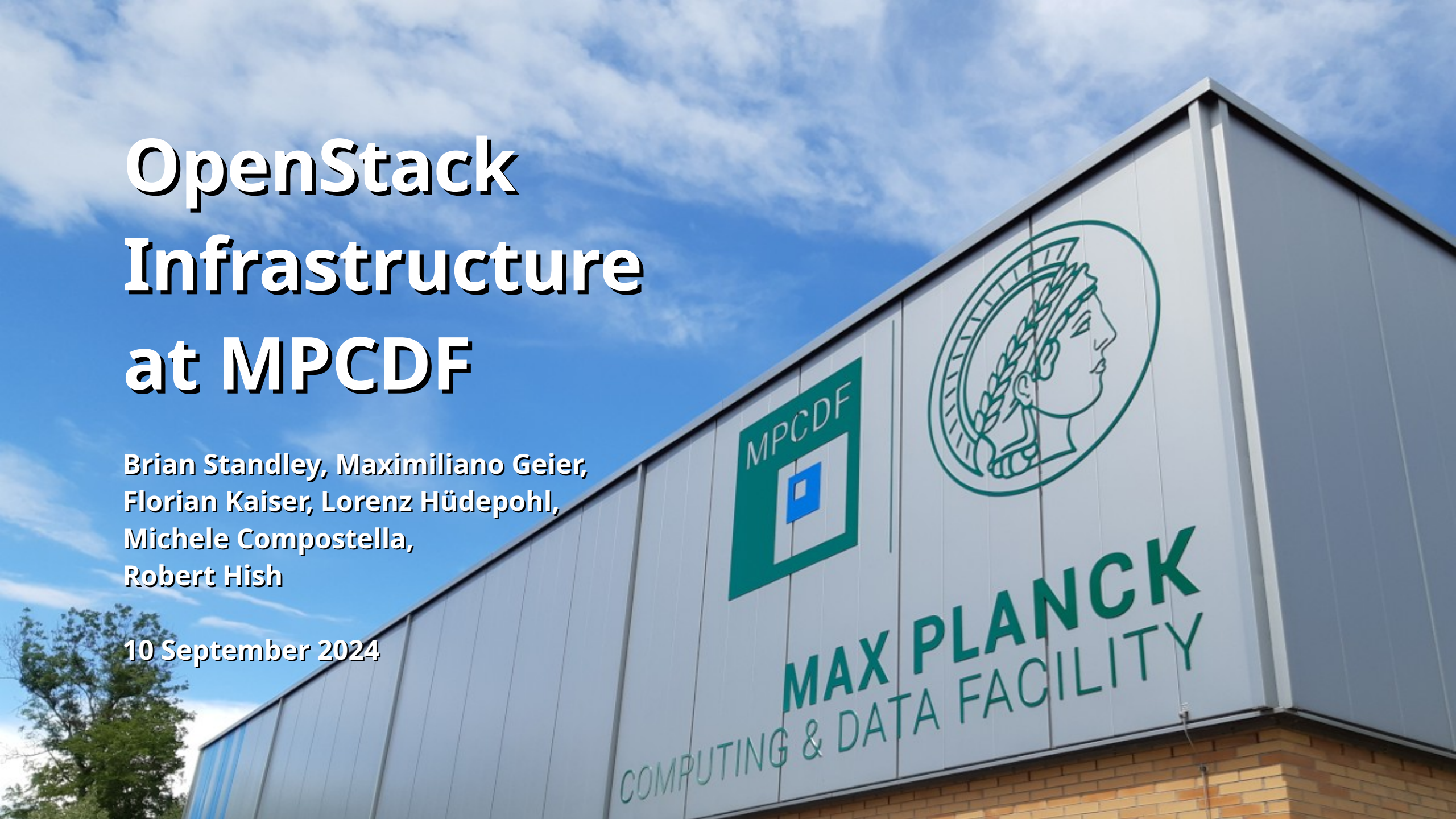
OpenStack Infrastructure at MPCDF

Brian Standley, Maximiliano Geier,
Florian Kaiser, Lorenz Hüdepohl,
Michele Compostella,
Robert Hish

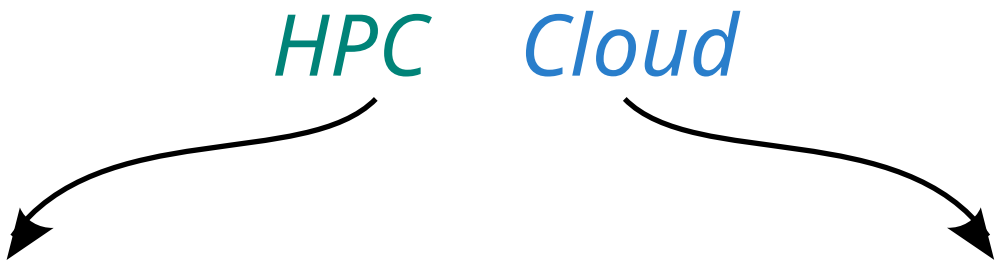
10 September 2024



CONCEPT (AND HISTORY)

