dCache.org

dCache - delegated storage solutions

Tigran Mkrtchyan for dCache Team

ISGC 2016, Taiwan

Fermilab

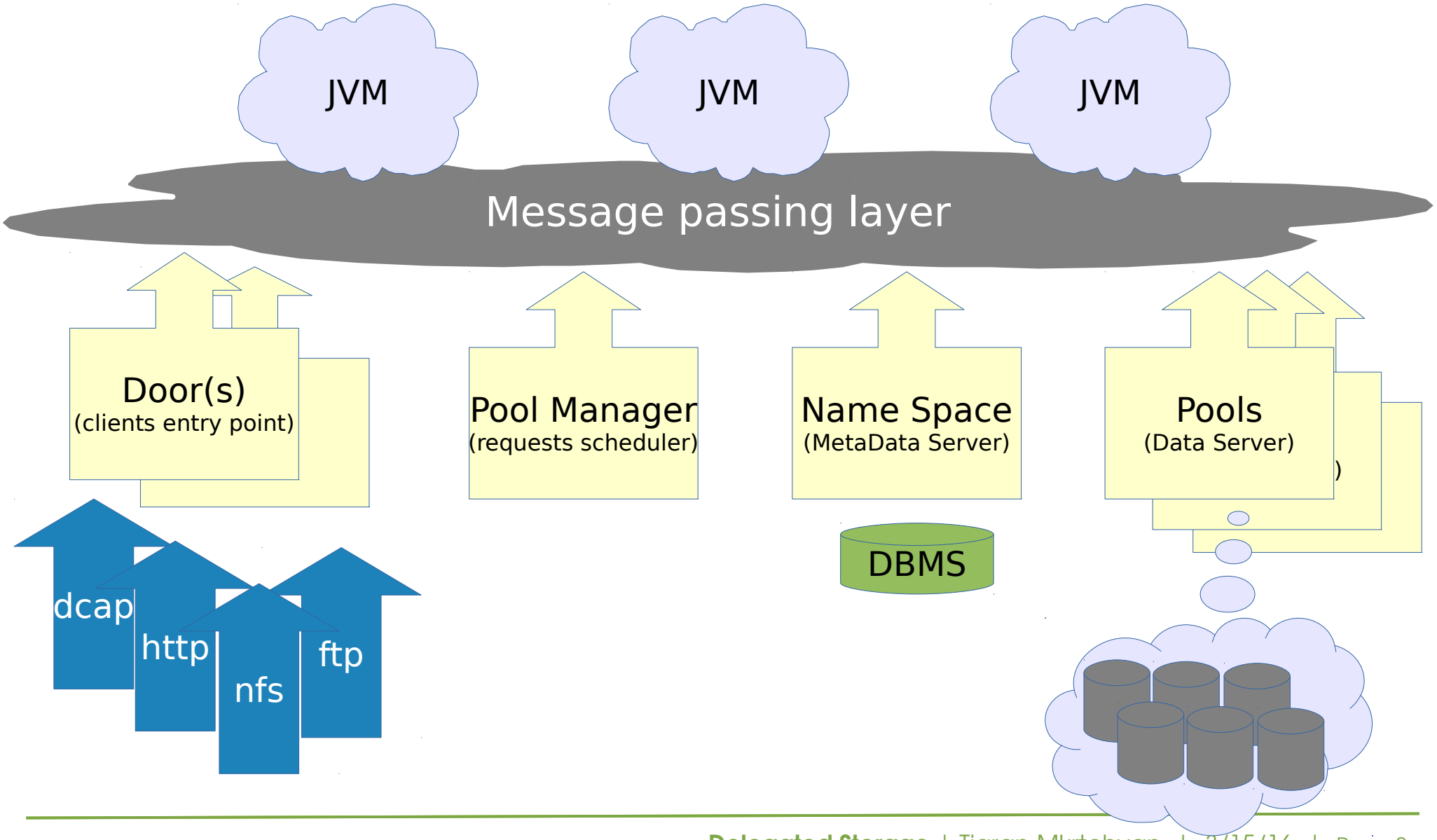
NDGF
Nordic DataGrid Facility

INDIGO - DataCloud

DESY

HELMHOLTZ
ASSOCIATION

dCache on one slide



Usage around the World

- ~ 80 installations
- > 50% of WLCG storage
- biggest 22 PB
- Typical ~100x nodes
- Typical ~ 10^7 files

dCache as Storage System

- Provides a single-rooted namespace.
- Metadata (namespace) and data locations are independent.
- Aggregates multipe storage nodes into a single storage system.
- Manages data movement, replication, integrity.
- Provides data migration between multiple tiers of storage (DISK, SSD, TAPE).
- Uniquely handles different Authentication mechanisms, like x509, Kerberos, login+password, auth tokens.
- Provides access to the data via variety of access protocols (WebDAV, NFSv4.1/pNFS, xxxFTP. DCAP, Xrootd, DCAP).

