MAX PLANCK
COMPUTING & DATA FACILITY

MPCDF

# GENERIC CLOUD STORAGE AT MPCDF

Florian Kaiser
MPCDF Storage Team

# ABOUT THE STORAGE TEAM

MAX PLANCK COMPUTING & DATA FACILITY

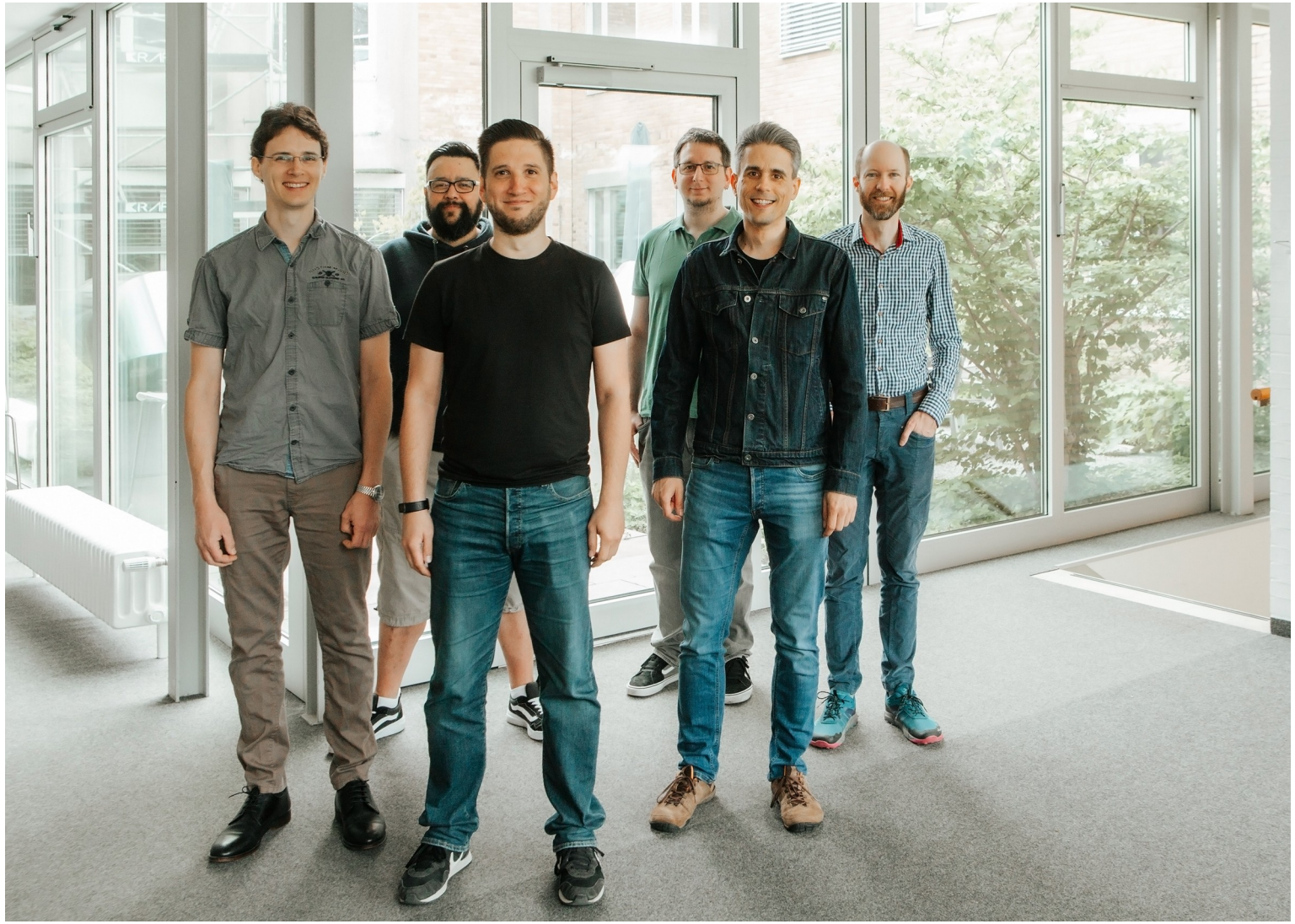
- About me: many different hats during the last ~10 years at MPCDF
  - Participated in the EUDAT project
  - Established first virtualization environments with Xen, KVM, VMware
  - Lead the MPCDF network team for a few years
- New storage team created ~2 years ago as an umbrella for existing services (DataShare, early Nexus POSIX) and to reduce the number of individual “silos”
  - Dedicated filesystems still make sense for larger HPC systems for best performance and to keep failure domains small
  - Procurement, installation and management becomes inefficient for smaller systems - which nowadays means single PB

