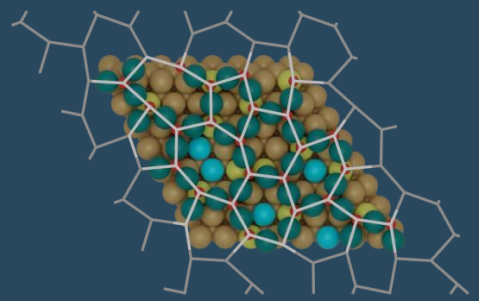
FRITZ-HABER-INSTITUT
MAX-PLANCK-GESELLSCHAFT

WORKSTATIONS, SLURM CLUSTER, ANALYSIS SERVER, USER STORAGE, NETWORK INTEGRATION

USE CASES AND CHALLENGES OF THE FHI



Maurits W. Vuijk[1,*], C. Scheurer[1], H. Junkes[2], Simeon D. Beinlich[1,2,**]

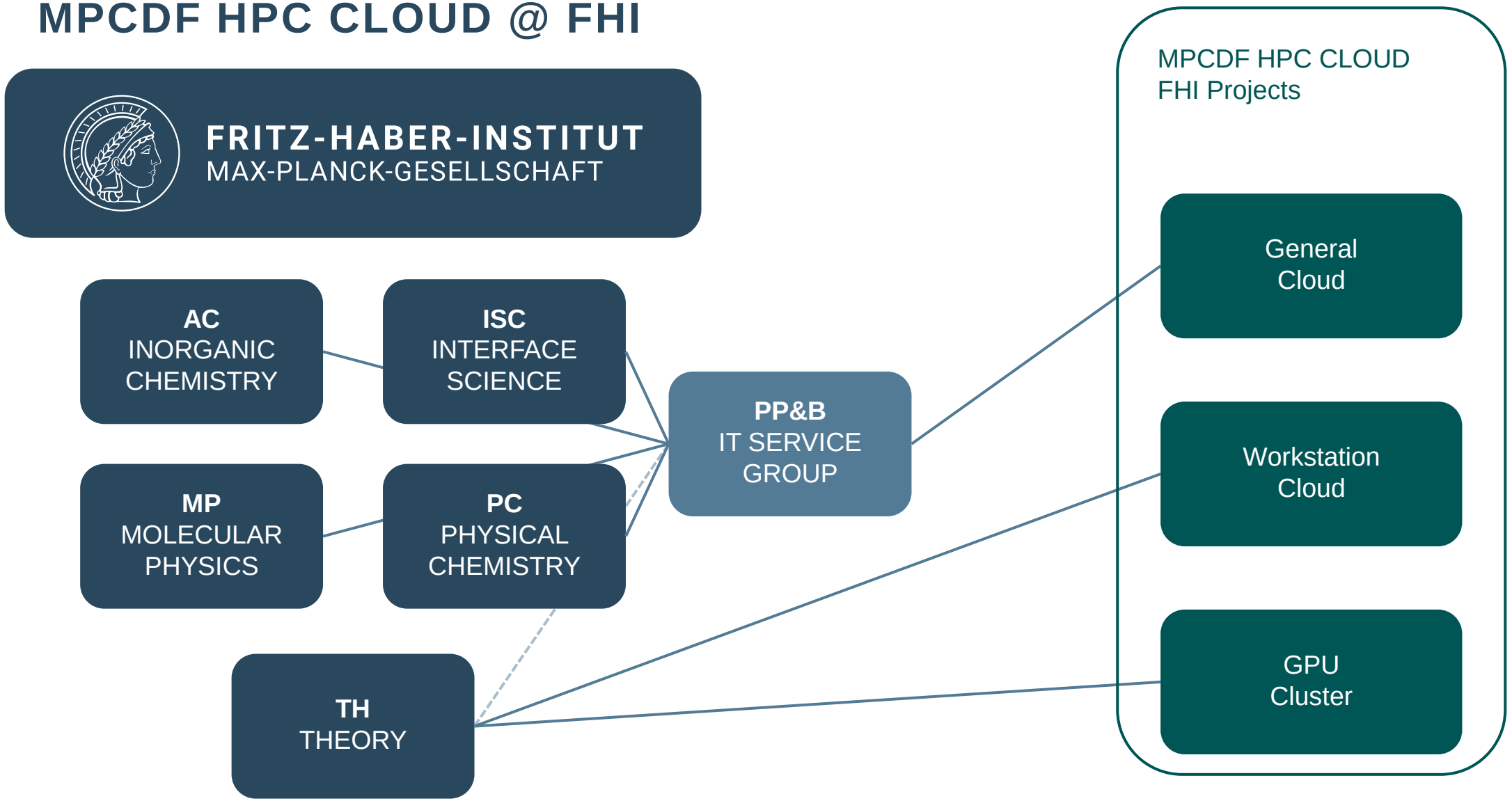
1 Theory Department
Fritz-Haber-Institut der Max-Planck-Gesellschaft, Berlin, Germany

2 PP&B – Computer Support Group
Fritz-Haber-Institut der Max-Planck-Gesellschaft, Berlin, Germany

*vuijk@fhi.mpg.de

**beinlich@fhi.mpg.de

MPCDF HPC CLOUD @ FHI



MPCDF HPC CLOUD @ FHI

FRITZ-HABER-INSTITUT
MAX-PLANCK-GESELLSCHAFT

AC
INORGANIC CHEMISTRY

ISC
INTERFACE SCIENCE

PP&B
IT SERVICE GROUP

MP
MOLECULAR PHYSICS

PHYSICAL CHEMISTRY

All Compute Resources at MPCDF

TH
THEORY

MPCDF HPC CLOUD
FHI Projects

General Cloud

Virtual Workstations

Workstation Cloud

Everyday Computing

GPU Cluster

Large GPU Nodes

MPCDF HPC CLOUD @ FHI

FRITZ-HABER-INSTITUT
MAX-PLANCK-GESELLSCHAFT

AC
INORGANIC CHEMISTRY

ISC
INTERFACE SCIENCE

PP&B
IT SERVICE GROUP

MP
MOLECULAR PHYSICS

PHYSICAL CHEMISTRY

All Compute Resources at MPCDF

TH
THEORY

MPCDF HPC CLOUD
FHI Projects

General Cloud

Virtual Workstations

Workstation Cloud

Everyday Computing

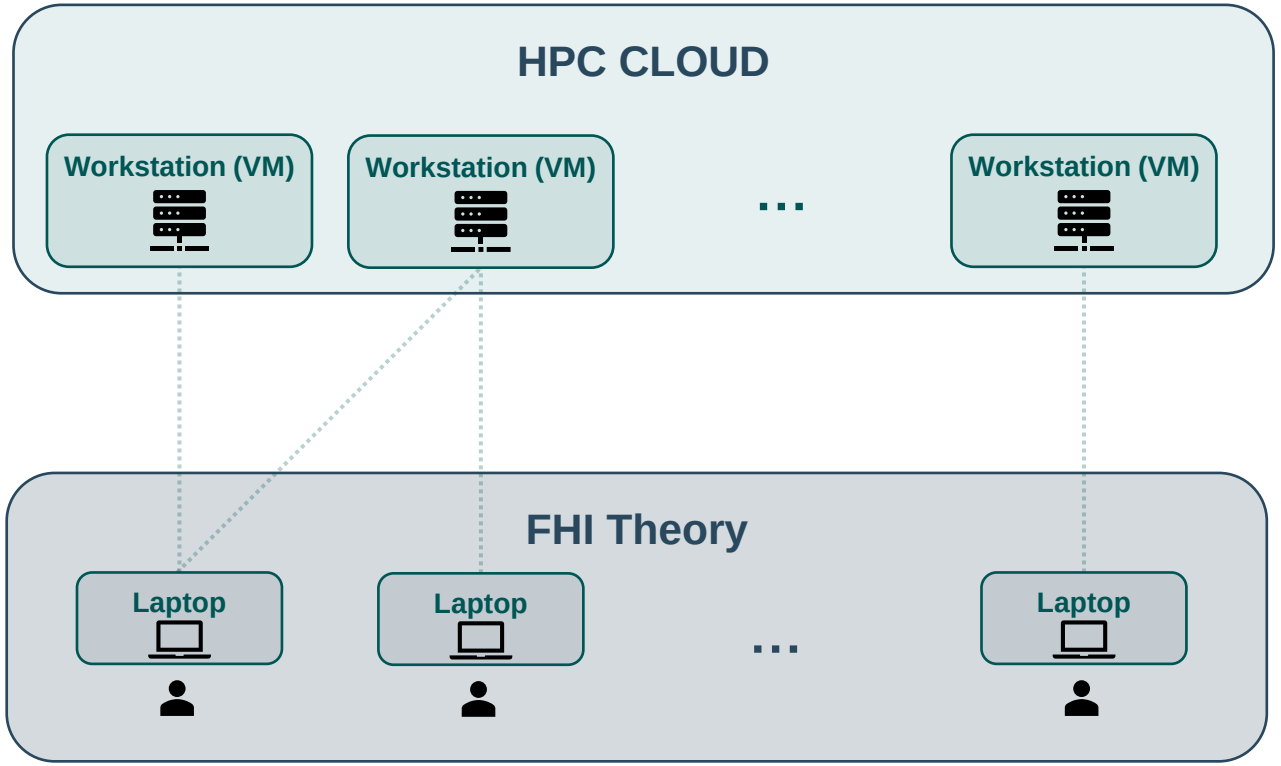
GPU Cluster

Large GPU Nodes

TH – PERSONAL AND PROJECT WORKSTATIONS

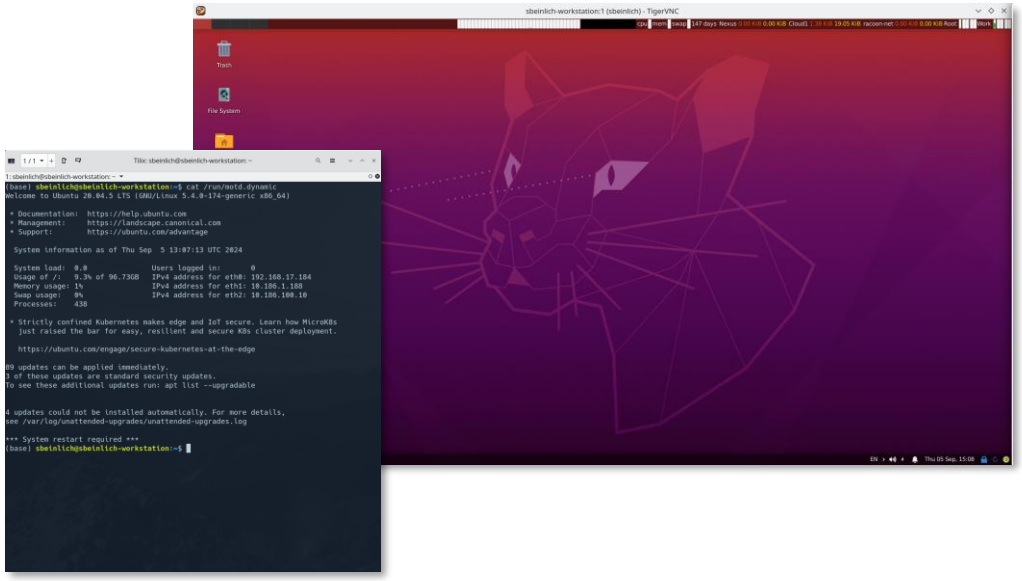
Use cases

- Virtual remote workstations
- *Everyday computing* – analysis, programming, simulations