MAX PLANCK COMPUTING & DATA FACILITY

A general solution for complex workflows, complementing the HPC systems (2020)

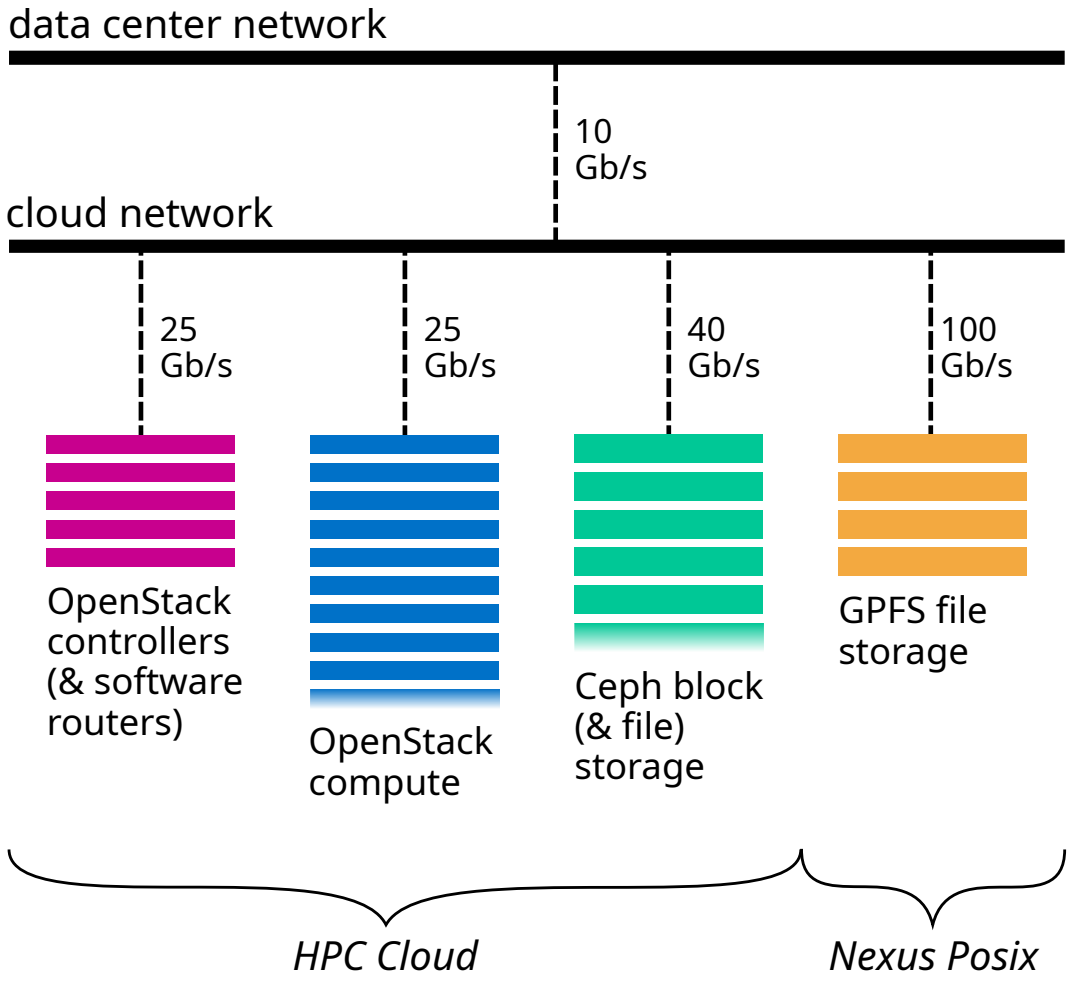


- ✔ Logically positioned near supercomputing resources, esp. Raven
- ✔ Contains significant compute power
- ✕ Not itself a parallel compute cluster…

- ✔ Flexible computing environment
- ✔ Infrastructure-as-a-Service inc. self-service dashboard, standard APIs
- ✕ Not intended for core IT services…

Predecessors: 2016-2018 "Testenv" Packstack-based (w/o Ceph), IBM BladeCenter compute
2018-2021 "Susecloud" SUSE-based (inc. Ceph), IBM dx360m4 compute, Lenovo SR650 storage