# THE CLOUD & STORAGE TEAM

MAX PLANCK COMPUTING & DATA FACILITY

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

(from left to right, core storage team members **bold**)



# GENERIC STORAGE SERVICES

MAX PLANCK COMPUTING & DATA FACILITY

- datashare.mpcdf.mpg.de (ownCloud)
  - Widely used sync&share service
  - Slow with large (100GB+) or many (100K+) files
- S3 compatible object storage (Ceph RadosGW)
  - objectstore.hpccloud.mpcdf.mpg.de
  - s3.nexus.mpcdf.mpg.de
  - “compromise” between DataShare and POSIX filesystem
  - More scalable and performant, still HTTP based and accessible from the internet
  - Powerful API that is supported by many third-party clients and tools
  - Unfortunatley no nice UI
- Nexus POSIX (IBM Storage Scale / GPFS)
  - More in BoF
- Experimental Nexus POSIX based on CephFS

- HPC-Cloud Ceph Cluster
  - RBD Images (OpenStack Glance)
  - RBD Volumes (OpenStack Cinder)
  - CephFS (OpenStack Manila)
    - *Generally mounted via NFS into VMs → not the fastest*
    - *Native CephFS mounts are being evaluated → some security and stability considerations → supported OS and/or ceph client versions may be limited*

# NEXUS CEPH HARDWARE

MAX PLANCK COMPUTING & DATA FACILITY

- 9 Storage Servers
  - 84x 16TB HDD
  - 3x 8TB NVMe
  - 72 CPU cores, 512GB RAM
- 2 Proxy Servers
  - 72 CPU cores, 256GB RAM
- Total Storage
  - 11 PiB brutto
  - 6.3 PiB with 4+3 Erasure Coding
  - 3.6 PiB with 3x Replication
- Separate ceph cluster for HPC cloud block storage

- Erasure Coding
  - RAID like redundancy implemented in software
  - 4+3 EC: data is split into 4 chunks, from which 3 additional parity chunks are calculated. Each chunk is stored on a separate machine.
  - Wider EC (with more data chunks) possible for increased efficiency with more machines.
  - Increased overhead for small objects and/or IO
- Replication
  - 3 copies of the entire object are stored on different machines
  - Faster with small objects, but high overhead → cost