dCache as Storage System

- Provides a single-rooted namespace.
- Metadata (namespace) and data locations are independent.
- Aggregates multipe storage nodes into a single storage system.
- Manages data movement, replication, integrity.
- Provides data migration between multiple tiers of storage (DISK, SSD, TAPE).
- Uniquely handles different Authentication mechanisms, like x509, Kerberos, login+password, auth tokens.
- Provides access to the data via variety of access protocols (WebDAV, NFSv4.1/pNFS, xxxFTP. DCAP, Xrootd, DCAP).

dCache's data management

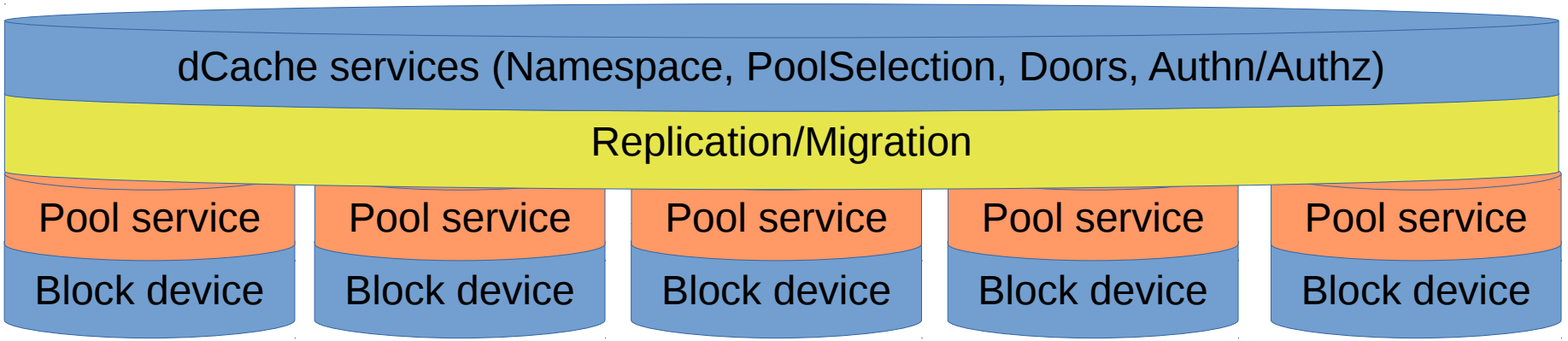
- Automatic migration
  - Tape/disk/disk
  - HotSpot detection
  - Permanent migration jobs
  - Checksumming on transfer
- Manual migration
- Data replication
  - multiple copies
  - same host/rack/site policy

Software-defined storage (or did you listen Patrick carefully?)

- Abstraction of logical storage services and capabilities from the underlying physical storage systems
- Automation with policy-driven storage provisioning with service-level agreements replacing technology details.
- Commodity hardware with storage logic abstracted into a software layer.