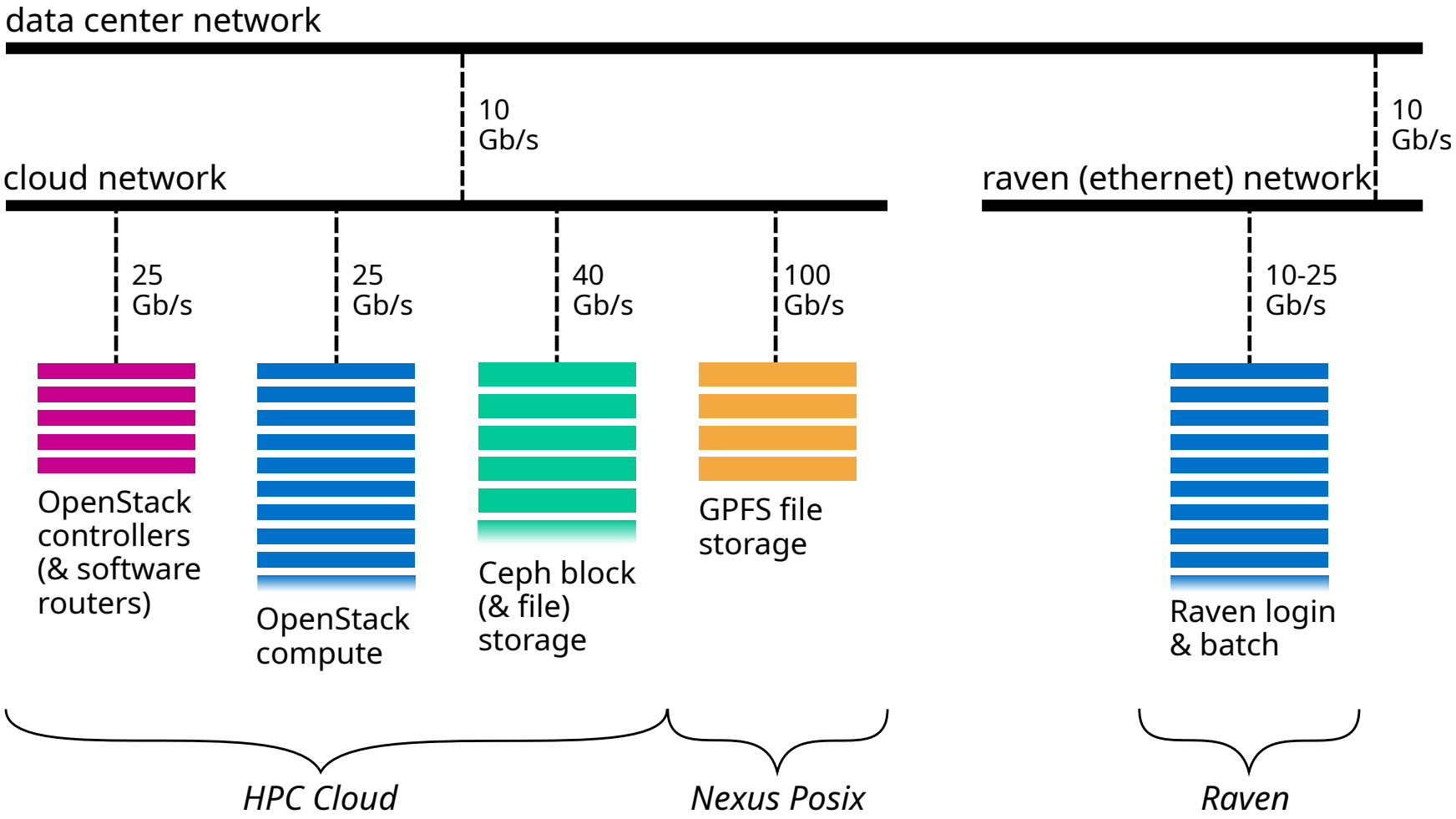
NETWORK ARCHITECTURE



2021

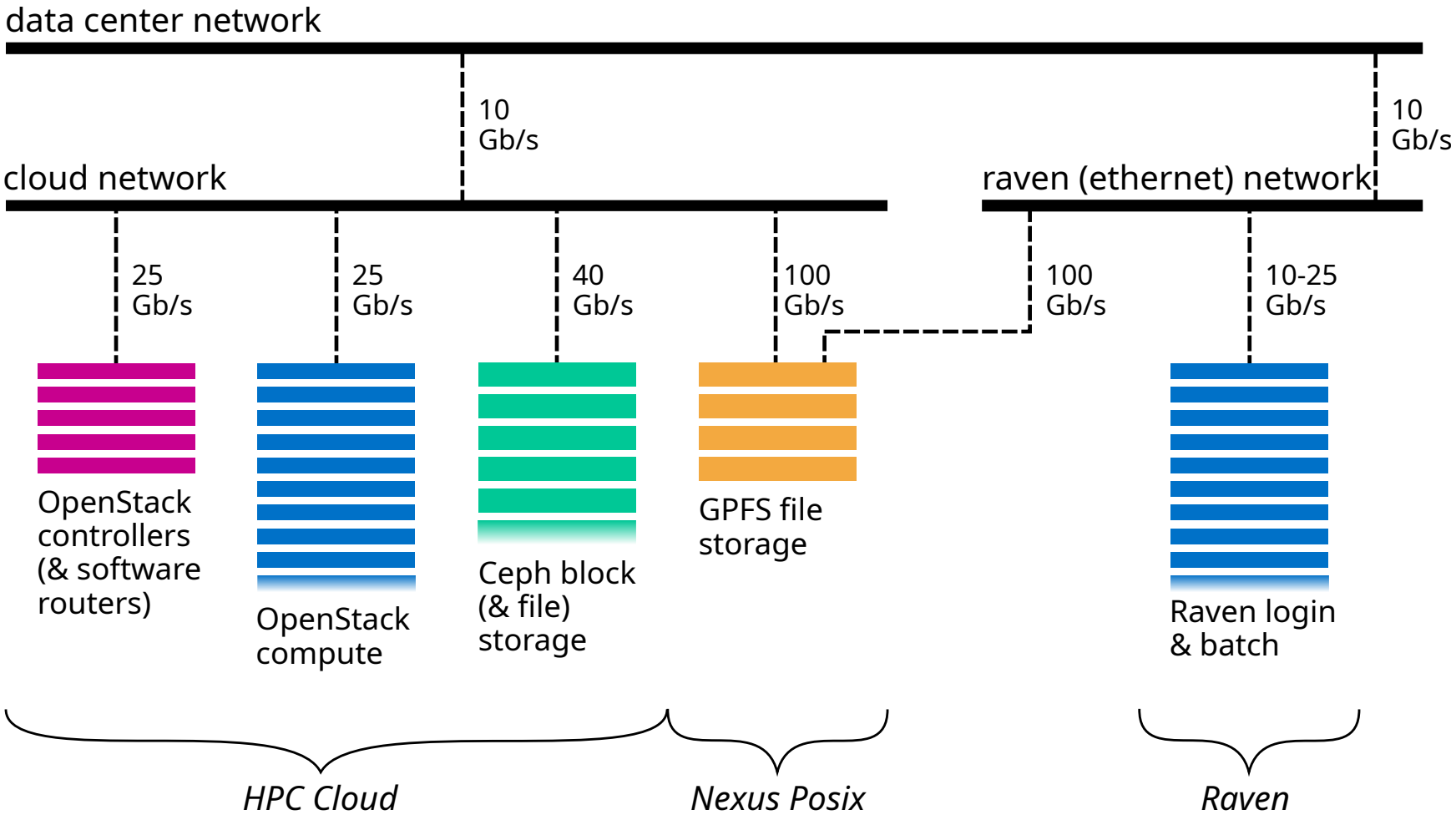
NETWORK ARCHITECTURE

MAX PLANCK
COMPUTING & DATA FACILITY



2021

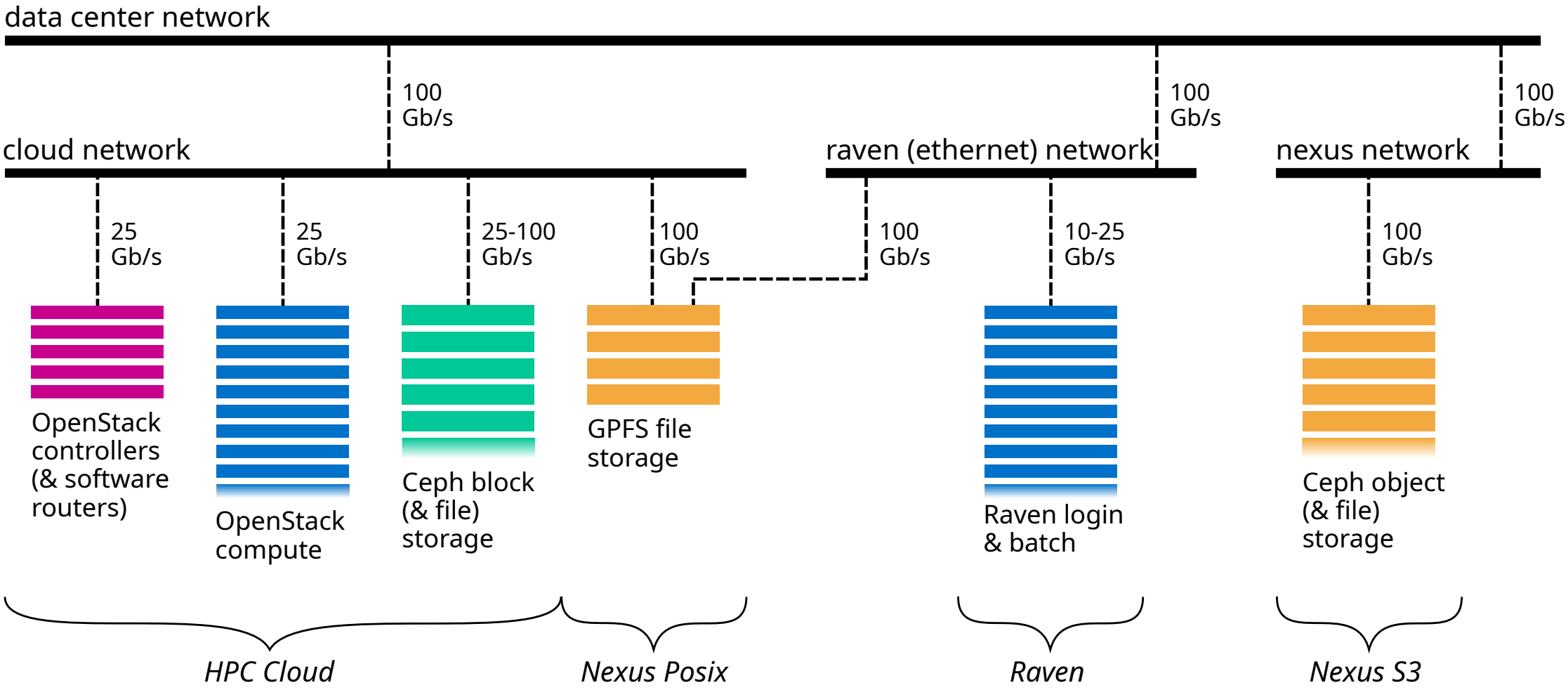
