synnefo

grnet

FLEXIBLE STORAGE FOR HPC CLOUDS WITH ARCHIPELAGO AND CEPH.

VANGELIS KOUKIS
TECHNICAL LEAD, SYNNEFO

synnefo

grnet

Outline

vkoukis@grnet.gr

Archipelago overview

Storage resources, clones/snapshots

HPC workflow with Archipelago

Resource Composition

Archipelago Implementation

Flexible I/O pipelines

Integration with Synnefo

Future directions

synnefo

grnet

Archipelago overview

vkoukis@grnet.gr

Distributed Storage System

- Powering storage in clouds

Decouples storage **resources** from storage **backends**

- Files / Images / Volumes / Snapshots

Unified way to provision, handle, and present resources

Decouples **logic** from actual physical **storage**

- Software-Defined Storage

synnefo

grnet

Archipelago logic

vkoukis@grnet.gr

Thin provisioning, with **clones** and **snapshots**

- Independent from the underlying storage technology

Hash-based data deduplication

Pluggable architecture

- Multiple endpoint (northbound) drivers
- Multiple backend (southbound) drivers

Multiple storage backends

- Unified management
- with storage migrations

synnefo

grnet

Unified view of resources

VHPC'13
vkoukis@grnet.gr

Files

- User files, with Dropbox-like syncing

Images

- Templates for VM creation

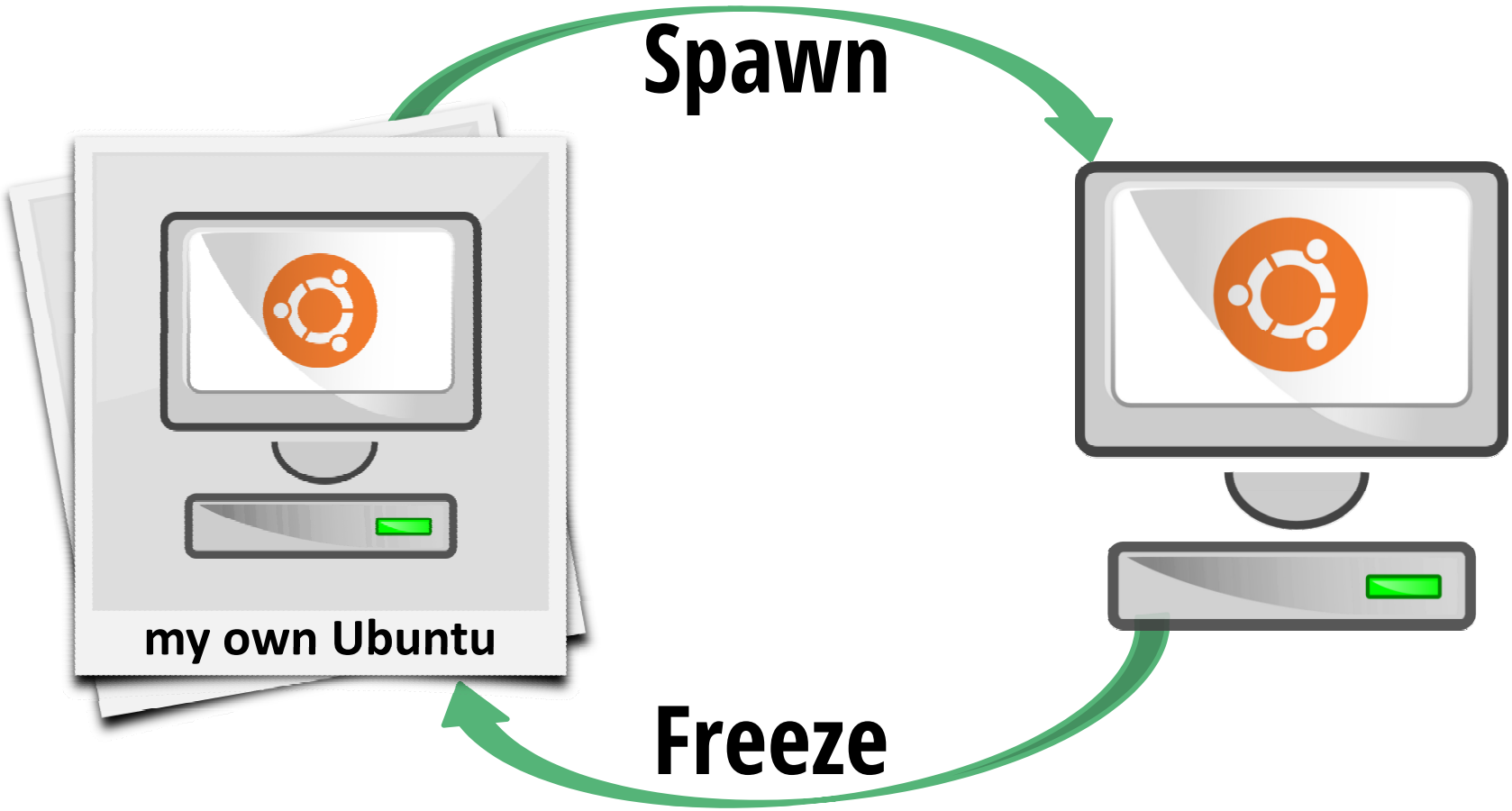
Volumes

- Live disks, as seen from VMs

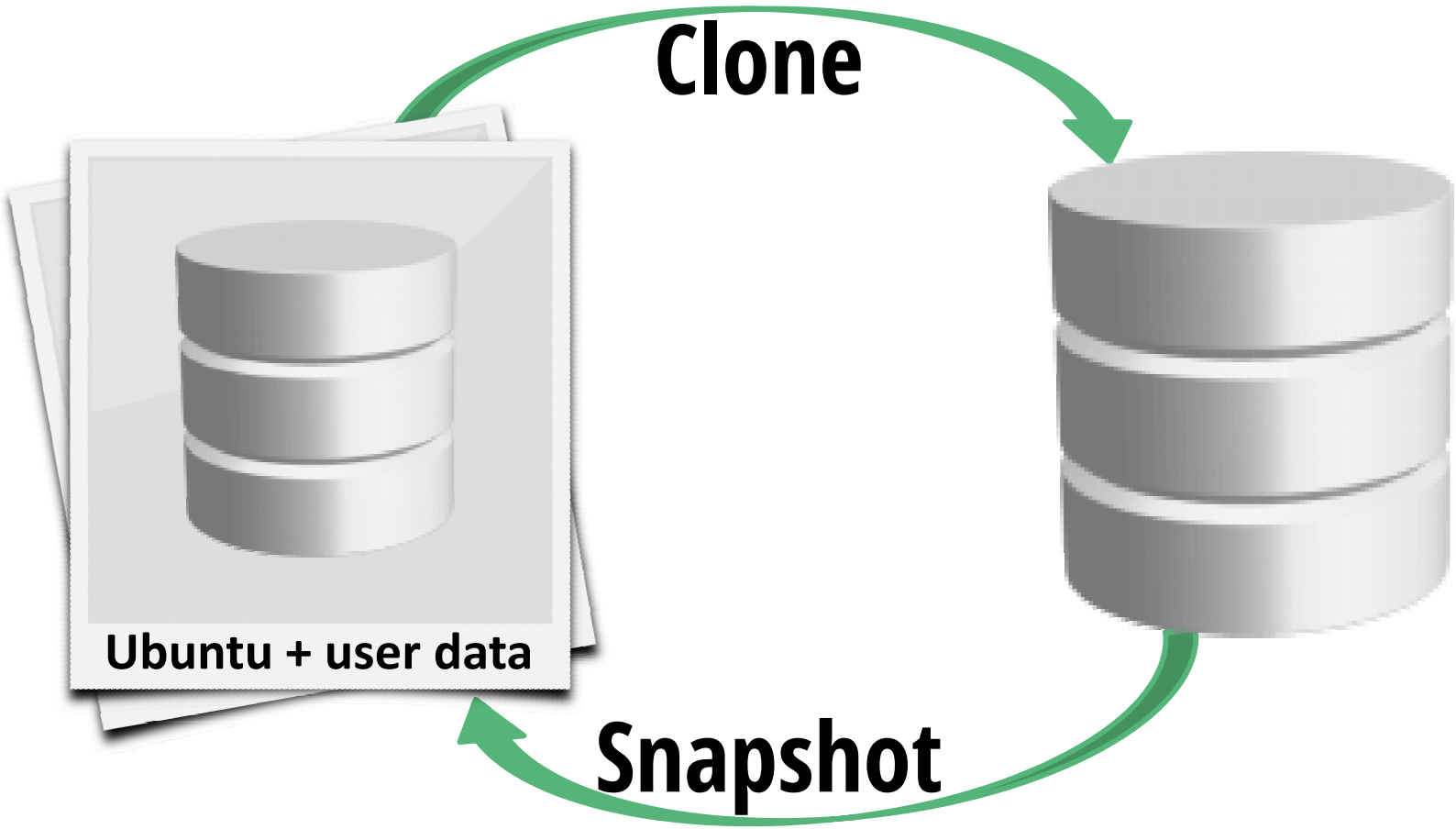
Snapshots

- Point-in-time snapshots of Volumes

VHPC'13
vkoukis@grnet.gr



VHPC'13
vkoukis@grnet.gr



synnefo

grnet

The big picture

vkoukis@grnet.gr

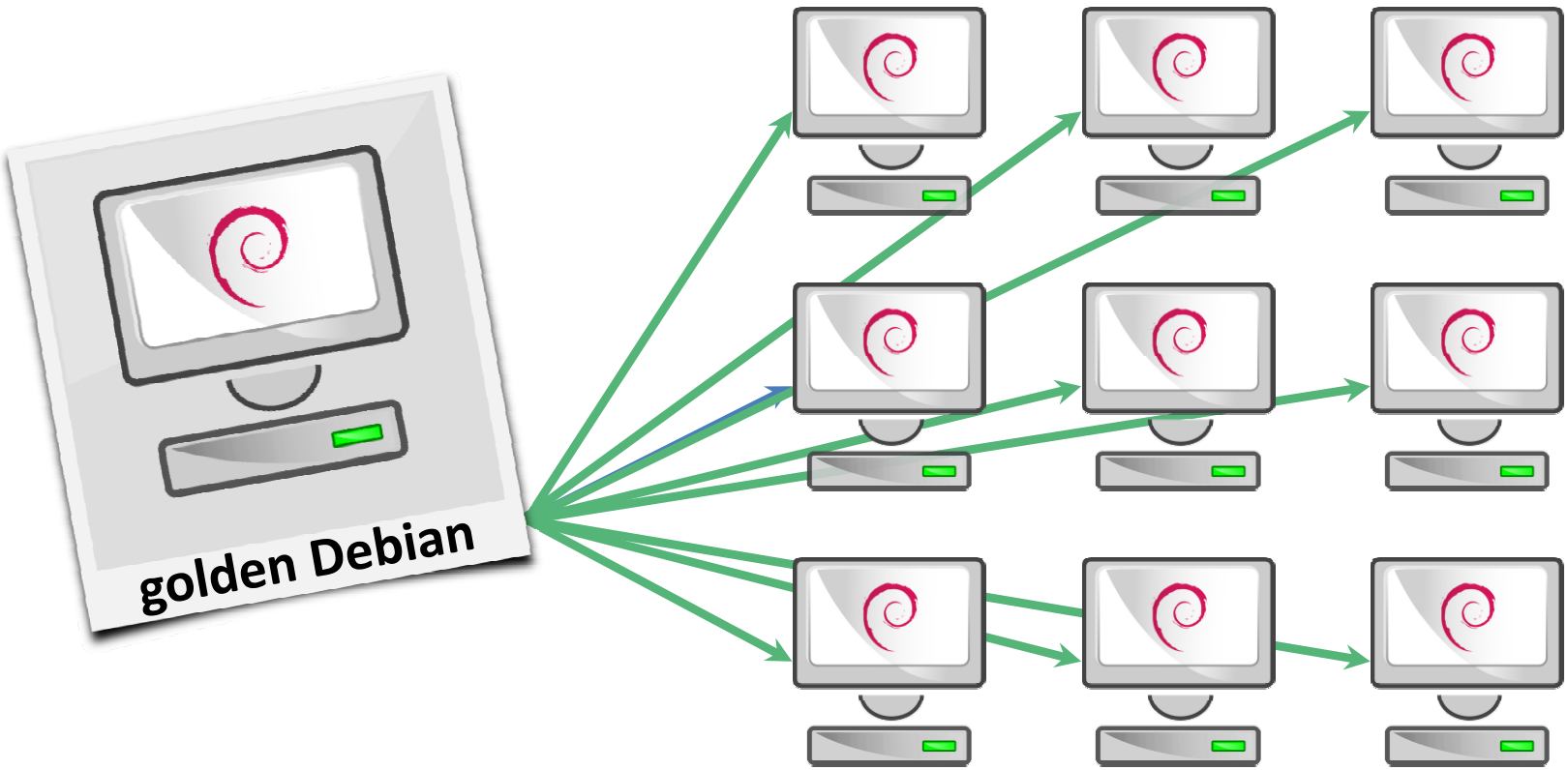
Archipelago Core

Storage backend 1
(e.g., Ceph cluster 1)

Storage backend 2
(e.g., Ceph cluster 2)

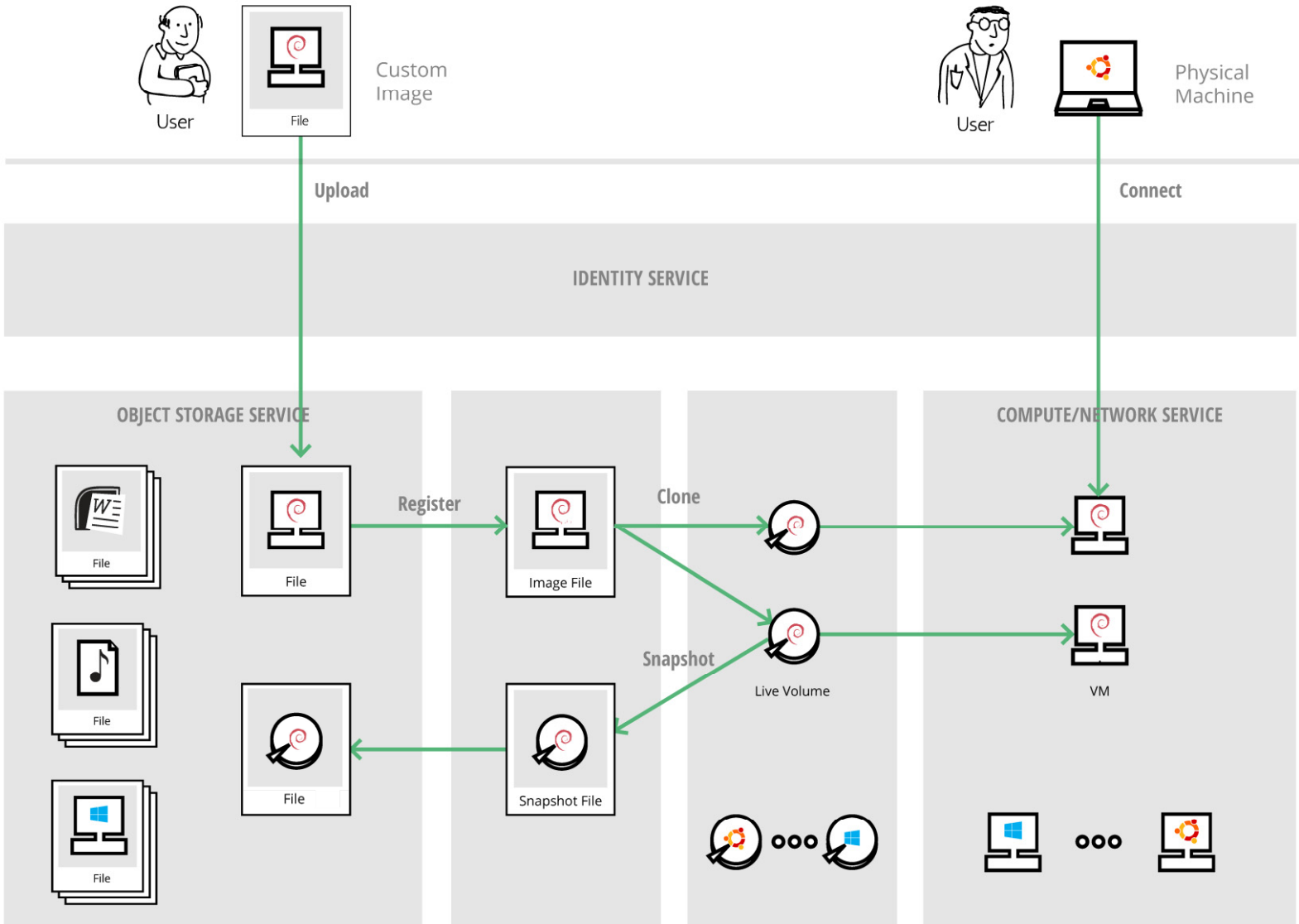
Storage backend 3
(e.g., NFS over NAS)