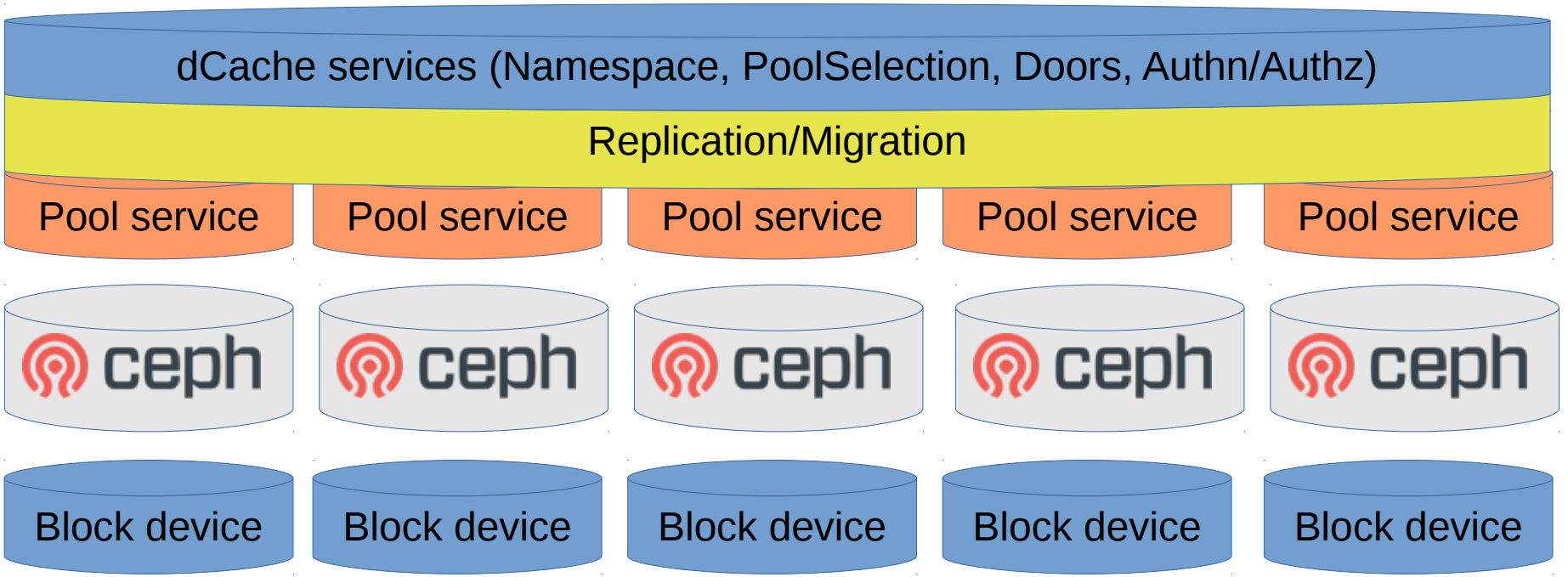
Storage in dCache (what we have)

- dCache provides high level service
- Data replication and management core dCache service
- Each pool attached to own disks



Storage in dCache (outsourcing, phase 1)

- dCache provides high level service
- Data replication and management core dCache service
- Each pool has it own 'partition' on shared storage
- Each 'partition' attached to it's own block device

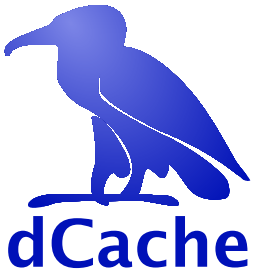


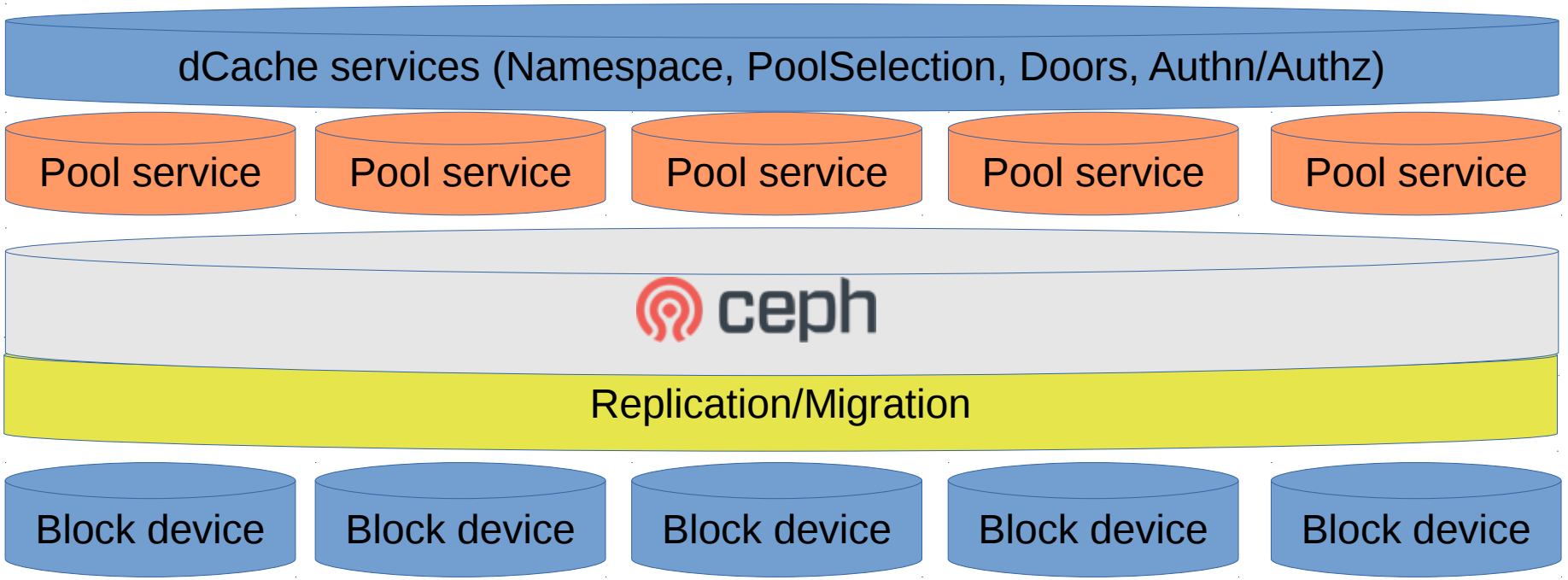
Phase 1 (changing IO layer)