NETWORK ARCHITECTURE



2021

NETWORK ARCHITECTURE

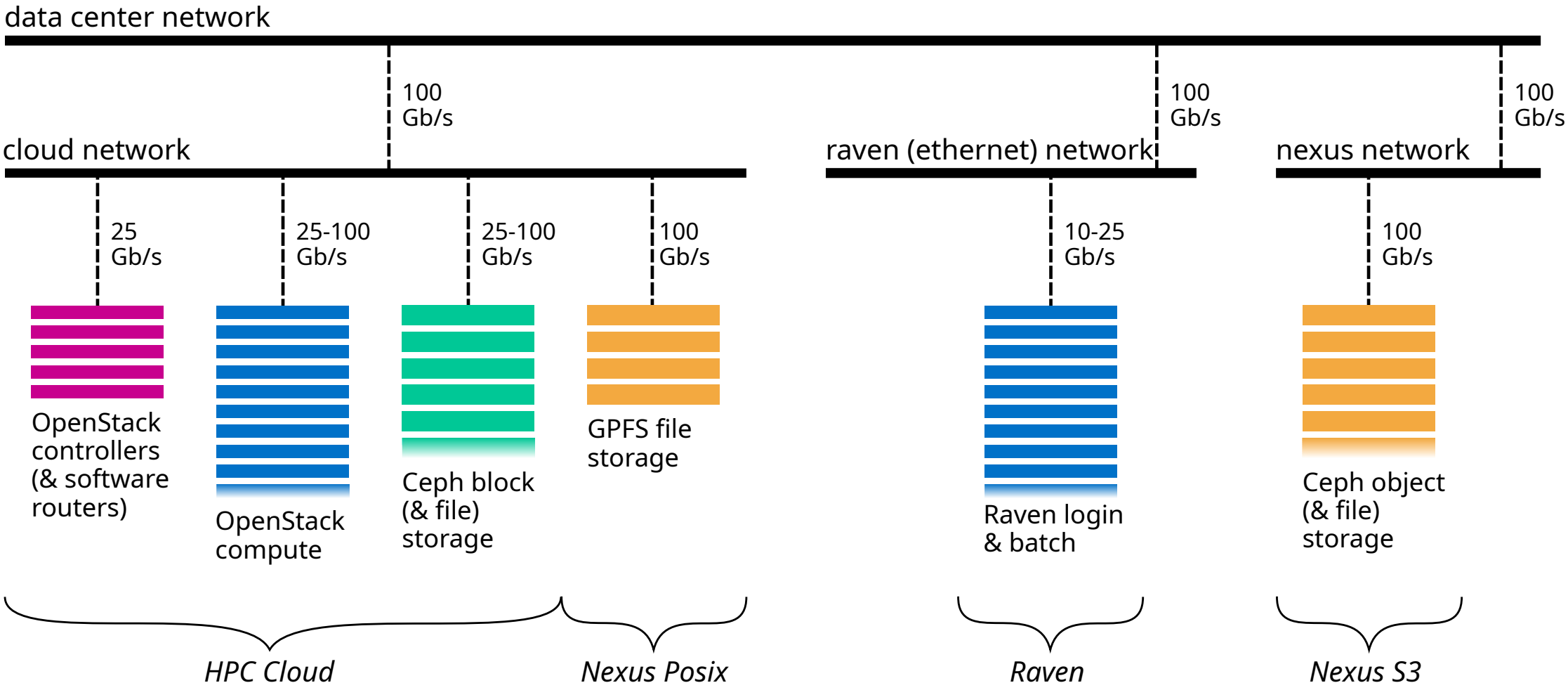
MAX PLANCK COMPUTING & DATA FACILITY



2022

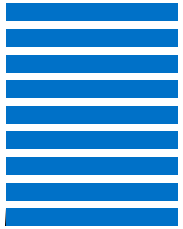
NETWORK ARCHITECTURE

MAX PLANCK COMPUTING & DATA FACILITY



HPC CLOUD IN NUMBERS

MAX PLANCK COMPUTING & DATA FACILITY



OpenStack compute

CPU: 2×Intel Icelake 32-36C – 110 nodes
2×AMD Genoa 48-64C – 22 nodes
Various OBS workers – 5 nodes