VHPC'13
vkoukis@grnet.gr



End-to-end workflow with unified storage

VHPC'13
vkoukis@grnet.gr

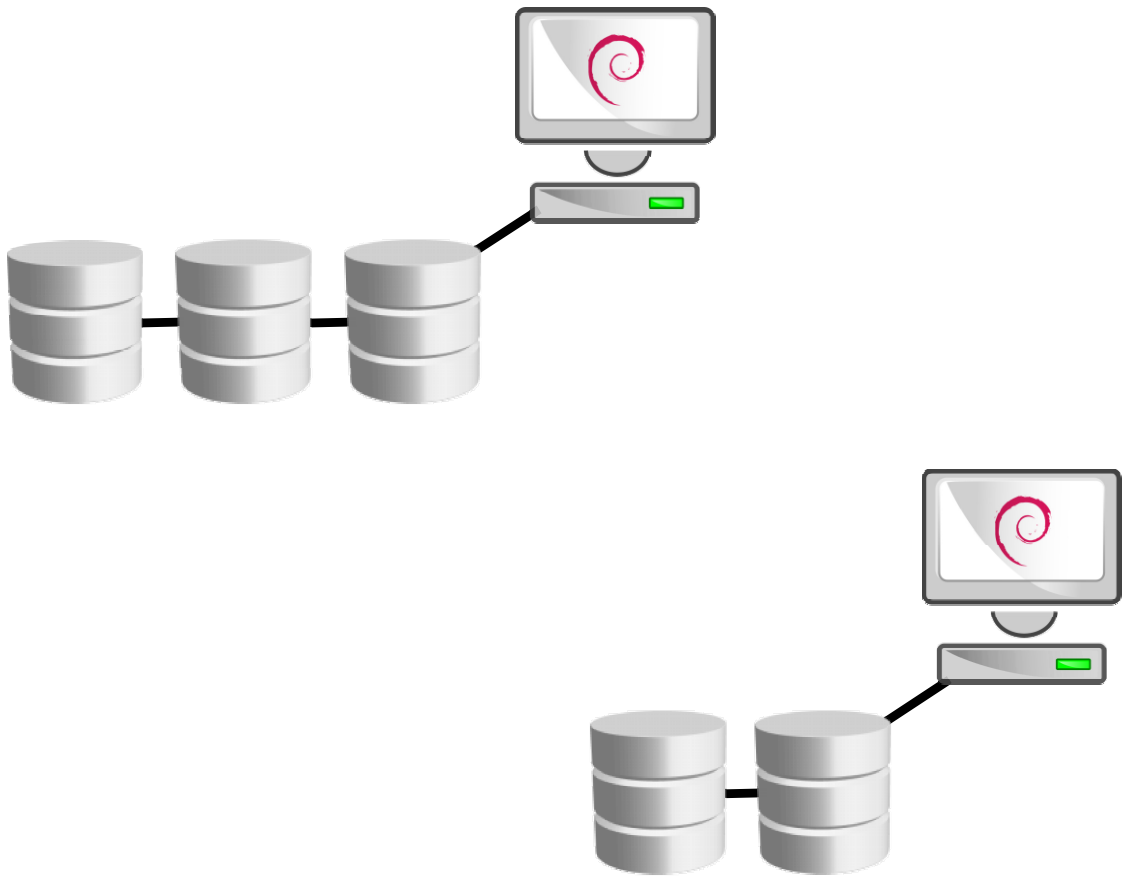


Live demo!