Resources (used)

- 118 VMs
- 3400 cores – 14.9TB RAM
- 196 volumes – 150TB storage
- Nexus share
  (mounted at Raven & Cobra)



TH – PERSONAL AND PROJECT WORKSTATIONS

Use cases

- Virtual remote workstations
- *Everyday computing* – analysis, programming, simulations

Current state

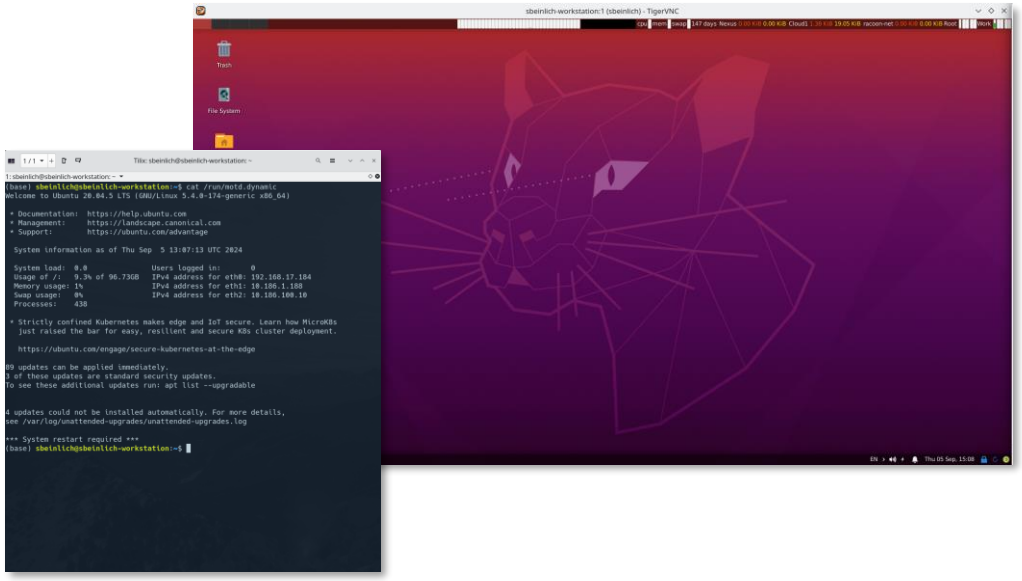
- Very mature (> 3 years in heavy use)
- Stable and low effort

Future plans

- Private subnet
- Additional Binder Hub
- Nexus often full (Block Size)
  → Manila NFS Share

Resources (used)

- 118 VMs
- 3400 cores – 14.9TB RAM
- 196 volumes – 150TB storage
- Nexus share
  (mounted at Raven & Cobra)



TH – PERSONAL AND PROJECT WORKSTATIONS

Use cases

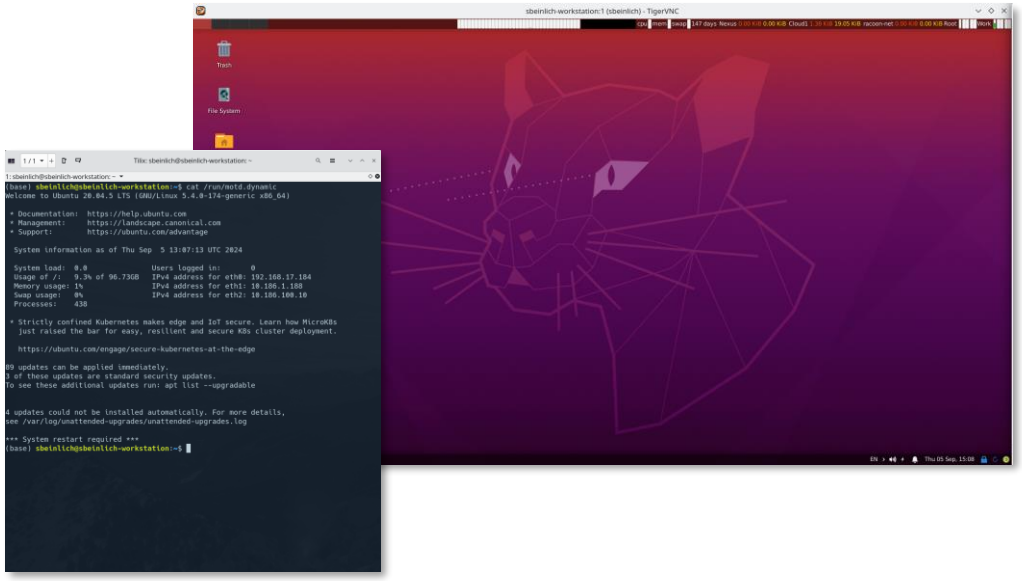
- Virtual remote workstations
- *Everyday computing* – analysis, programming, simulations

Implementation

- Personal & project VMs (36c/120g, user-administrated, root access)
- MPCDF user accounts
- Ubuntu 20.04 + XFCE desktop
- SSH & VNC access (via gate1.mpcdf.mpg.de)
- 1-5TB volumes (BTRFS snapshotted & compressed)

Resources (used)

- 118 VMs
- 3400 cores – 14.9TB RAM
- 196 volumes – 150TB storage
- Nexus share
  (mounted at Raven & Cobra)



TH – PERSONAL AND PROJECT WORKSTATIONS

Use cases

- Virtual remote workstations
- *Everyday computing* – analysis, programming, simulations

Implementation

- Personal & project VMs (36c/120g, user-administrated, root access)
- MPCDF user accounts
- Ubuntu 20.04 + XFCE desktop
- SSH & VNC access (via gate1.mpcdf.mpg.de)
- 1-5TB volumes (BTRFS snapshotted & compressed)

Advantages

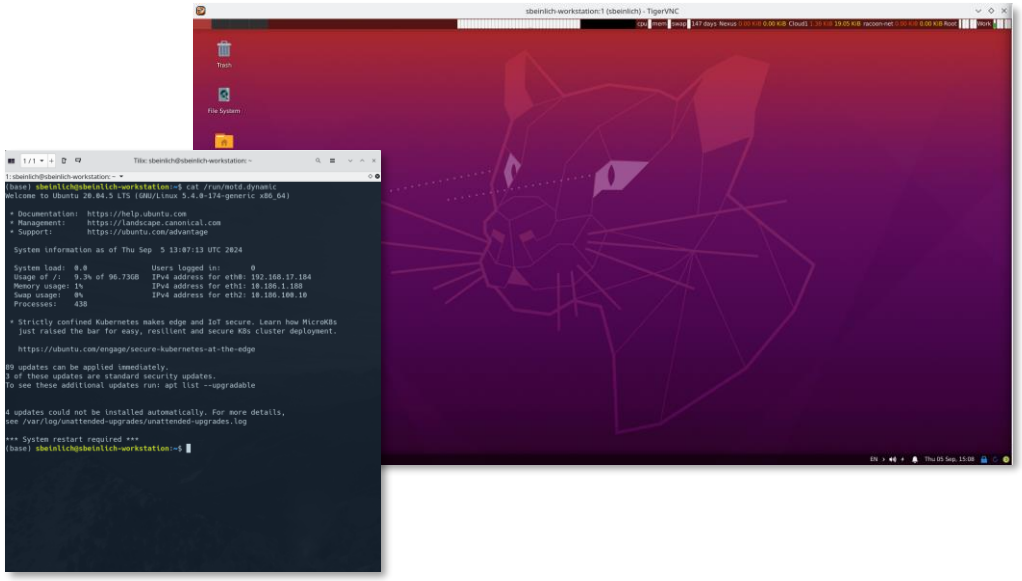
- Low maintenance & uniform setup
- High performance

Disadvantages

- (Some) administrative effort

Resources (used)

- 118 VMs
- 3400 cores – 14.9TB RAM
- 196 volumes – 150TB storage
- Nexus share
  (mounted at Raven & Cobra)