RAM: 256-384GB – 5 nodes
512-768GB – 60 nodes
1024GB – 12 nodes
1536GB – 10 nodes
2048GB – 44 nodes
4096GB – 6 nodes

GPU: 12×Nvidia A30-24GB
24×Nvidia A40-48GB
60×Nvidia A100-80GB
8×Nvidia H100-94GB

NVMe: 60×1.6TB
12×3.2TB

10016 Cores
172 TB RAM
104 GPU
134 TB NVMe

Ceph block (& file) storage

2×AMD Rome 16C, 256 GB – 5 nodes
2×Intel Sap. Rapids 12C, 192 GB – 8 nodes
12×3.84TB SATA SSD – 5 nodes
4×7.68GB NVMe SSD – 8 nodes
24×22TB SATA HDD – 8 nodes

1.2 PB HDD pool (net, rep3)
70 TB SSD pool (net, rep3)
50 TB NVMe pool (net, rep3)

HPC CLOUD IN PHOTOS

MAX PLANCK COMPUTING & DATA FACILITY

OpenStack staging/ test cluster

Monitoring cluster

Ceph cluster HDD/NVMe

OpenStack controllers

OpenStack compute (HA)

Ceph staging/ test cluster

Ceph cluster SSD

Controllers/ Nexus Posix/ Ceph cluster

