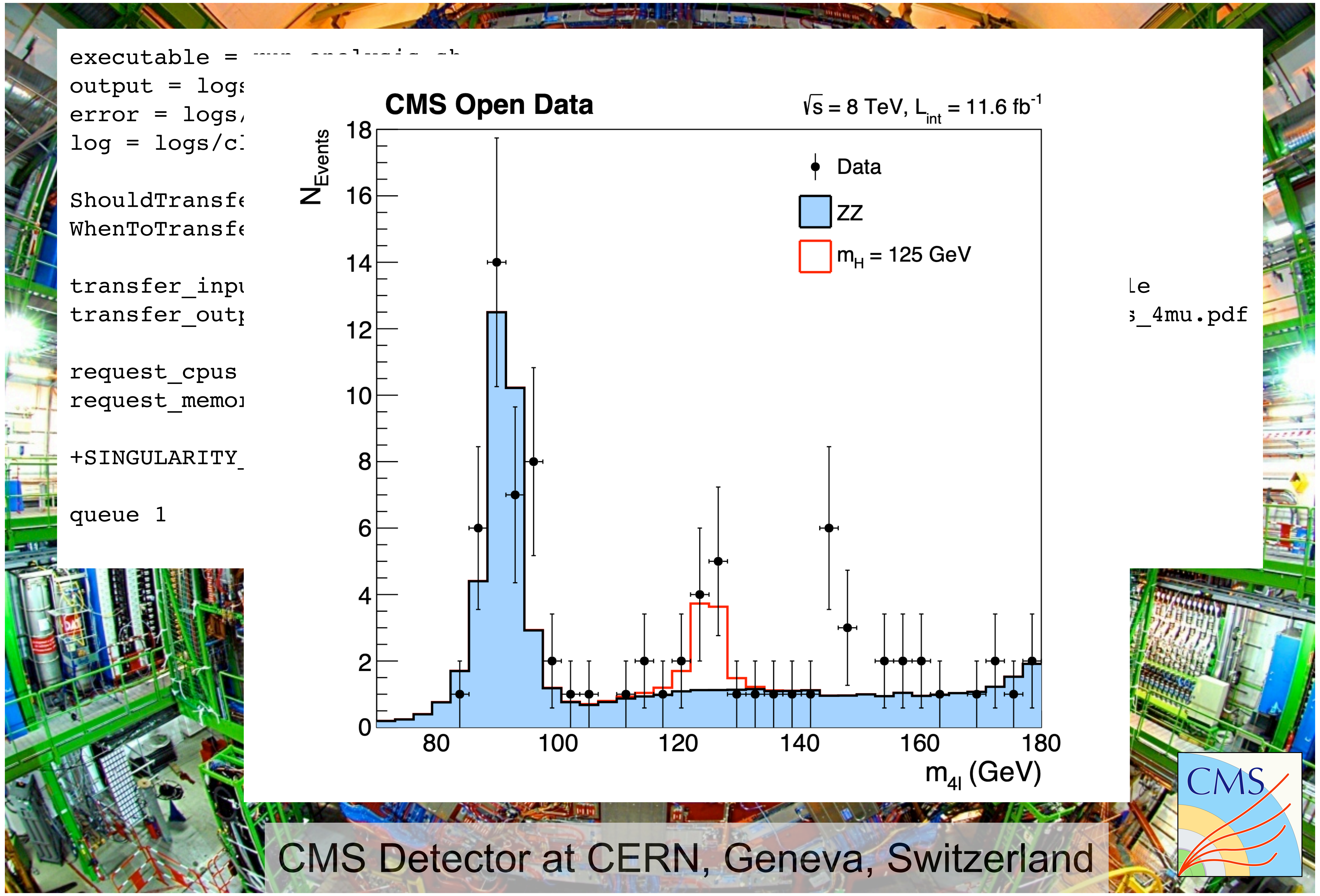
```
executable = run_analysis.sh
output = logs/$(cluster).$(process).out
error = logs/$(cluster).$(process).err
log = logs/cluster.log

ShouldTransferFiles = YES
WhenToTransferOutput = ON_SUCCESS

transfer_input_files = df103_NanoAODHiggsAnalysis.C, PrintHistos.C, Snakefile
transfer_output_files = higgs_2el2mu.pdf, higgs_4el.pdf, higgs_4l.pdf, higgs_4mu.pdf

request_cpus = 8
request_memory = 20000

+SINGULARITY_JOB_CONTAINER = "snakemake-wn:latest"

queue 1
```

CMS Detector at CERN, Geneva, Switzerland

Workflows on Compute4PUNCH & Storage4PUNCH

CERN Open Data Workflow

- Data taken at the **C**ompact **M**uon **S**olenoid (CMS) detector at LHC, CERN
- Perform simplified Higgs analysis using data taken back in 2012
- Software from CVMFS
- Data directly streamed from EOS Filesystem at CERN (~13 GB)



```
executable =
output = logs
error = logs/
log = logs/cl

ShouldTransfe
WhenToTransfe

transfer_inpu                                                   le
transfer_outp                                                   s_4mu.pdf

request_cpus
request_memor

+SINGULARITY_

queue 1
```

CMS Detector at CERN, Geneva, Switzerland

Conclusions

- Demonstrator of the federated Compute4PUNCH infrastructure is available
- Demonstrator of a distributed Storage4PUNCH infrastructure based on dCache and XRootD is available (no federation yet)
- Access to compute and storage resource possible utilizing access tokens provided by PUNCH AAI (based on Unity IdM)
- Demonstrator of the automated access token refresh workflow based on MyToken, HTCondor Credd & CredMon available
- Several demonstration workflows of different communities available
  - Astronomy Workflows: e.g. LOFAR Radio imaging workflow
  - HEP Workflows: e.g. CERN Open Data Workflow ($H \rightarrow ZZ \rightarrow 4l$)