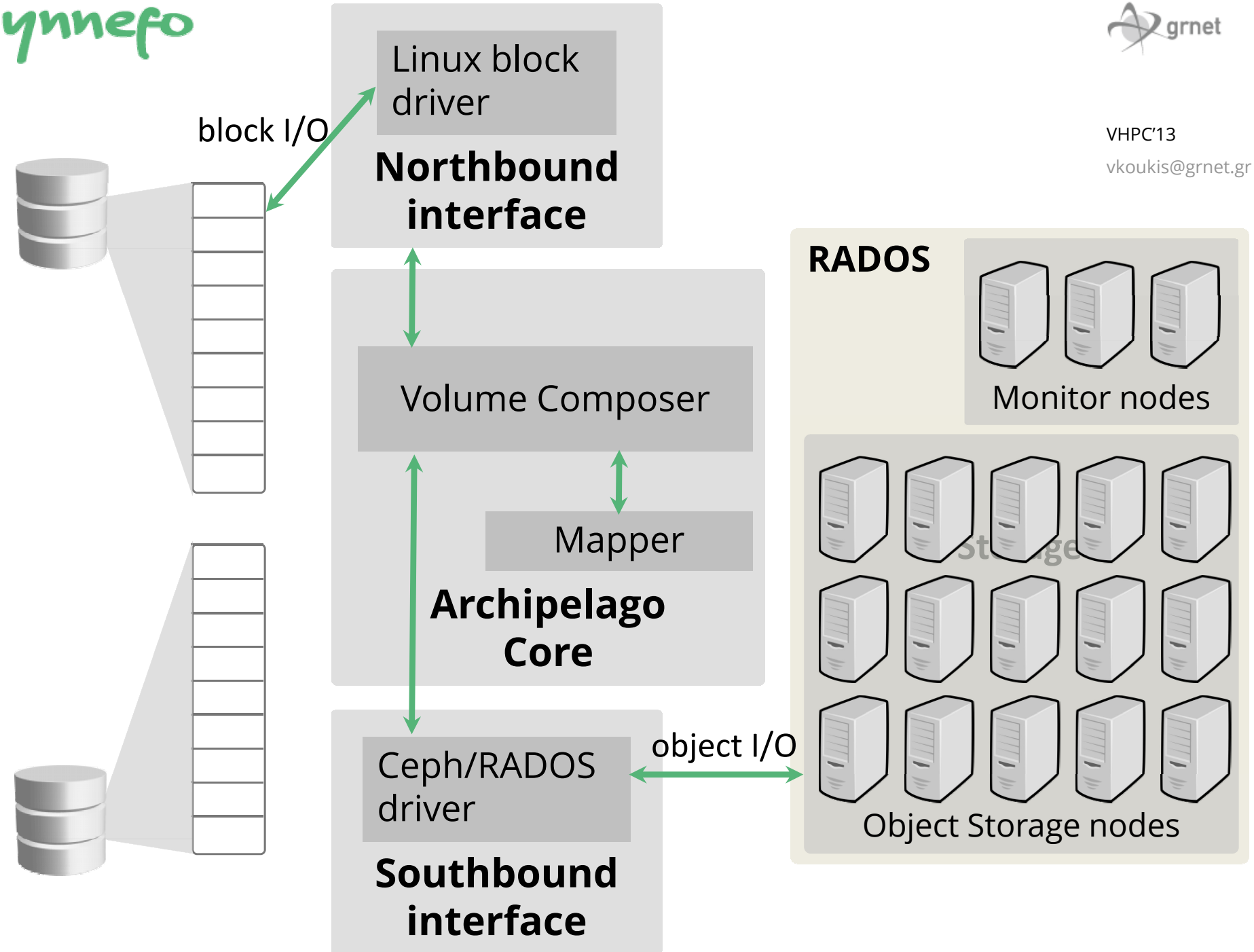
VHPC'13
vkoukis@grnet.gr

Login, view/upload files
Unified image store: Images as files
View/create/destroy servers from Images
...on multiple storage backends
...on Archipelago, for thin, super-fast creation
...with per-server customization, e.g., file injection
Take a point-in-time snapshot of a VM's disk, in seconds
Share it with collaborators, with fine-grained Access Control
Create a virtual cluster from this Snapshot
...from the command-line, and in Python scripts