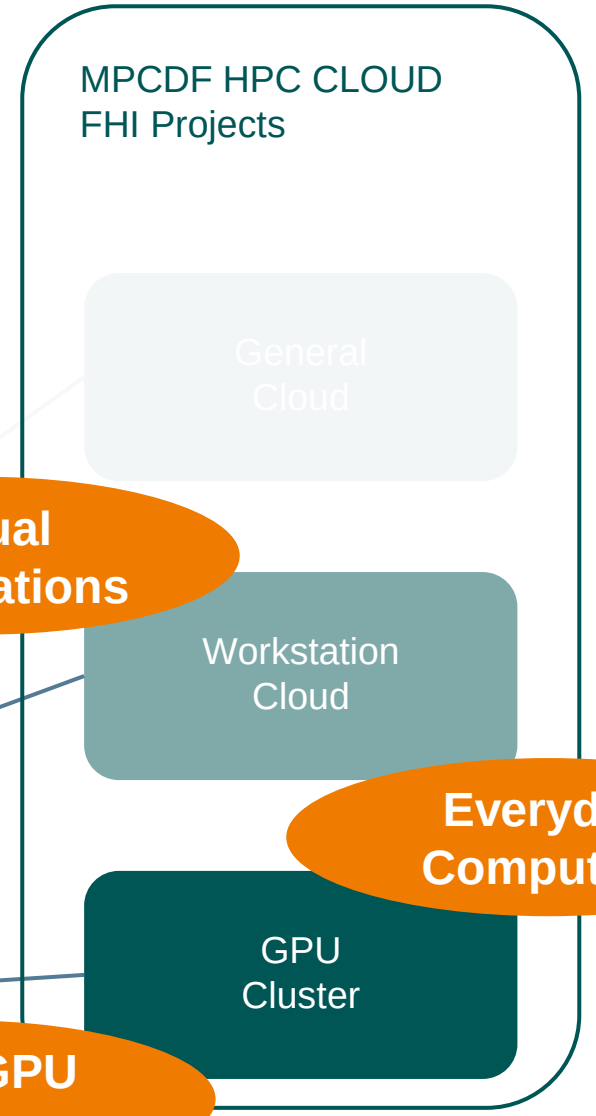


MPCDF HPC CLOUD @ FHI

FRITZ-HABER-INSTITUT
MAX-PLANCK-GESELLSCHAFT

All Compute Resources at MPCDF

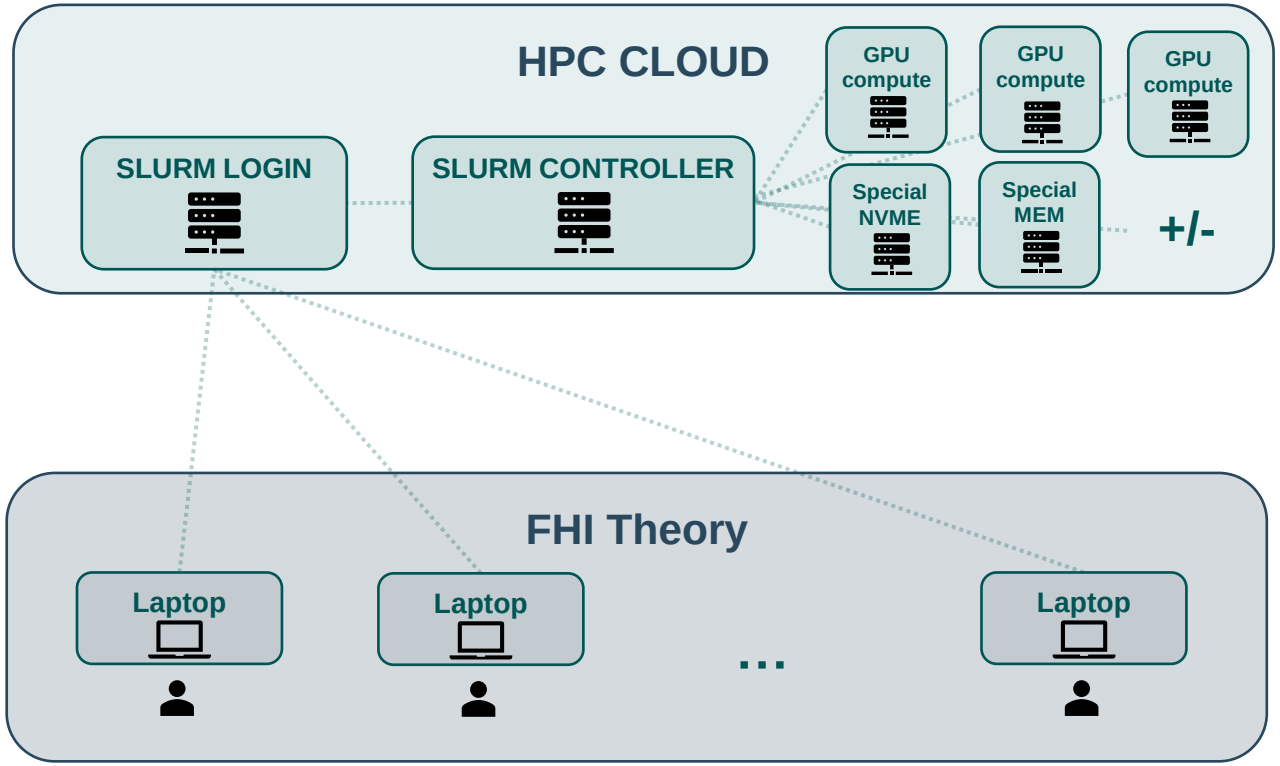
TH
THEORY



TH – GPU CLUSTER RACCOON

Use cases

- *GPU computing* – AI, ML, rendering, …
- *Specialty computing – if Raven doesn't fit*



Resources (total)

- 1408 cores – 56TB RAM
- 44 GPUs (NVIDIA A100)
- 140TB (local) NVME storage
- Manila NFS
- Nexus share
  (mounted at Raven & Cobra)

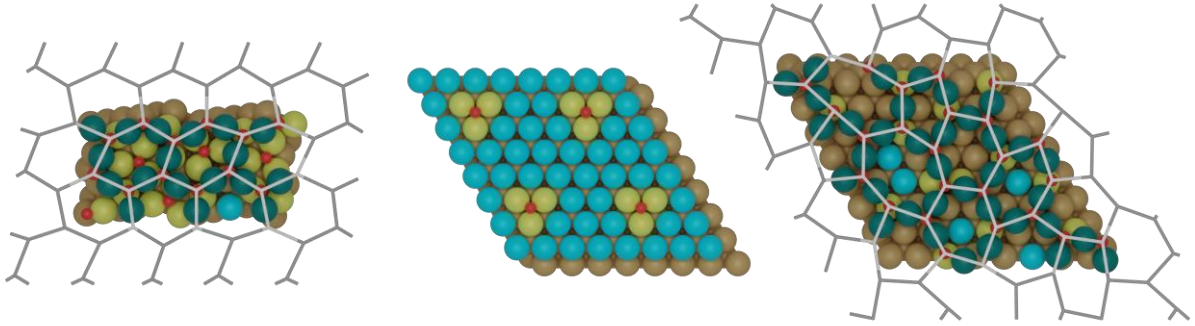
TH – GPU CLUSTER RACCOON

Use cases

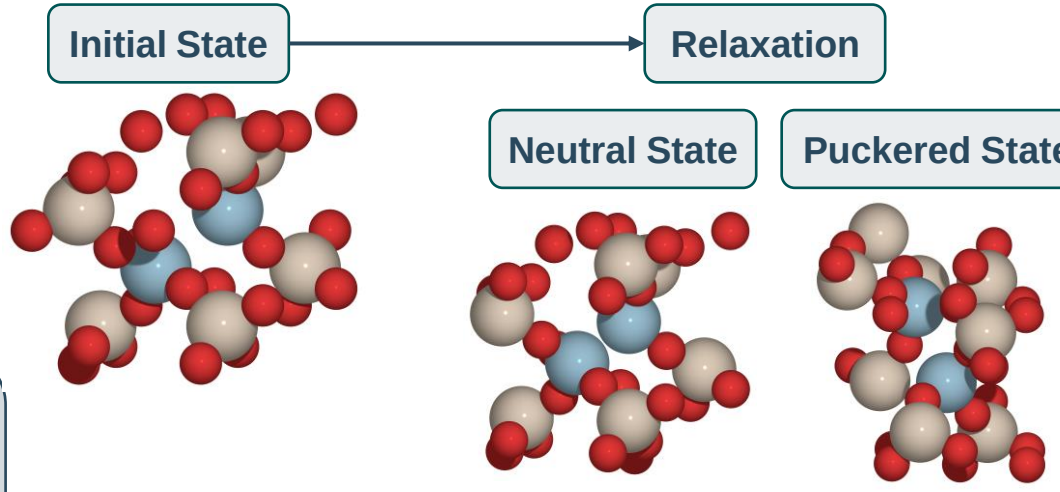
- *GPU computing* – AI, ML, rendering, …
- *Specialty computing – if Raven doesn't fit*

FHI Theory

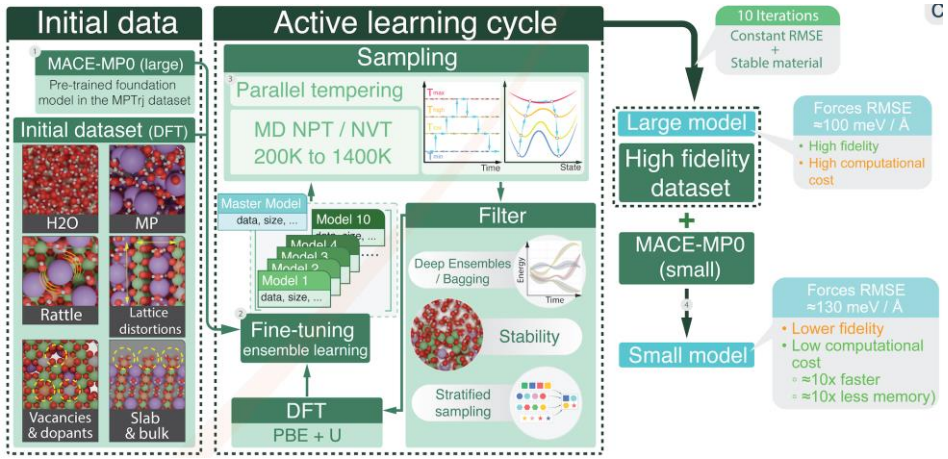
- Machine Learned Interatomic Potentials
- Large Language Models
- Computer Vision
- Generative Adversarial Networks