Downlinks

Uplinks

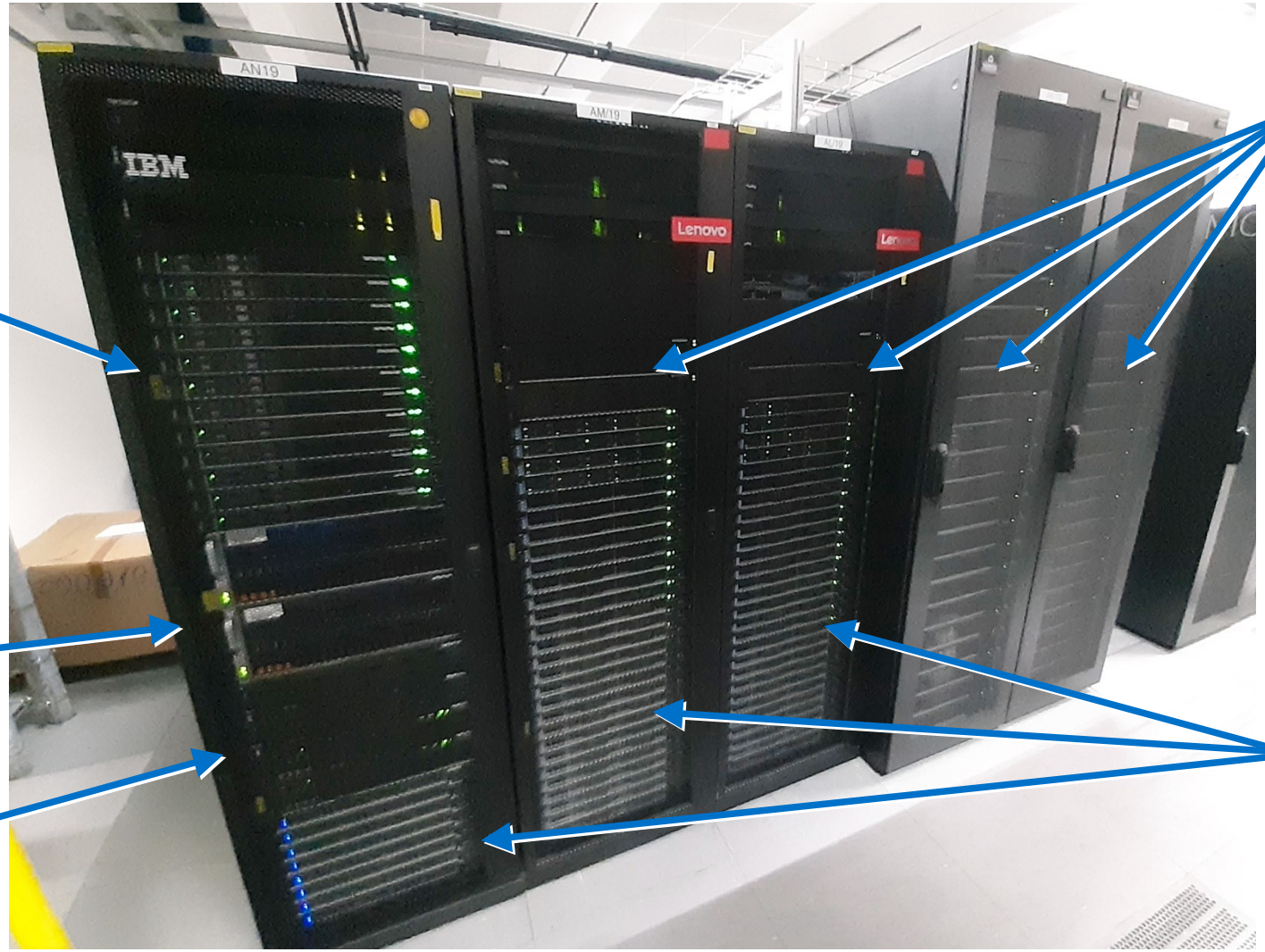
HPC CLOUD IN PHOTOS

MAX PLANCK COMPUTING & DATA FACILITY

Genoa
w/o GPU

Genoa
w/ GPU

Haswell,
Skylake
(OBS workers)



Icelake
w/ GPU

Icelake
w/o GPU

SUPPORTING SOFTWARE

MAX PLANCK COMPUTING & DATA FACILITY

- Software stack
  - OpenStack Wallaby
    - *2021.04 release*
    - *On par with latest RHOSP (v17.1)*
- Microservice architecture
  - Containers (*podman*)
- HA-enabled control plane
- KVM-based virtual machines
- NFV provided by OVN

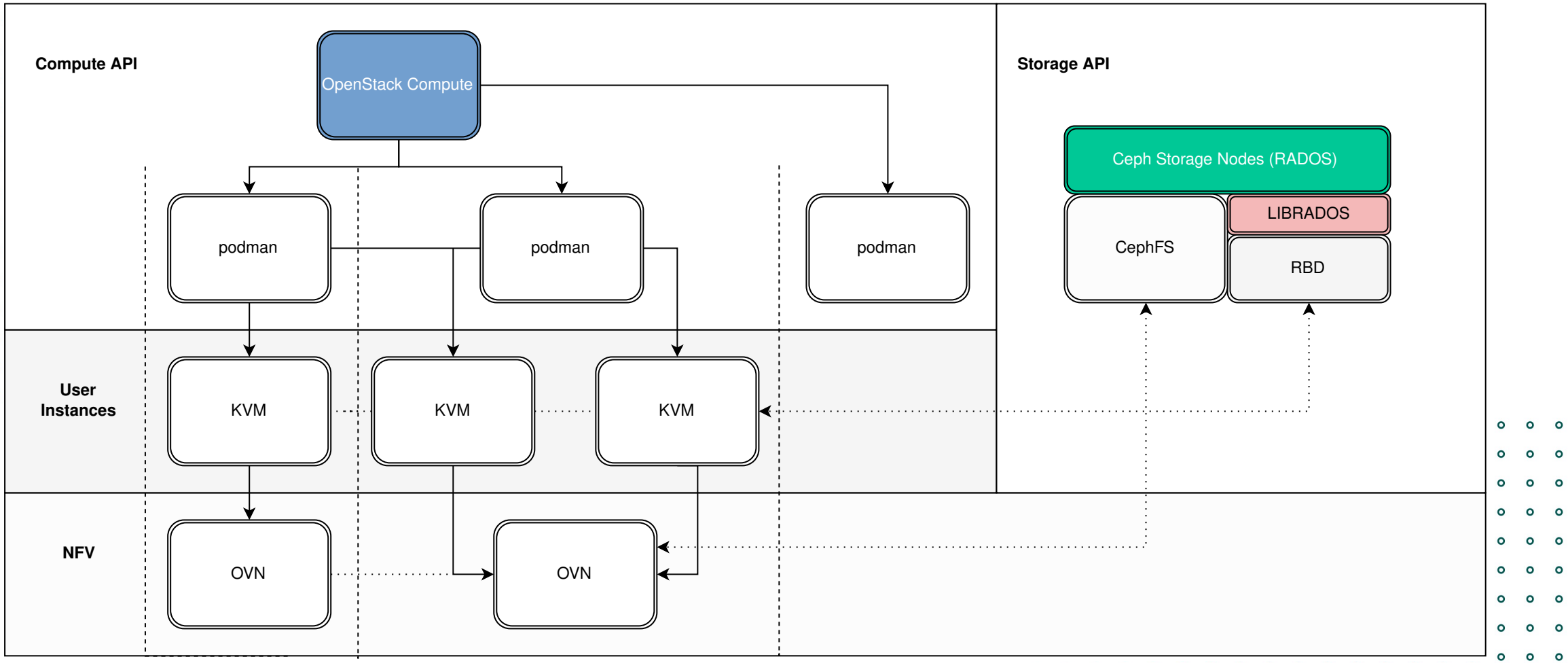
Compute

- Software stack
  - Ceph Quincy (v17, 2022.04)
- Block Storage
  - Block devices for instances (RBD)
- File Storage
  - POSIX-like filesystem (native or NFS) for shared storage
- Local Storage
  - Local disk space (not Ceph!)