VHPC'13
vkoukis@grnet.gr



VHPC'13
vkoukis@grnet.gr

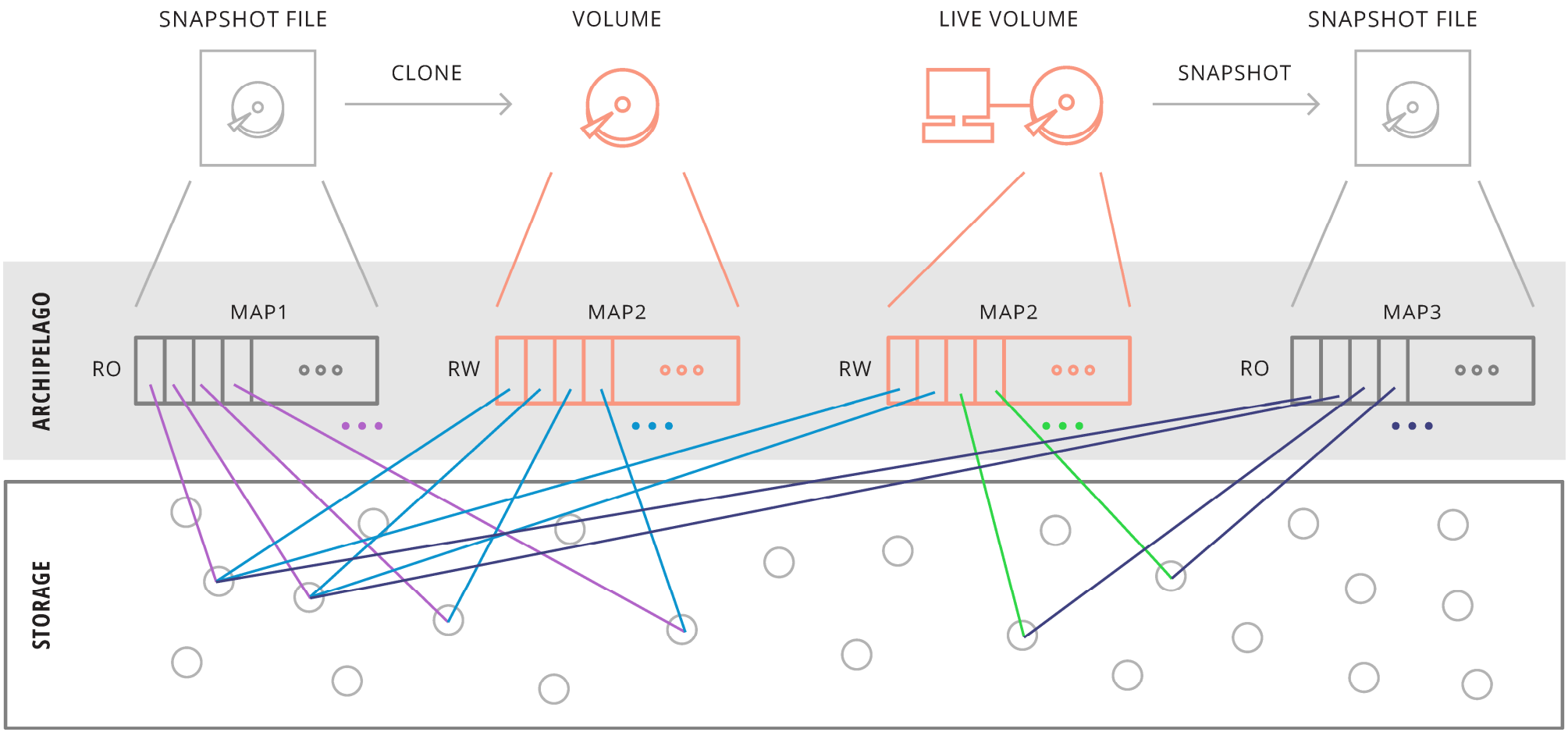


synnefo

grnet

Resource composition

VHPC'13
vkoukis@grnet.gr