LOCAL CHEMICAL ENVIRONMENTS – F. RICCIUS

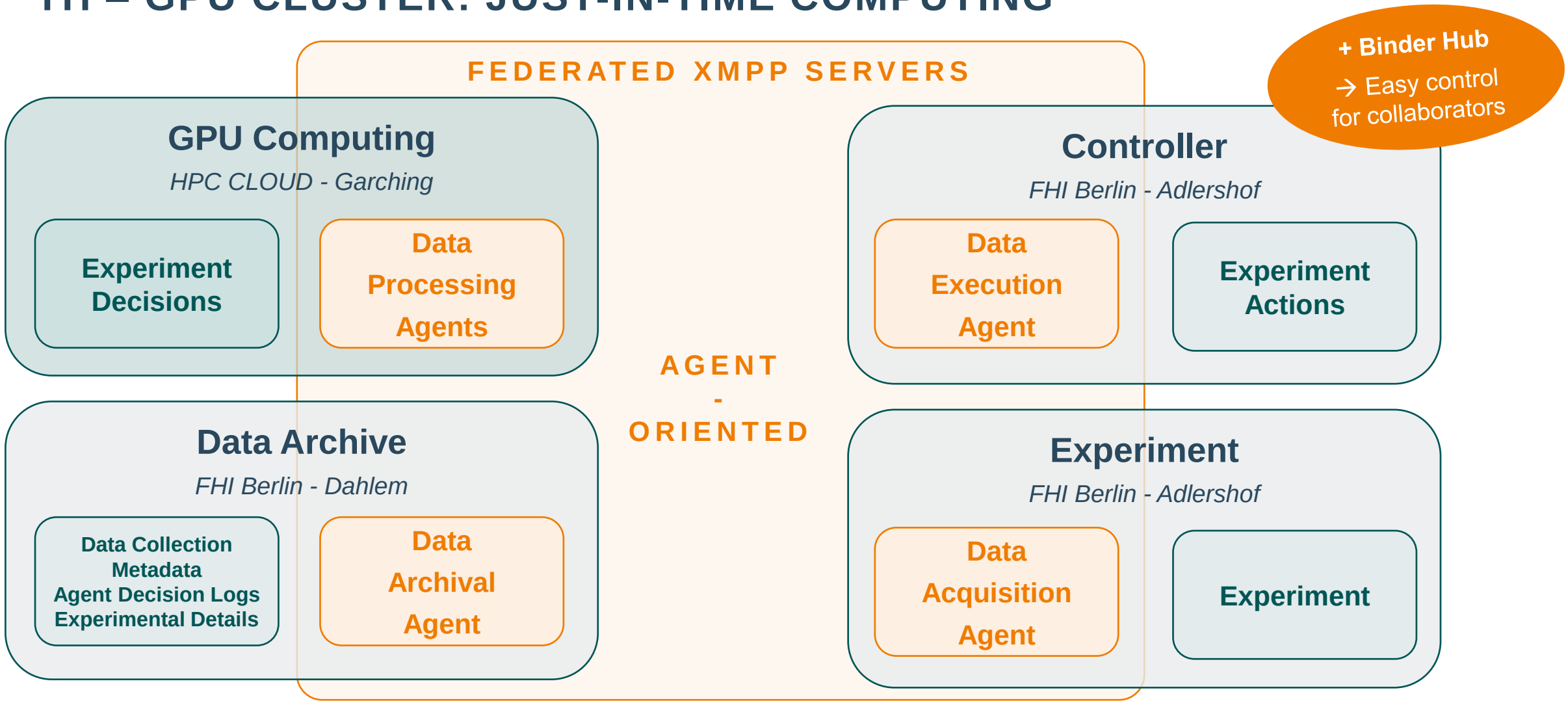


MACE WITH CHARGES – M. VONDRÁK



MACE ACTIVE LEARNING – J.M. LOMBARDI

TH – GPU CLUSTER: JUST-IN-TIME COMPUTING



TH – GPU CLUSTER RACCOON

Use cases

- *GPU computing* – AI, ML, rendering, …
- *Specialty Computing – if Raven doesn't fit*

Goal

- Raven-like compute cluster (SLURM, module system, Nexus)
- Dynamically scalable
- Flexible (any use case):
  - Ultra long walltime, shared-node jobs, …
  - A100 GPUs, huge memory, local NVMEs, …

Current state

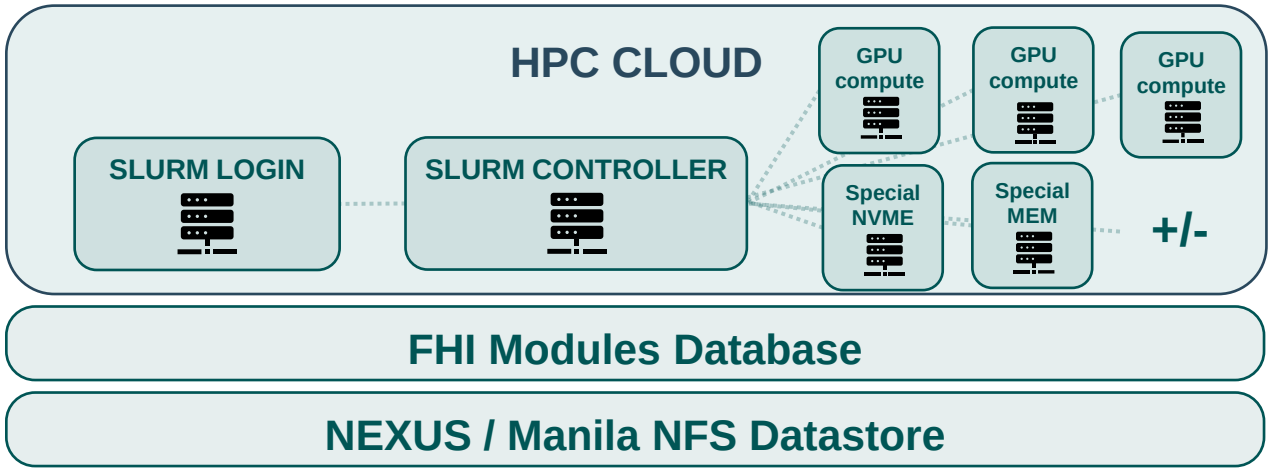
- Setup / testing
- High effort to set up

Future plans

- Direct network-link FHI-MPCDF
  → Controlled link latency

Resources (total)

- 1408 cores – 56TB RAM
- 44 A100 GPUs
- 140TB (local) NVME storage
- Manila NFS
- Nexus share
  (mounted at Raven & Cobra)



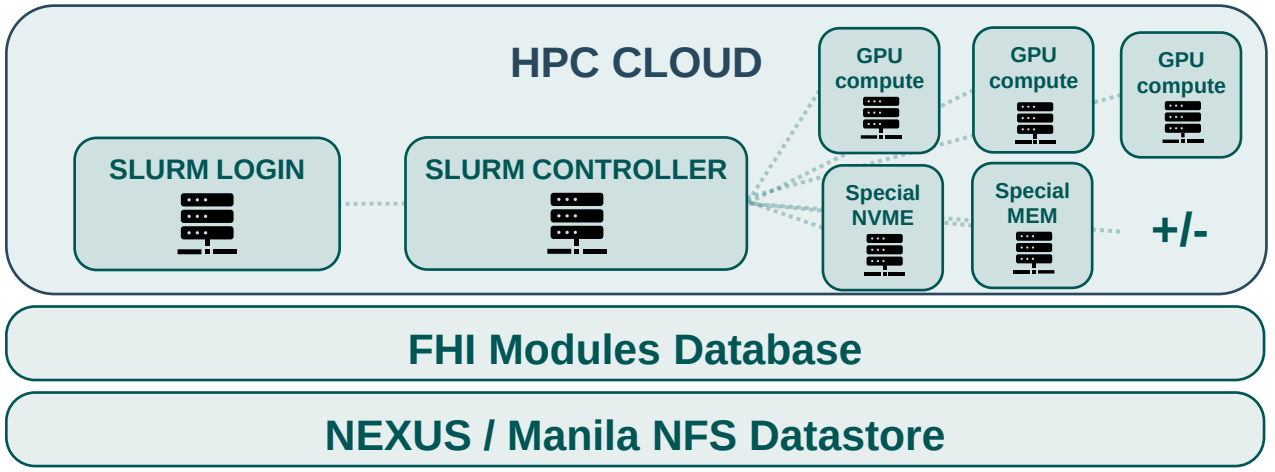
TH – GPU CLUSTER RACCOON

Konstantin Jakob
jakob@fhi.mpg.de

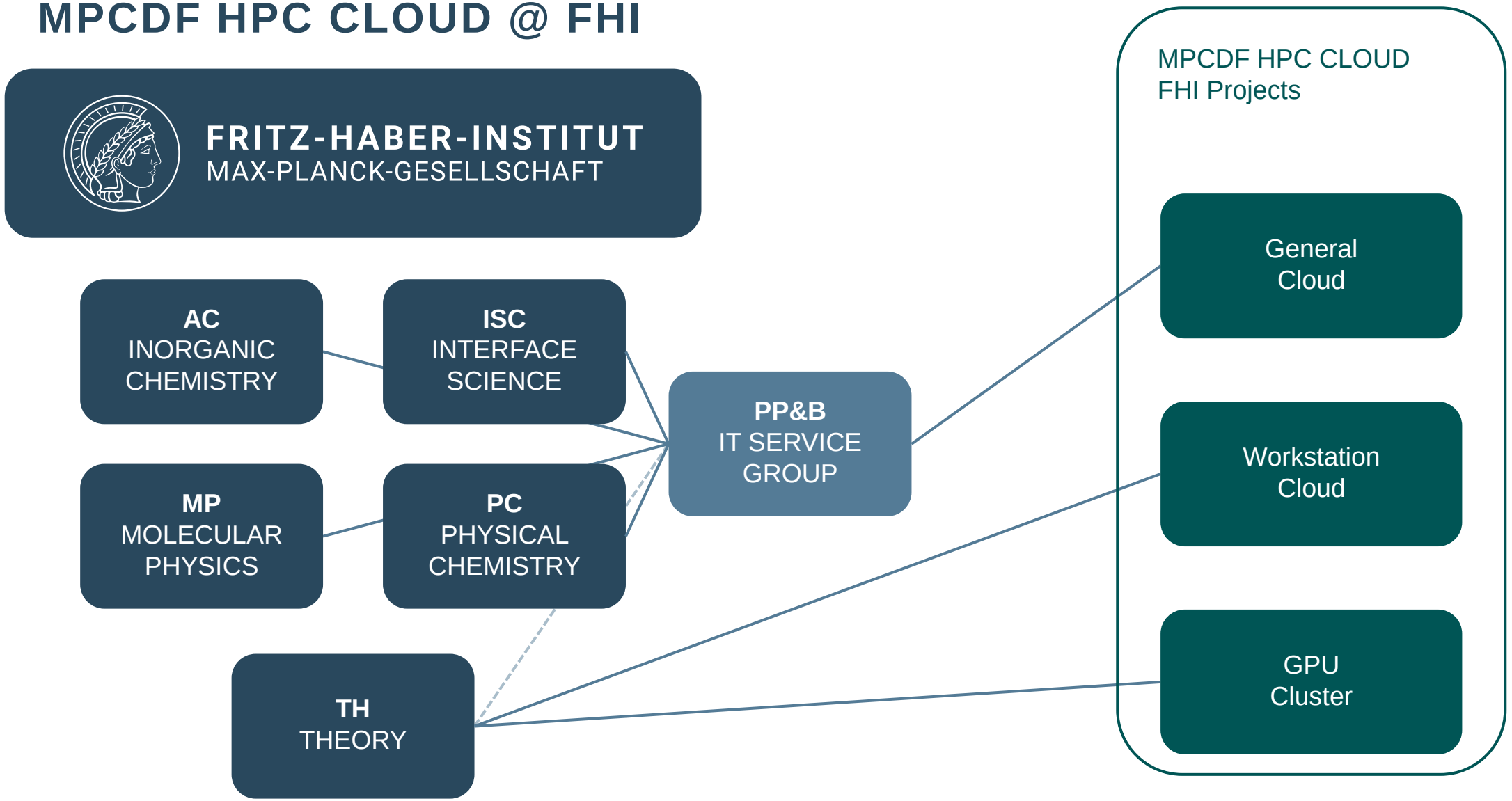
Implementation

- SLURM cluster (+ dedicated GPU workstations for testing)
- GPU nodes: 32c960g + A100 / 64c3840g + NVME + A100 / …
- Custom image (Packer):
  - Ubuntu 22.04, MPCDF user accounts, Nexus, Manila NFS
  - GPU packages, Slurm config, …
- Orchestrated (Jade / Terraform):
  - Flexible scaling (node count & type)
- Modules system:
  - Loading modules works by setting PATH variable
  - Adapted / extended from Raven, e.g. CASTEP
  - Shared Manila NFS → All compute nodes

```
module purge
module load gcc mkl/2022.2gsl/2.4openmpi/4.1fftw-mpi/3.3.9
module load anaconda/3/2023.03
module load cuda/12.1cudnn/8.9.2
module load cmake/3.22
```