- Single data server owns the data
- Single data server manages data
  - flush to tape
  - restore from tape
  - removal
  - garbage collection

Storage in dCache (outsourcing, phase 2)

- dCache provides high level service
- All pool see all 'partition' on shared storage
- Any pool can deliver data from any partition
- Object store takes care about replication



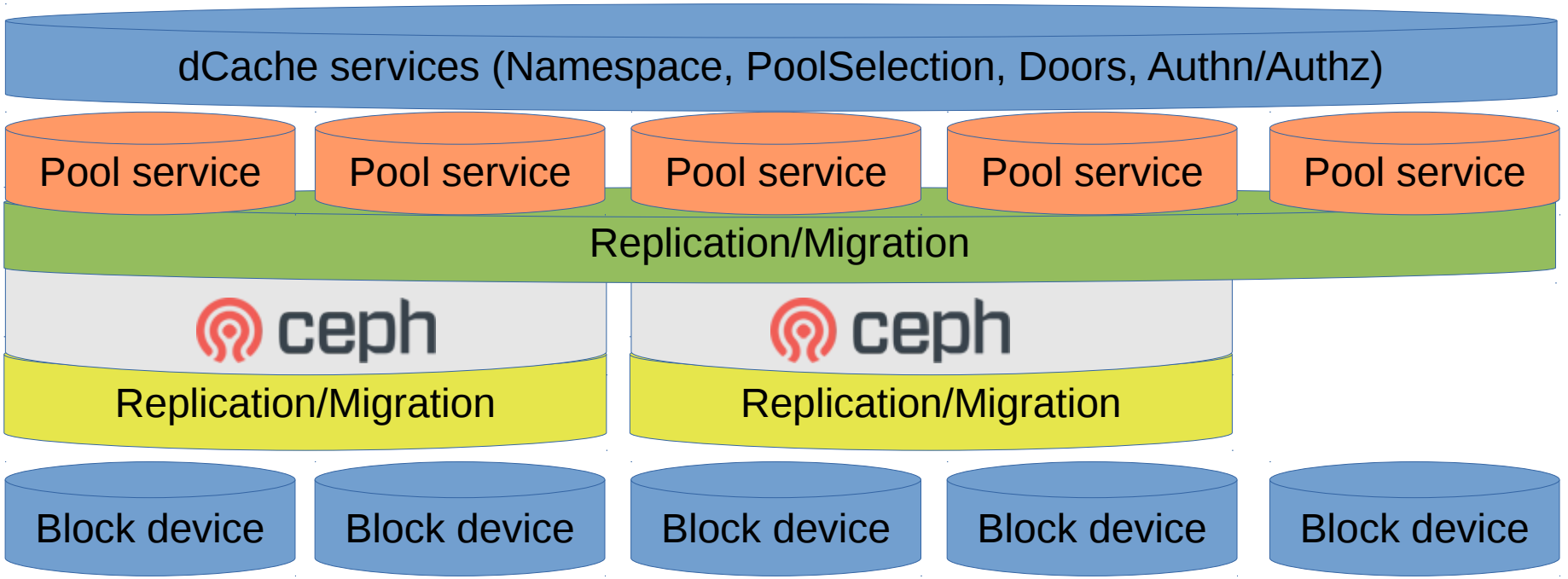


Phase 2 (Changing core philosophy)