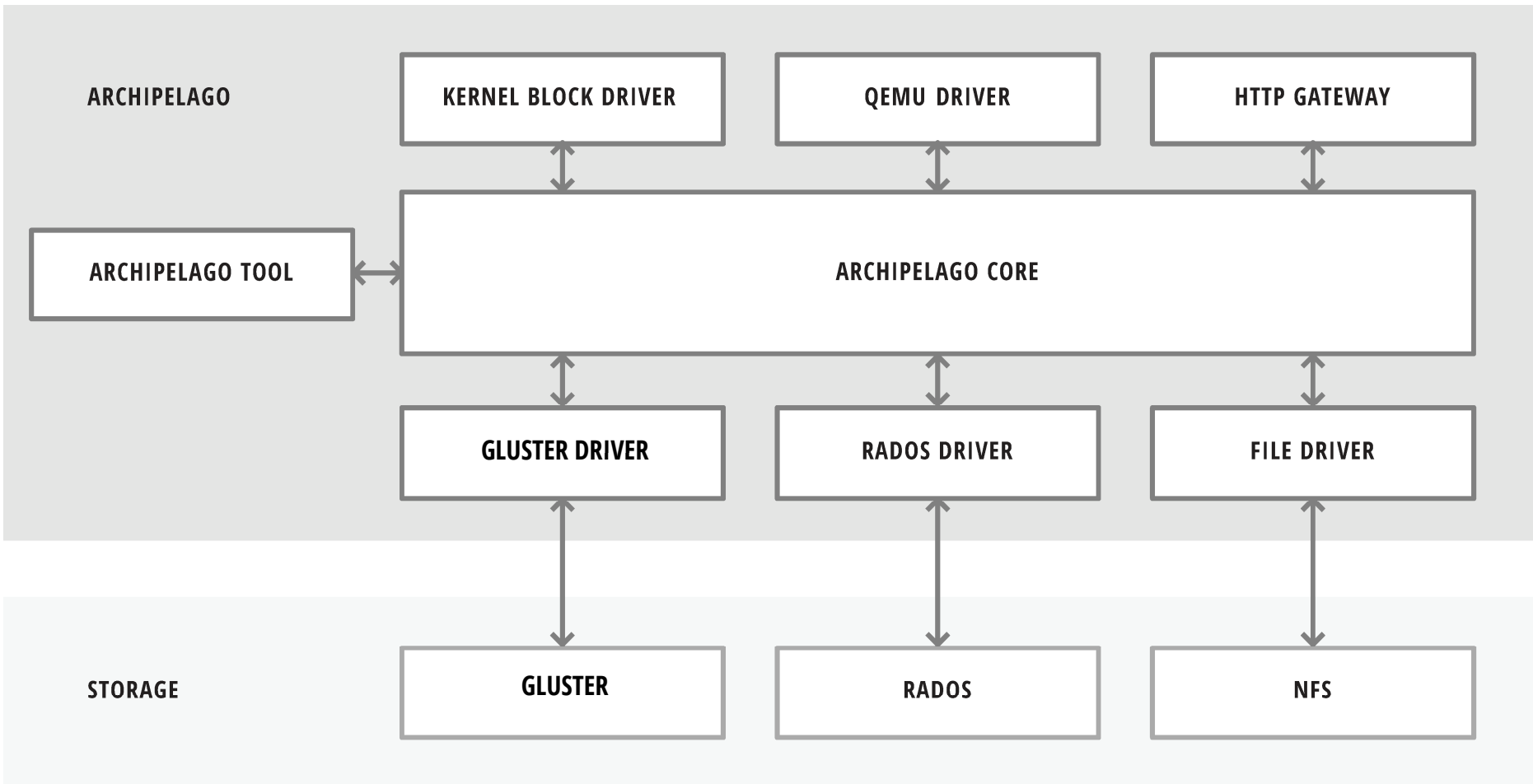


synnefo

grnet

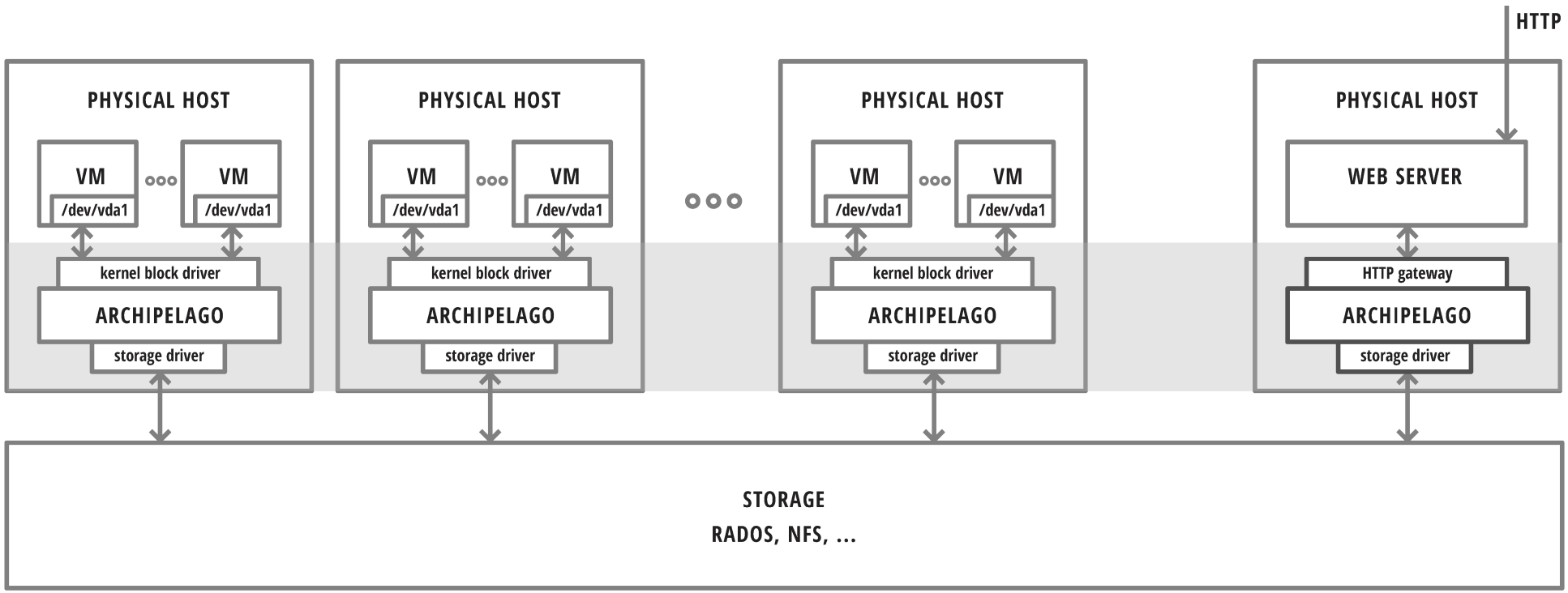
Archipelago interfaces

VHPC'13
vkoukis@grnet.gr



Running Archipelago

VHPC'13