MPCDF HPC CLOUD @ FHI



MPCDF HPC CLOUD @ FHI

FRITZ-HABER-INSTITUT
MAX-PLANCK-GESELLSCHAFT

AC
INORGANIC CHEMISTRY

ISC
INTERFACE SCIENCE

MP
MOLECULAR PHYSICS

PC
PHYSICAL CHEMISTRY

PP&B
IT SERVICE GROUP

Linux Workstations

MPCDF HPC CLOUD
FHI Projects

GPU Workstations

General Cloud

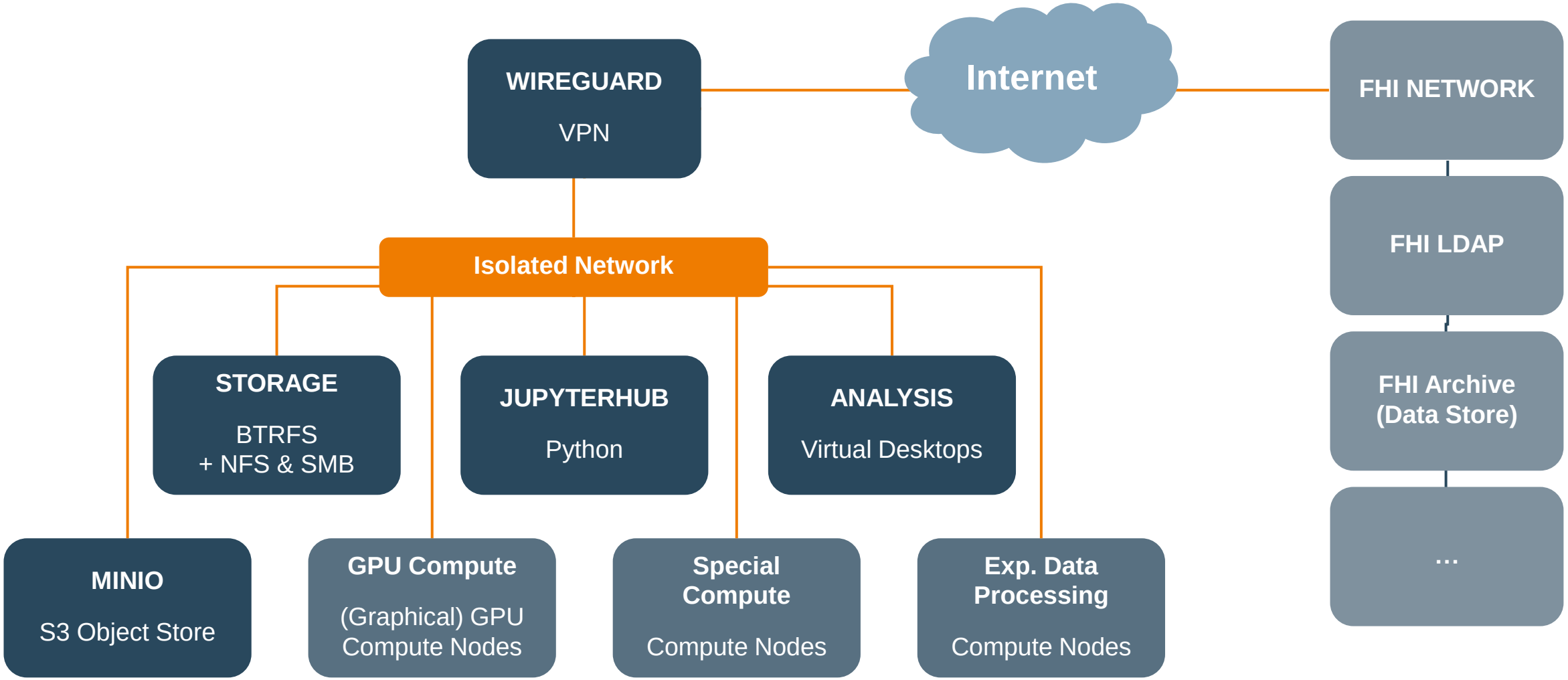
Compute Servers

Any use case where Windows PCs are not sufficient

GPU Cluster

All Compute Resources at MPCDF

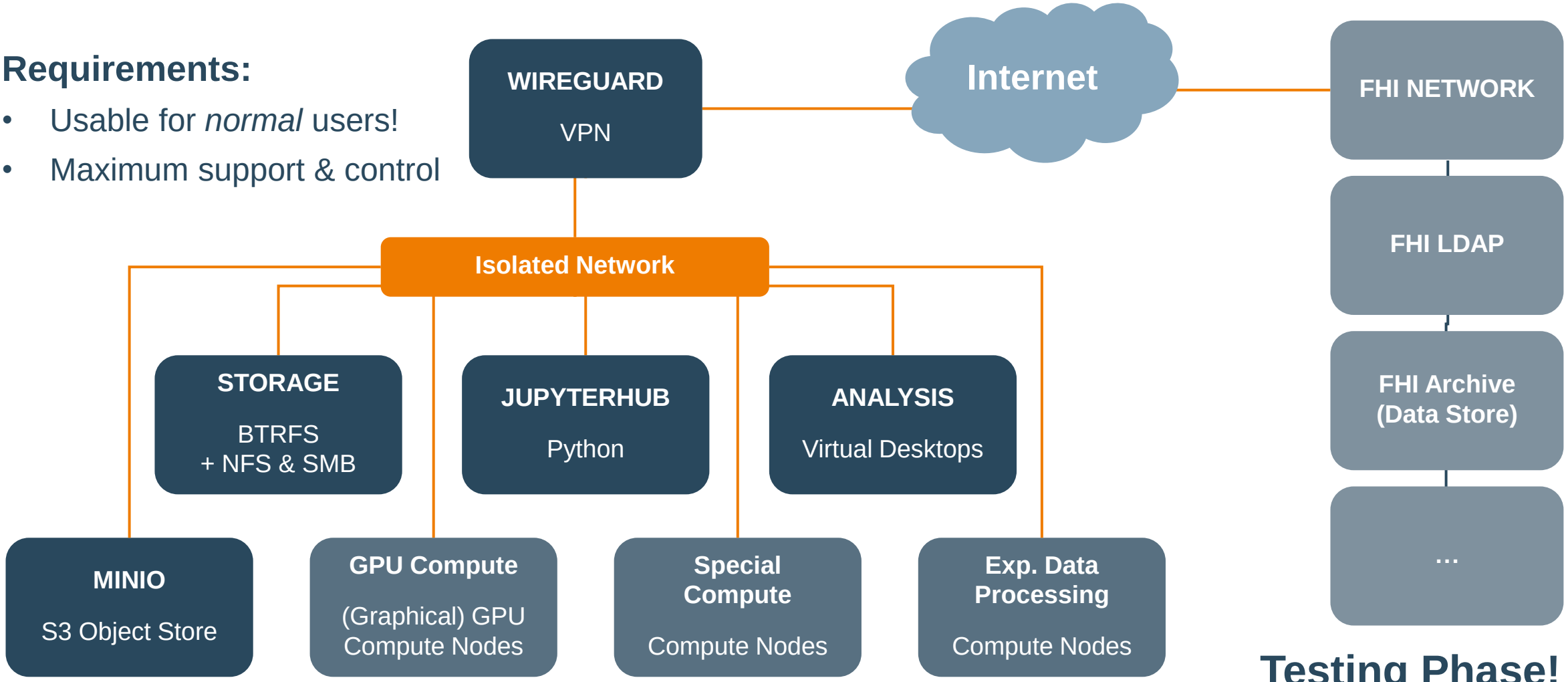
FHI – GENERAL CLOUD OVERVIEW



FHI – GENERAL CLOUD OVERVIEW

Requirements:

- Usable for *normal* users!
- Maximum support & control



Testing Phase!

FHI – JUPYTERHUB

Use cases

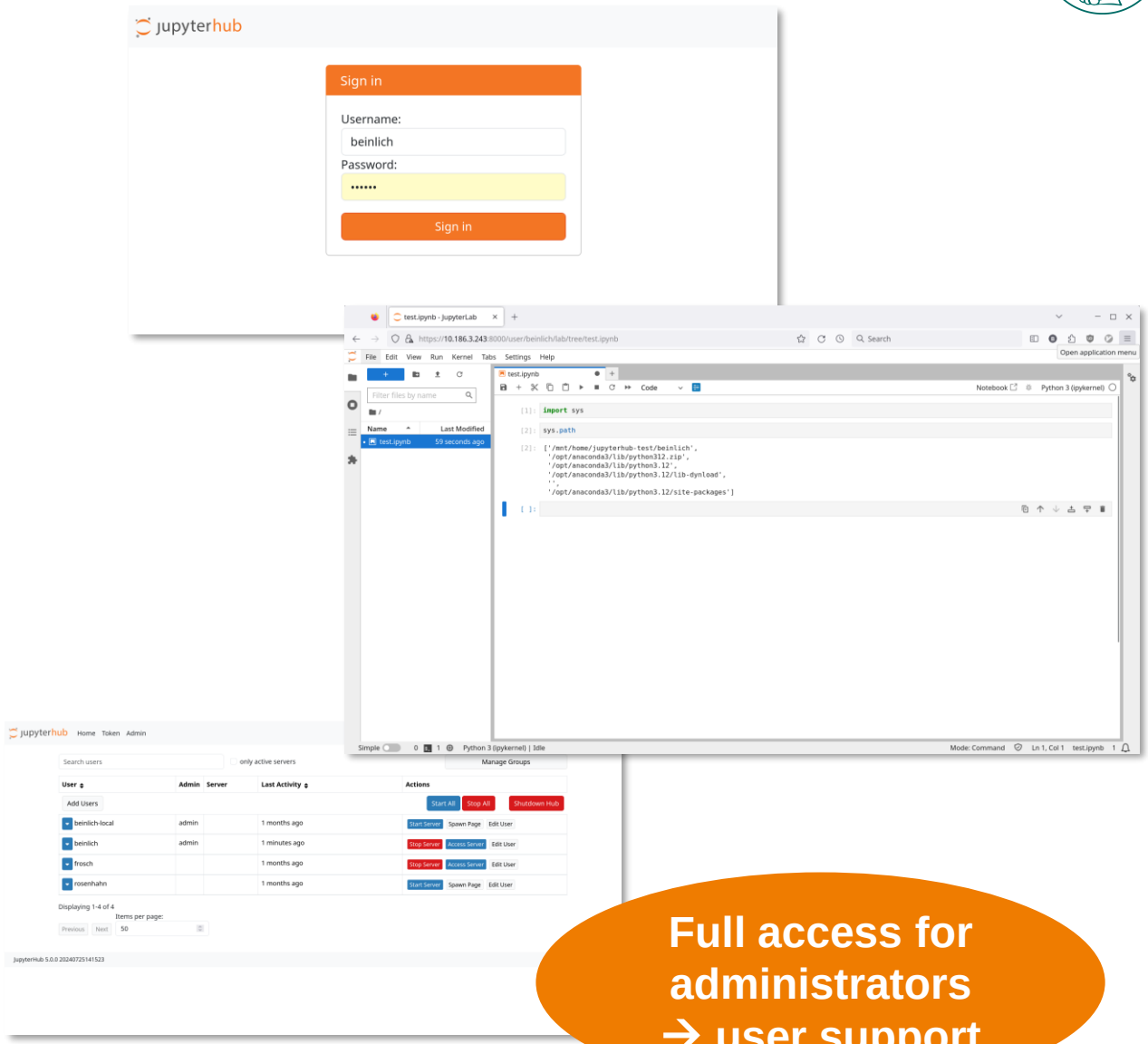
- Ready-to-run Python server
- Analysis – computing – programming