- All data managed by 'quorum'
  - group decision who interact with tape
  - group decision who/when file is removed
  - File location is always 'known'

Storage in dCache (outsourcing, phase 3)

- dCache provides high level service
- dCache can move data between regular and OS pools



Phase 3 (mixed environment)

- Mixed setup
- Islands of storage servers
- Replication and data movement between islands

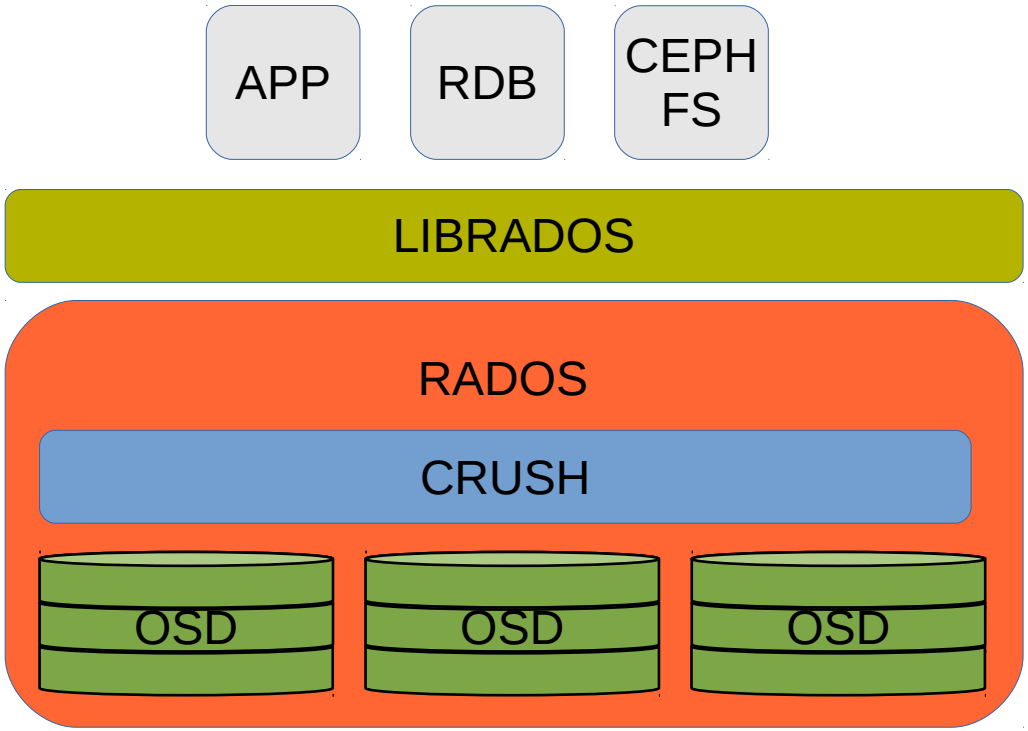
Why CEPH

- No specific hardware support
- Runs on commodity hardware
- Scalable to exabytes of data
- Deployed at sites as storage system for OpenStack
- Provides Object, Block and File interfaces

And not only CEPH

- Other object store can be adopted
  - DDN WOS
- Swift/S3/CDMI
- Cluster file systems (as a side effect)
  - Luster
  - GPFS
  - GlusterFS

CEPH (extremely simplified)



- OSD ~ a physical disk
- CRUSH - determines how to store and retrieve data by computing data storage locations.
- RADOS - distributes objects across the storage cluster and replicates objects
- librados - provides low-level access to the RADOS service.

Current work

- Functional prototype only
- Focus on stability first
- RBD based
  - striping
  - alterable content
- Object interface will be evaluated as well

Roadmap

- Phase 1
  - running prototype is available today
    - some sites volunteer to help with testing
  - cleaning up to make generally available
- Phase 2/3
  - depends on user demand
  - operational overhead, if any
  - support overhead, if any

Summary

- dCache is demanded storage system.
- New technology provides required building blocks.
- Combination on both makes us to concentrate on missing parts.
- Working prototype available for testing.

Links

- https://www.dcache.org/
- https://en.wikipedia.org/wiki/Software-defined_storage
- http://ceph.com/