Ongoing development project. More will follow soon …

References

[1] Max Fischer, Eileen Kuehn, Manuel Giffels, Matthias Schnepf, Stefan Kroboth, Thorsten M., & Oliver Freyermuth. (2022). MatterMiners/cobald: v0.13.0 (0.13.0). Zenodo. https://doi.org/10.5281/zenodo.7032186

[2] Manuel Giffels, Max Fischer, Alexander Haas, Stefan Kroboth, Matthias Schnepf, Eileen Kuehn, PSchuhmacher, Rene Caspart, Florian von Cube & Peter Wienemann. (2023). MatterMiners/tardis: The Escape (0.7.0). Zenodo. https://doi.org/10.5281/zenodo.7680164

[3] CVMFS, https://cernvm.cern.ch/portal/filesystem, accessed on 2023-05-04

[4] UnitIdm, https://unity-idm.eu/, accessed on 2023-05-04

[5] Martin Barisits, et al, Rucio: Scientific Data Management, Computing and Software for Big Science (2019) 3:11 https://doi.org/10.1007/s41781-019-0026-3

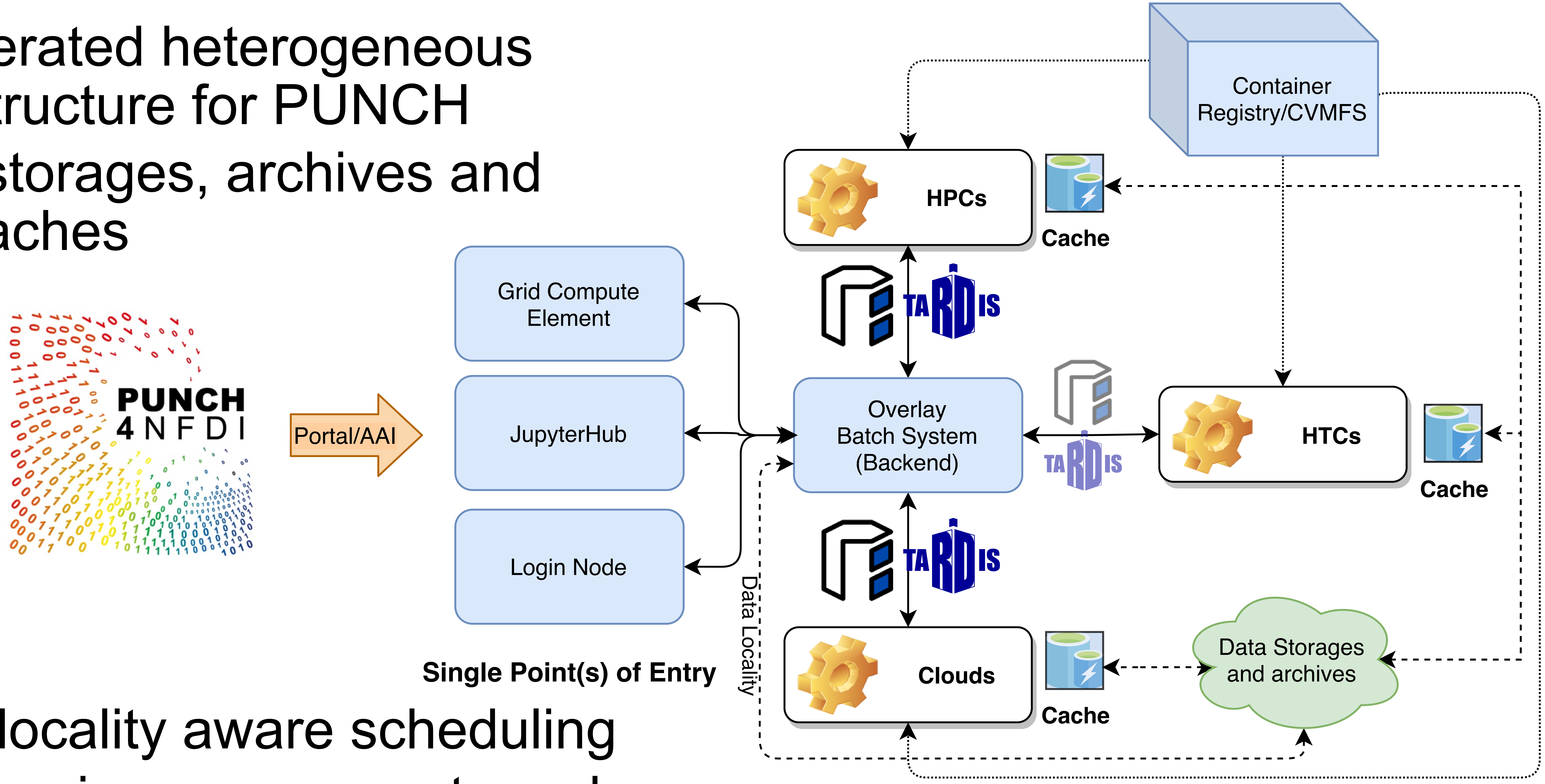
[6] Mark G. Beckett, et al, Building the International Lattice Data Grid, Comput. Phys. Commun, 182:1208-1214, 2011

[7] MyToken, https://mytoken-docs.data.kit.edu/, accessed on 2023-05-04

Backup

Towards the Compute4PUNCH Infrastructure

- Establish a federated heterogeneous compute infrastructure for PUNCH
- Integrate data storages, archives and opportunistic caches



- Introduce data-locality aware scheduling
- Benefit from experiences, concepts and tools available in HEP community