Implementation

- Debian 12 + JupyterHub
- LDAP accounts + NFS $HOME
- High resource (Memory, CPU, …)
- Access: SSH / WEB
- Soon: live collaboration

Advantages

- Direct admin access & support
- Ready-to-run preparation



FHI – GRAPHICAL VIRTUAL SHARED WORKSTATIONS

Use cases

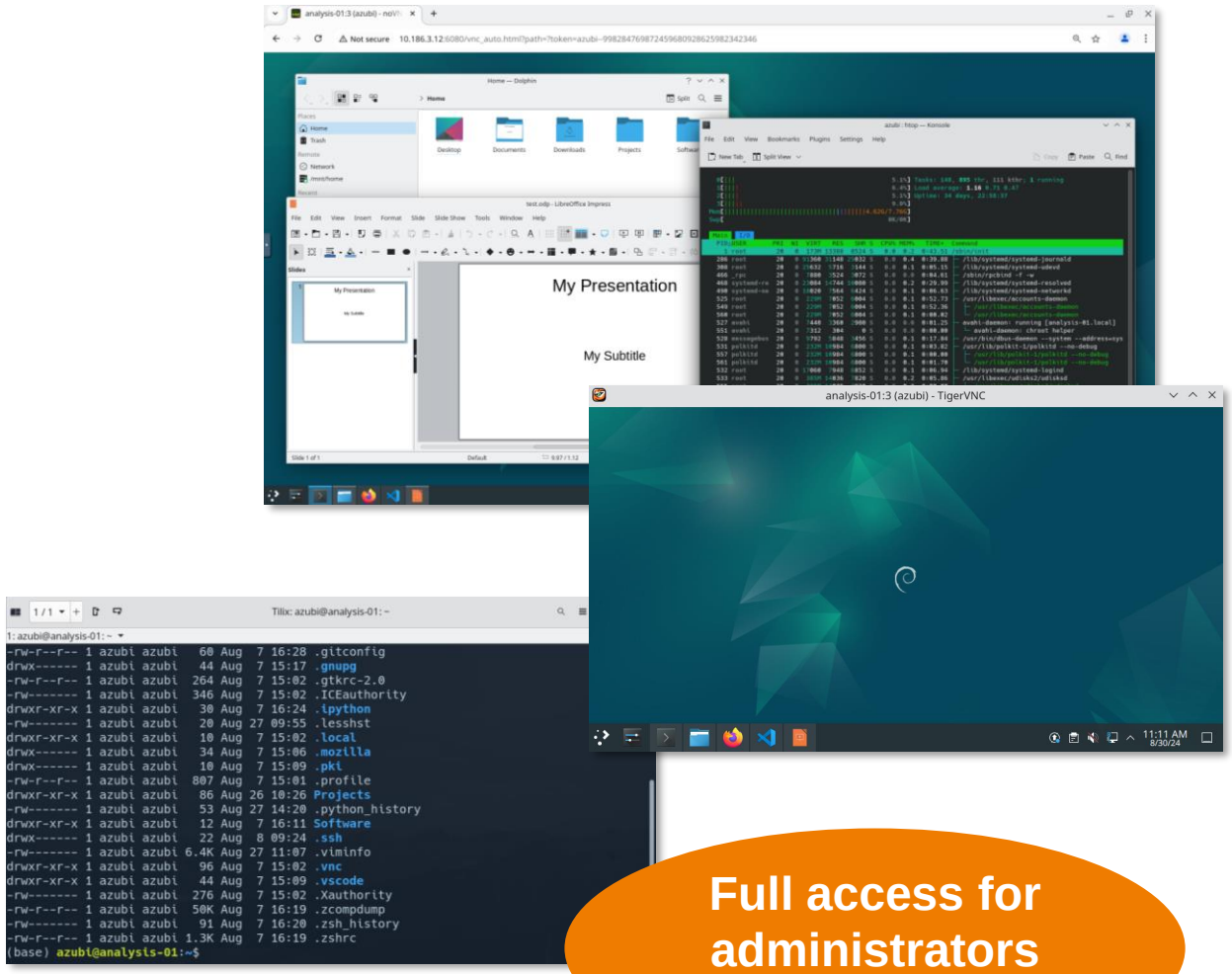
- Access to powerful Linux systems (shared)
- Analysis – computing – programming

Implementation

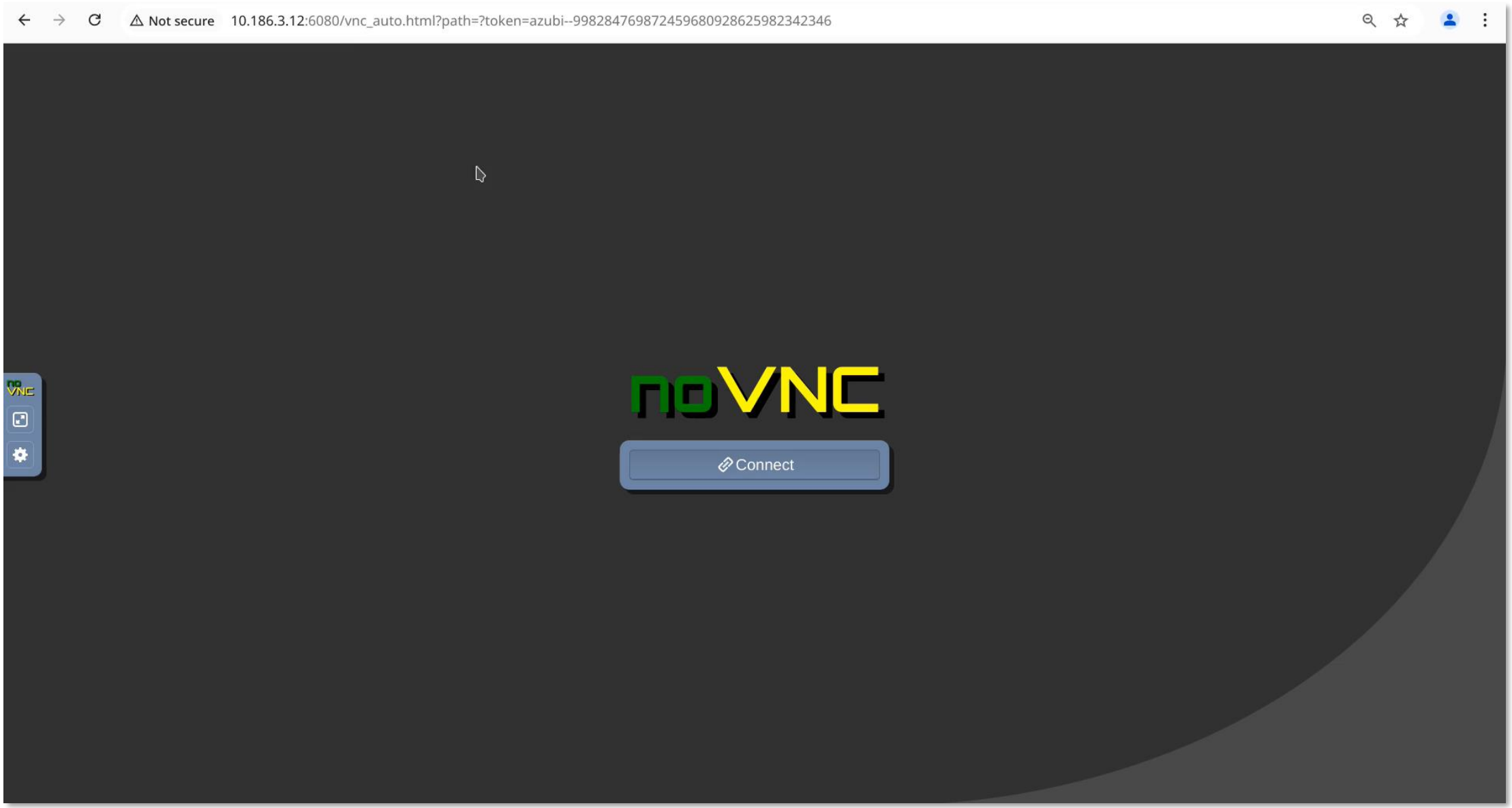
- Debian 12 + KDE Plasma
- LDAP accounts + NFS $HOME
- VNC server (tigervnc) + noVNC browser access
- Live collaboration (one ‘shared’ virtual screen)
- High resource (Memory, CPU, …)
- Access: SSH / SSH+VNC / WEB (noVNC)

Advantages

- Direct admin access & support
- Ready-to-run preparation



Full access for administrators → user support


Not secure
10.186.3.12:6080/vnc_auto.html?path=?token=azubi--998284769872459680928625982342346
noVNC
Connect