Storage

SUPPORTING SOFTWARE

MAX PLANCK
COMPUTING & DATA FACILITY



SELECTED FEATURES

MAX PLANCK COMPUTING & DATA FACILITY

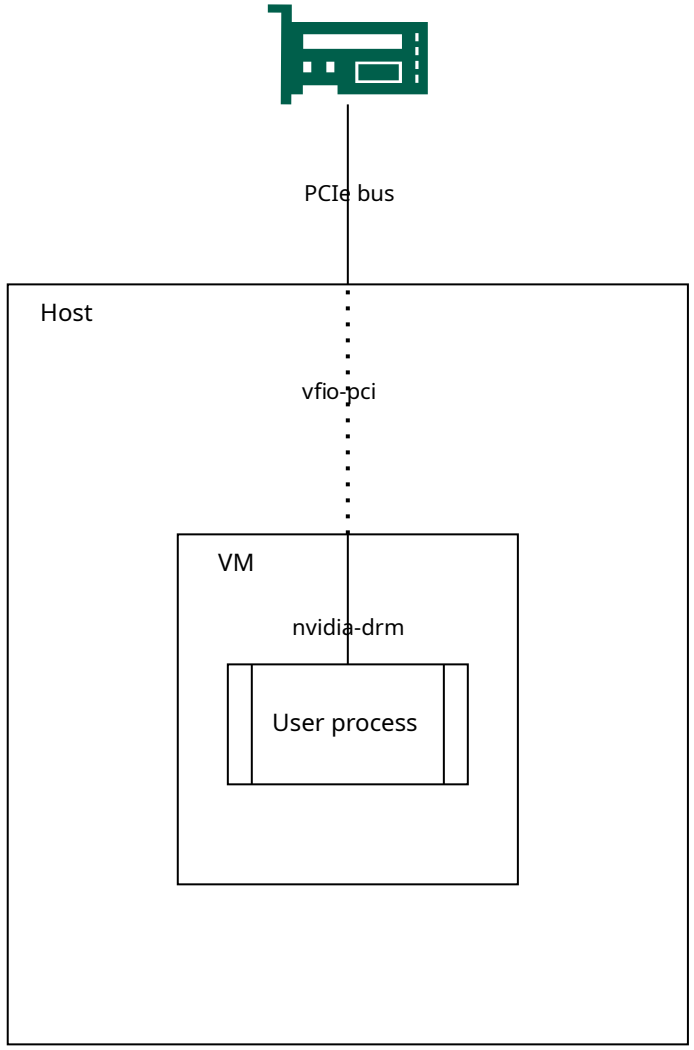
- File sharing
  - Manila
- Loadbalancers
  - Octavia
- Object storage
  - Swift
- Orchestration
  - Heat

- Live migration
  - Ceph-backed RBD blocks
  - Local disks
- Additional pluggable hardware
  - NVMes
  - **GPUs**

PCI PASSTHROUGH

MAX PLANCK COMPUTING & DATA FACILITY

- Expose physical PCI hardware to instances
  - Pre-configured hardware types
  - GPUs, disks
- Exclusive Access
  - Available hardware “unplugged” from host
- OpenStack-managed with caveats
  - Keeps track of hardware allocations
  - Hardware-agnostic
  - Client-side drivers
  - Instances cannot be live migrated



GPU INSTANCES

MAX PLANCK COMPUTING & DATA FACILITY

MPCDF

- Exclusive mode
  - ➔ Instance tied to hypervisor where allocated GPU resides (PCI Passthrough)
  - ➔ Inefficient resource allocation under some configurations

GPU INSTANCES

MAX PLANCK COMPUTING & DATA FACILITY

MPCDF

- Exclusive mode
  - Instance tied to hypervisor where allocated GPU resides (PCI Passthrough)
  - Inefficient resource allocation under some configurations

- Why not vGPUs?
  - Supported by OpenStack/KVM
  - Requires driver support on the *hypervisor*
  - Extra licensing costs with NVIDIA

GPU INSTANCES

MAX PLANCK COMPUTING & DATA FACILITY

- Exclusive mode
  - Instance tied to hypervisor where allocated GPU resides (PCI Passthrough)
  - Inefficient resource allocation under some configurations

- Why not vGPUs?
  - Supported by OpenStack/KVM
  - Requires driver support on the *hypervisor*
  - Extra licensing costs with NVIDIA
- Why not MIG? (Partitioning)
  - Partitions cannot be attached to instances (no virtual hardware device)
  - Support through Cyborg (Accelerator as a Service); untested
  - Cyborg service not available in our current setup (yet!)
  - Containers

WRAPPING UP

MAX PLANCK COMPUTING & DATA FACILITY

- Server OS Upgrades
- OpenStack Upgrades
- More OpenStack services
  - Cyborg: Accelerator as a Service
  - Ironic: Baremetal instances
- IPv6 on user instances
- More hardware
- *Your suggestion here!*

THANK YOU