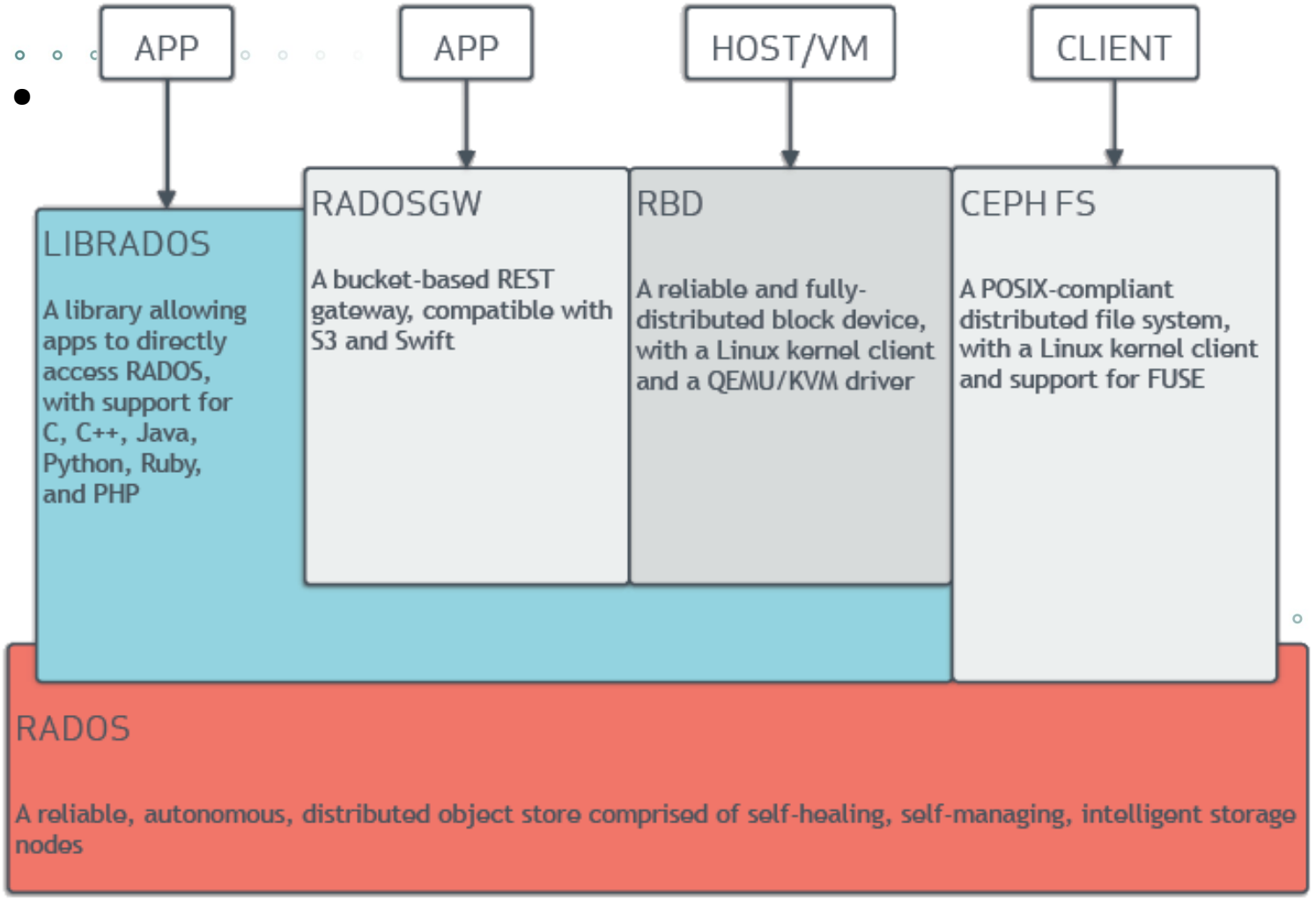
# NEXUS CEPH HARDWARE

MAX PLANCK
COMPUTING & DATA FACILITY

MPCDF

# CEPH ARCHITECTURE

MAX PLANCK
COMPUTING & DATA FACILITY



# NEXUS POSIX (BOF?)

MAX PLANCK COMPUTING & DATA FACILITY

MPCDF

- Some comments and background on Nexus POSIX based on yesterday's talks and questions:
- Parallel HPC filesystem, natively mounted on 1000+ HPC nodes
  - All have root access to the entire filesystem: security considerations
  - Distributed locking for performance: Deadlocks can (seldomly) happen
  - Strict inode limits due to TSM backups and to limit impact of accidental abuse
- Individual project shares mounted via NFS on HPC cloud
  - NFS exports generally configured root-squash to limit accidental "rm -rf /"
  - Root can still impersonate users
- New IBM ESS storage put into operation this summer
  - 20PB HDD, 400TB NVMe
  - NVMe capacity offers the possibility of creating filesystems/filesets with less strict inode limits / smaller average file size (e.g. PIROL)
  - TSM backups (and restores!) still a concern if needed
  - Details and pricing still to be hashed out for non-trivial projects

# DISCUSSION

MAX PLANCK COMPUTING & DATA FACILITY

- Things we are considering
  - OwnCloud Infinite Scale (OCIS), CERNbox (Reva):
    Unified POSIX like FS, WebDAV based sync&share and web interface
  - Generic CephFS
    - *as an alternative backend for Nexus POSIX*
    - *Ideally integrated with OCIS*
  - Some Nexus POSIX like filesystem with less strict inode limits
- Frequent blocker: scalable backups to tape
  - Alternative backups and/or async replication to disk
  - Still needs an efficient way to discover what to backup (~snapdiff)
- What are experiences with our object storage?
  Any features that are you missing?