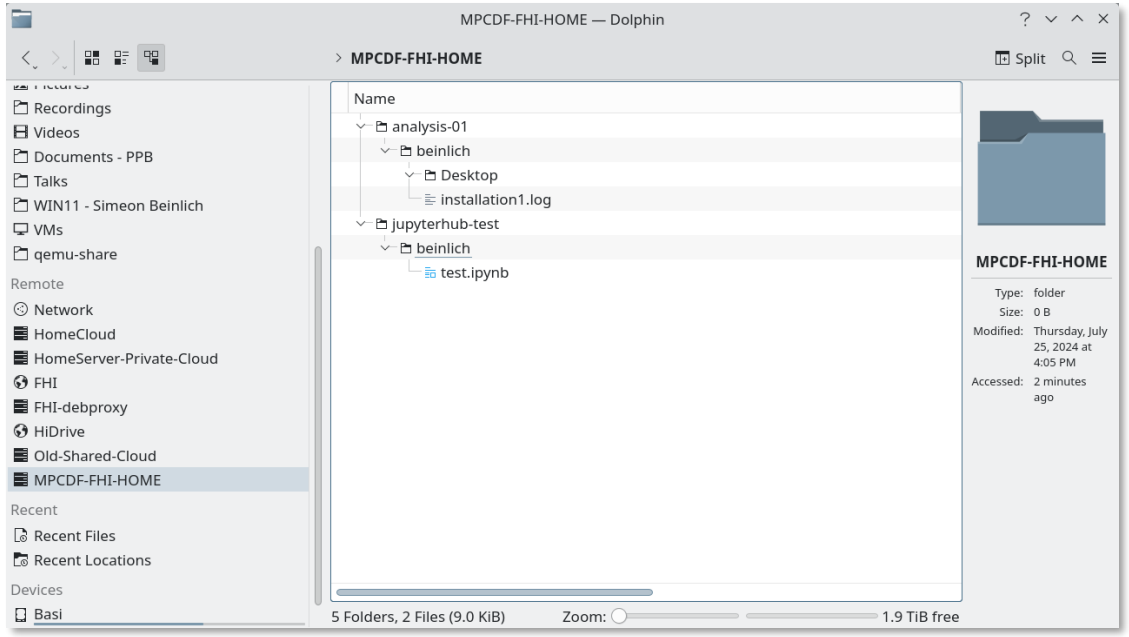
FHI – NFS STORAGE FOR MPCDF SYSTEMS

Use Cases

- Network storage for MPCDF-side FHI service

Implementation

- Debian 12
- BTRFS (snapshotted and compressed)
- NFS (MPCDF Linux clients)
- SMB (FHI Windows clients)
- LDAP accounts



Advantages

- Full storage control, simple backup, simple snapshotting / versioning
- For users:
  - *Just another SMB share on their local PC*
  - Access to filesystems of all MPCDF-side FHI services

Maybe:

- Replace with Manila NFS + SMB Bridge?

FHI – S3 BUCKETS FOR USERS AND SYSTEM

HPC CLOUD CEPH Buckets

+ Simple bucket & credential creation
+ High performance

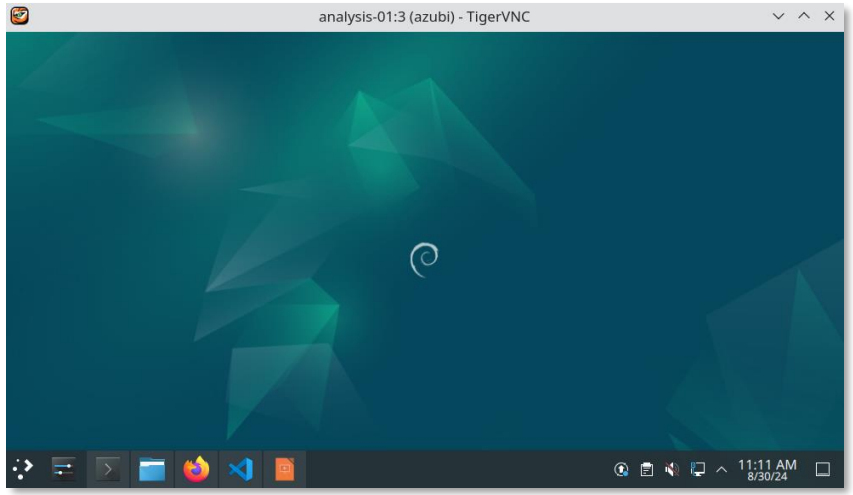
- Object-size restrictions (>4MB/obj)
  → But we have no control over user usage!
- No per-bucket / per-user credentials
  → No end-user buckets

- Used for system / administrative buckets
  - Mainly Restic backups of servers and important data storages

MINIO Server (VM)

+ Simple bucket & credential creation (if scripted)
+ Per-bucket credentials
+ High flexibility & scalability & replication
+ No object-size limitations
- Some setting-up effort
- Object storage on top of virtual block storage
  → performance?
- Used for end-user buckets:
  - Restic backups (self-administrated devices)
  - Data stores …

FHI – OTHER PLANNED SERVICES



GPU / specialty compute VMs

- Virtual desktop-like GPU nodes for AI / ML
- Huge memory nodes / NVME nodes for computing
- Single nodes with unlimited walltime/direct execution.

Processing nodes for experimental outputs

- VMs for automated compute-intensive data postprocessing

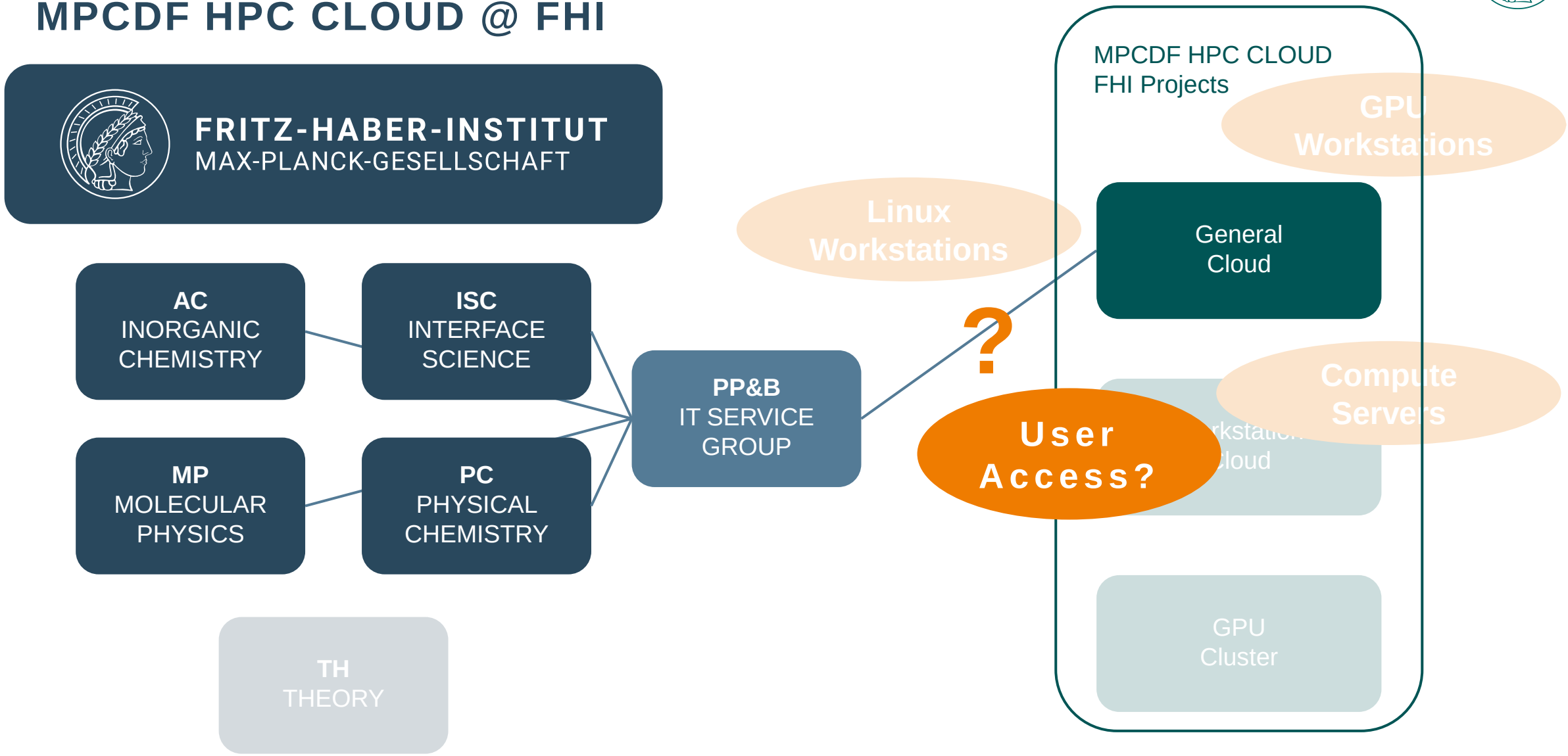
Raven access / dedicated SLURM cluster

- Raven access for general FHI users ← vs. → cedicated virtual SLURM cluster

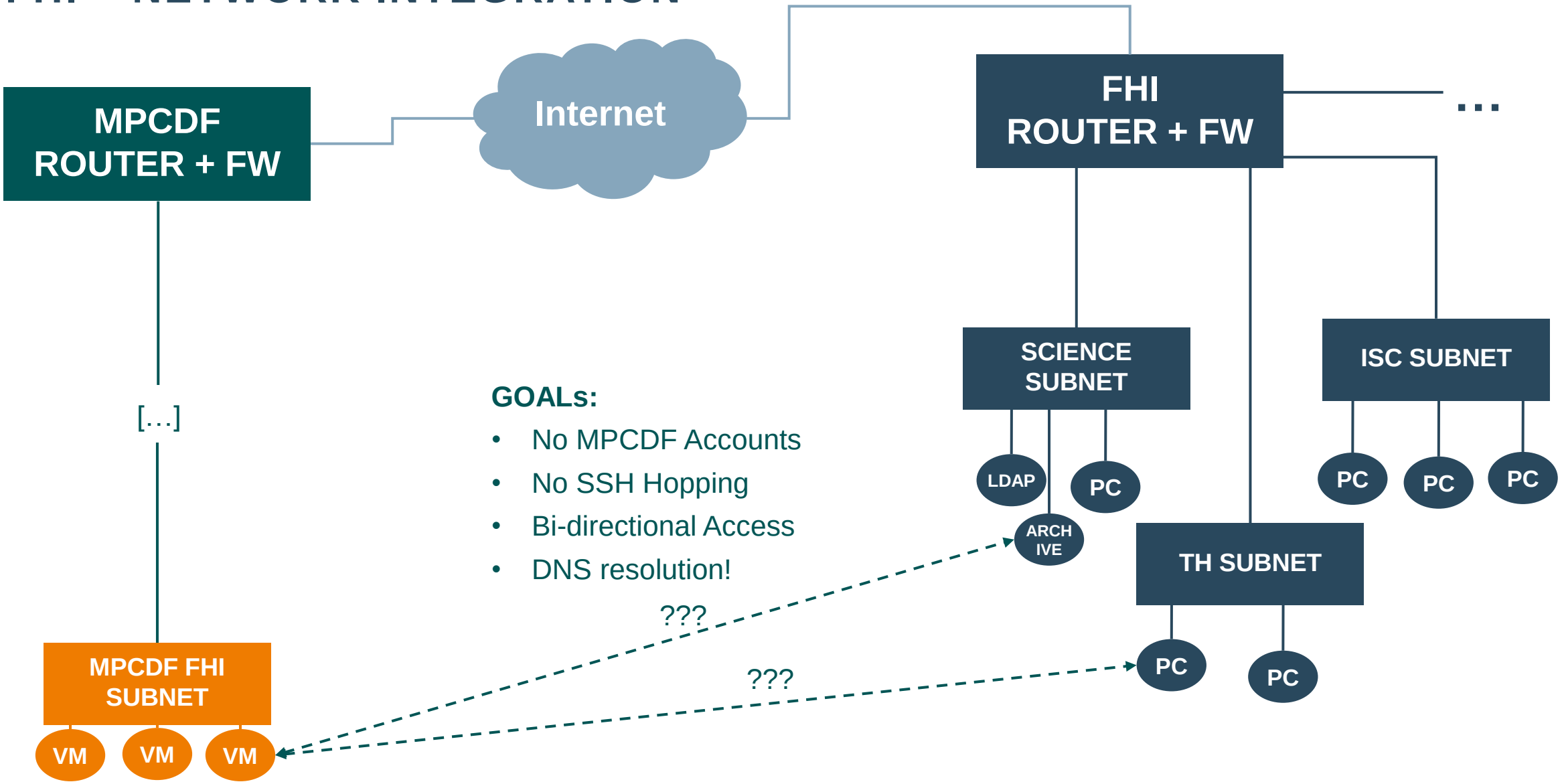
[…]

➡ **All Compute Resources at MPCDF**

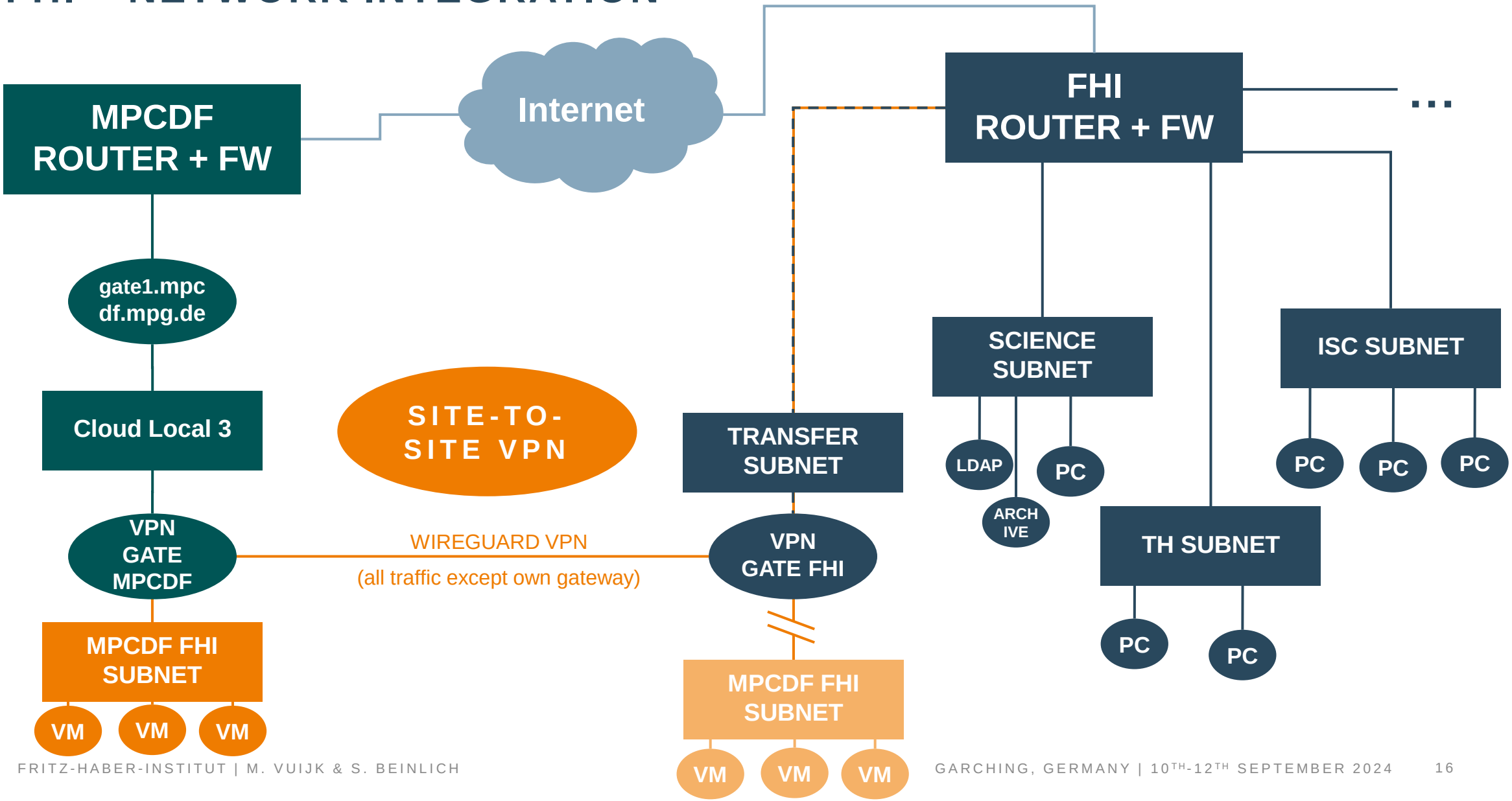
MPCDF HPC CLOUD @ FHI



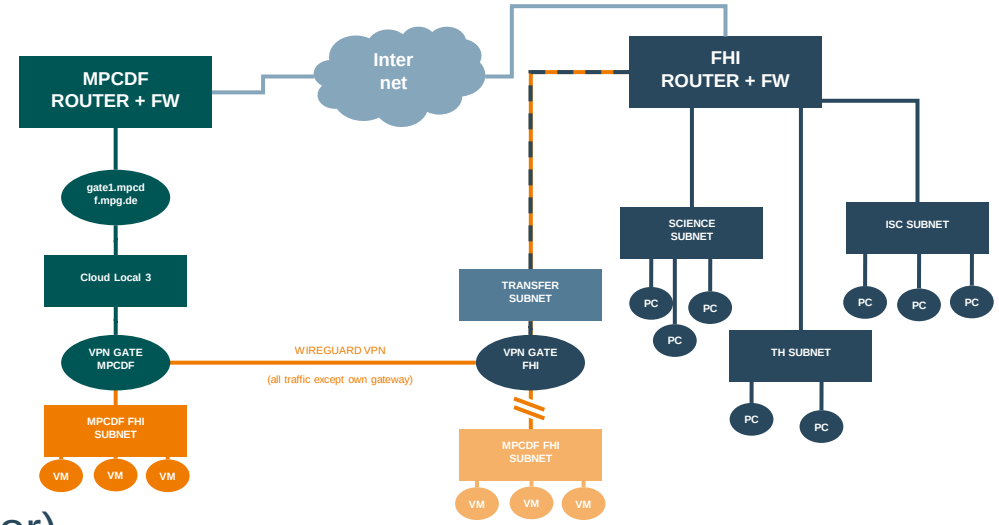
FHI – NETWORK INTEGRATION



FHI – NETWORK INTEGRATION



FHI – NETWORK INTEGRATION



Use Cases

- Direct access to (our) Virtual Resources at MPCDF
- Direct access to resources at FHI (LDAP, Backup, …)

Implementation

- Wireguard Site-to-Site VPN (via forwarding)
- MPCDF subnet: private isolated subnet (all traffic via FHI)
- FHI-side routing via transfer subnet (VLAN tagged to hypervisor)
- MPCDF-side routing by construction:
  MPCDF-side WG == MPCDF FHI subnet gateway
- Virtual subnet behaves like ‘real’ subnet → DNS, Firewall, etc. from FHI DNS

Advantages

- Direct full access (bi-directional)
- No hopping → No SSH knowledge required
- User don’t recognize any difference to a ‘real’ server at FHI

Disadvantages

- Public IPs?
  (not planned anyway)
- MPCDF FHI subnet → MPCDF services
  Munich → Berlin → Munich

SUMMARY

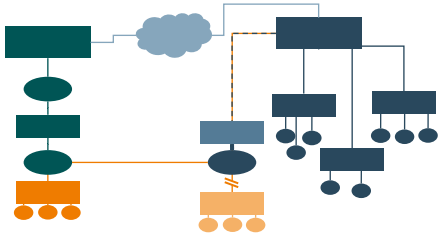
THANK YOU!

MPCDF HPC CLOUD
FHI Projects

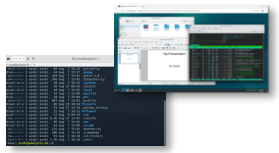
General Cloud

Workstation Cloud

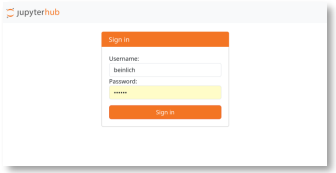
GPU Cluster



Integrated Virtual Subnet

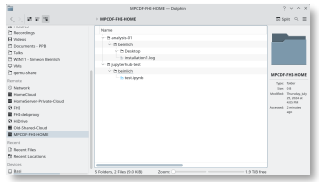


Virtual Desktops



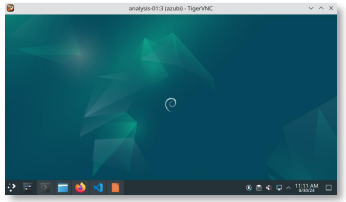
JupyterHub

S3 Buckets
(Ceph / Minio)

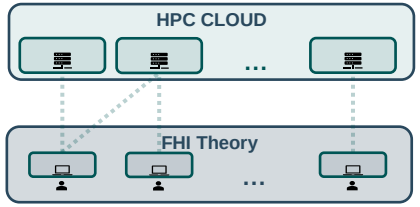


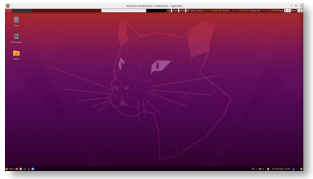
NFS / SMB Server

. . .



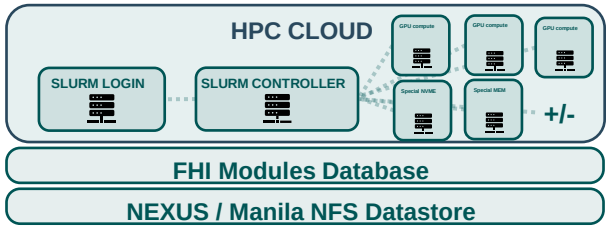
GPU / Compute VMs



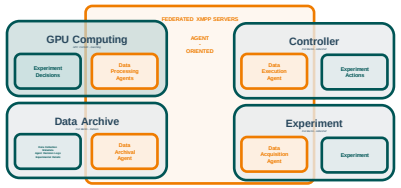


+ Binder Hub

Virtual Personal & Project Workstations



Scalable GPU / Specialty SLURM Cluster



Agent-Oriented Just-in-Time Computing