vkoukis@grnet.gr

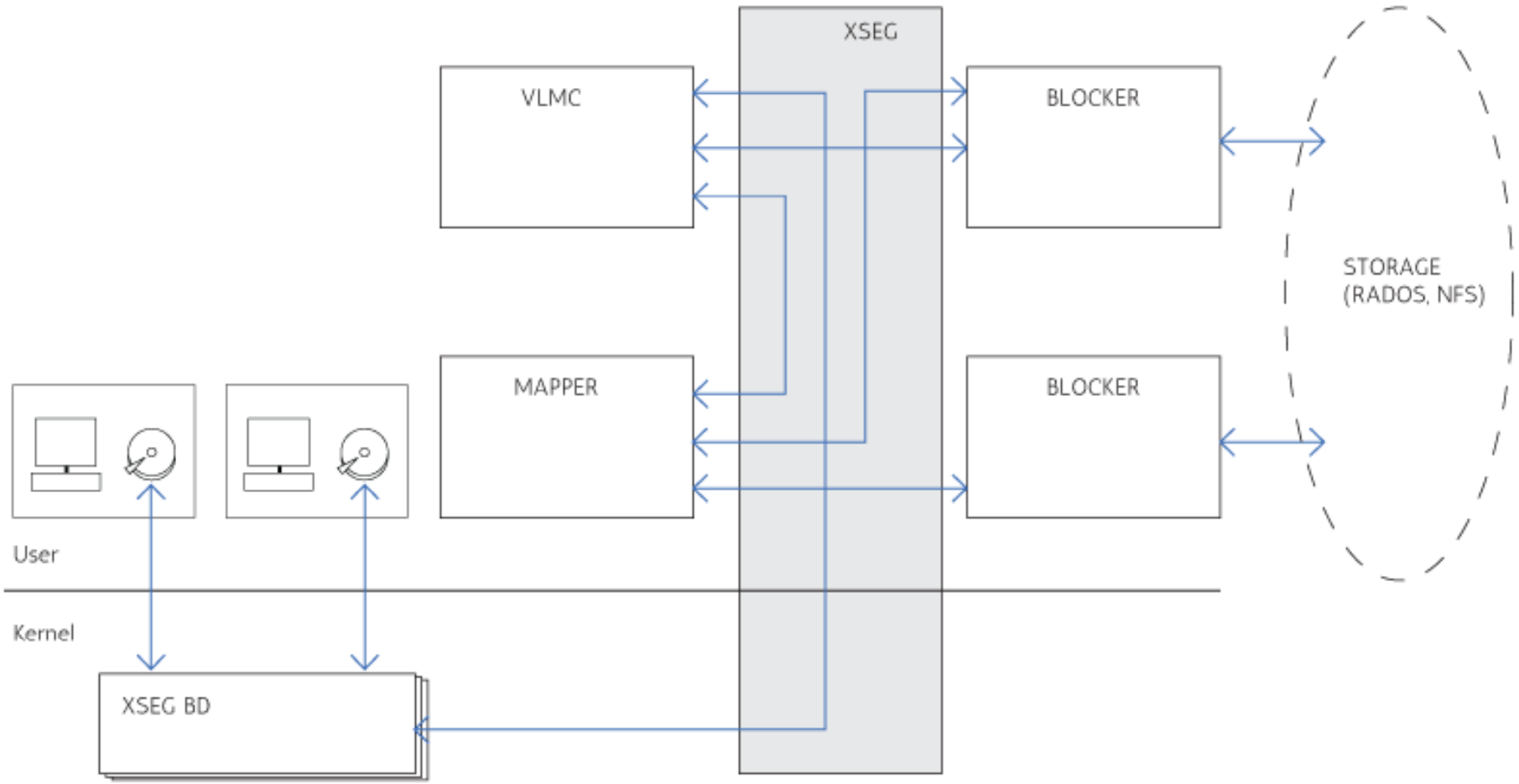


synnefo

grnet

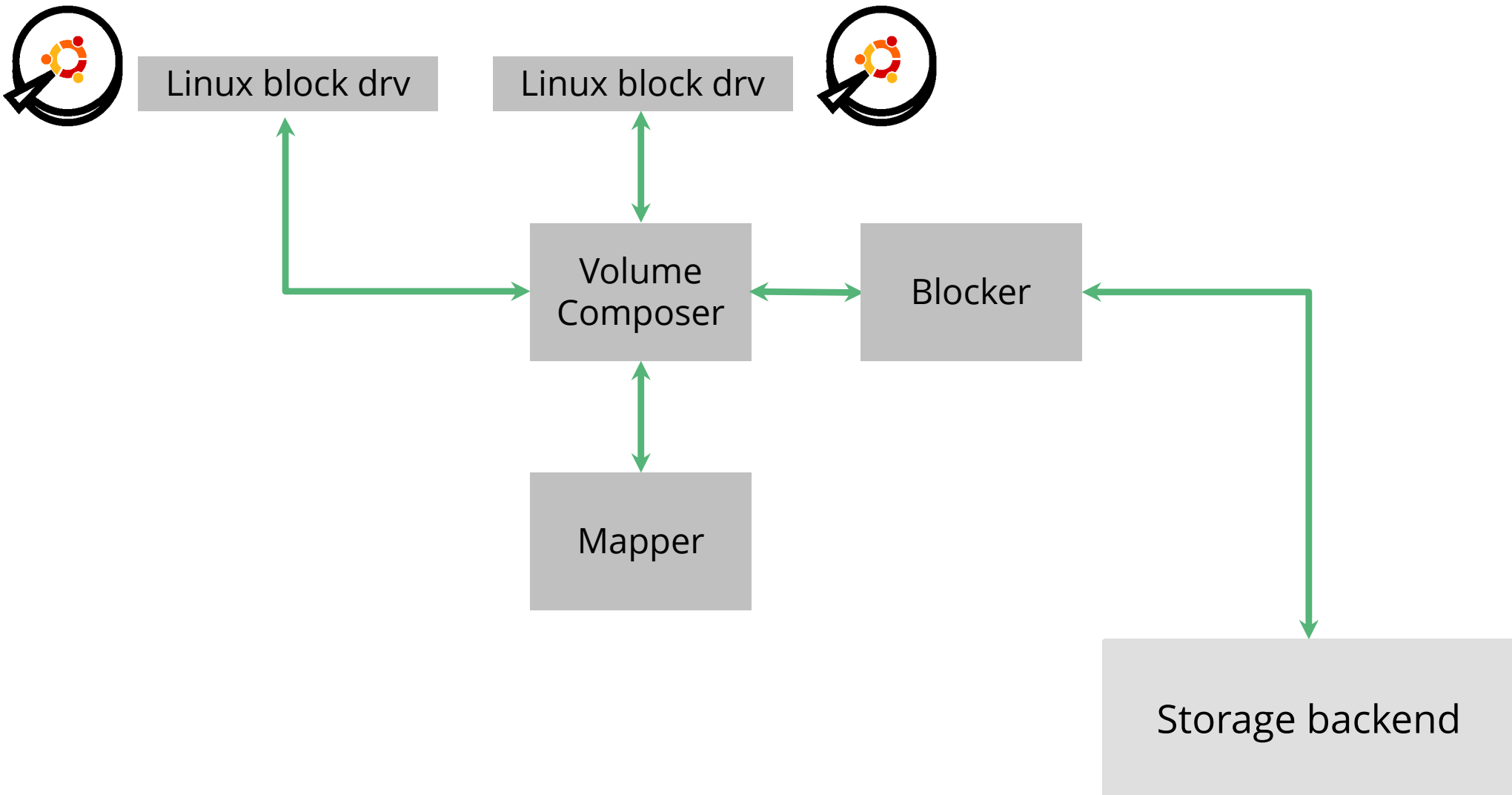
Implementation Internals

VHPC'13
vkoukis@grnet.gr



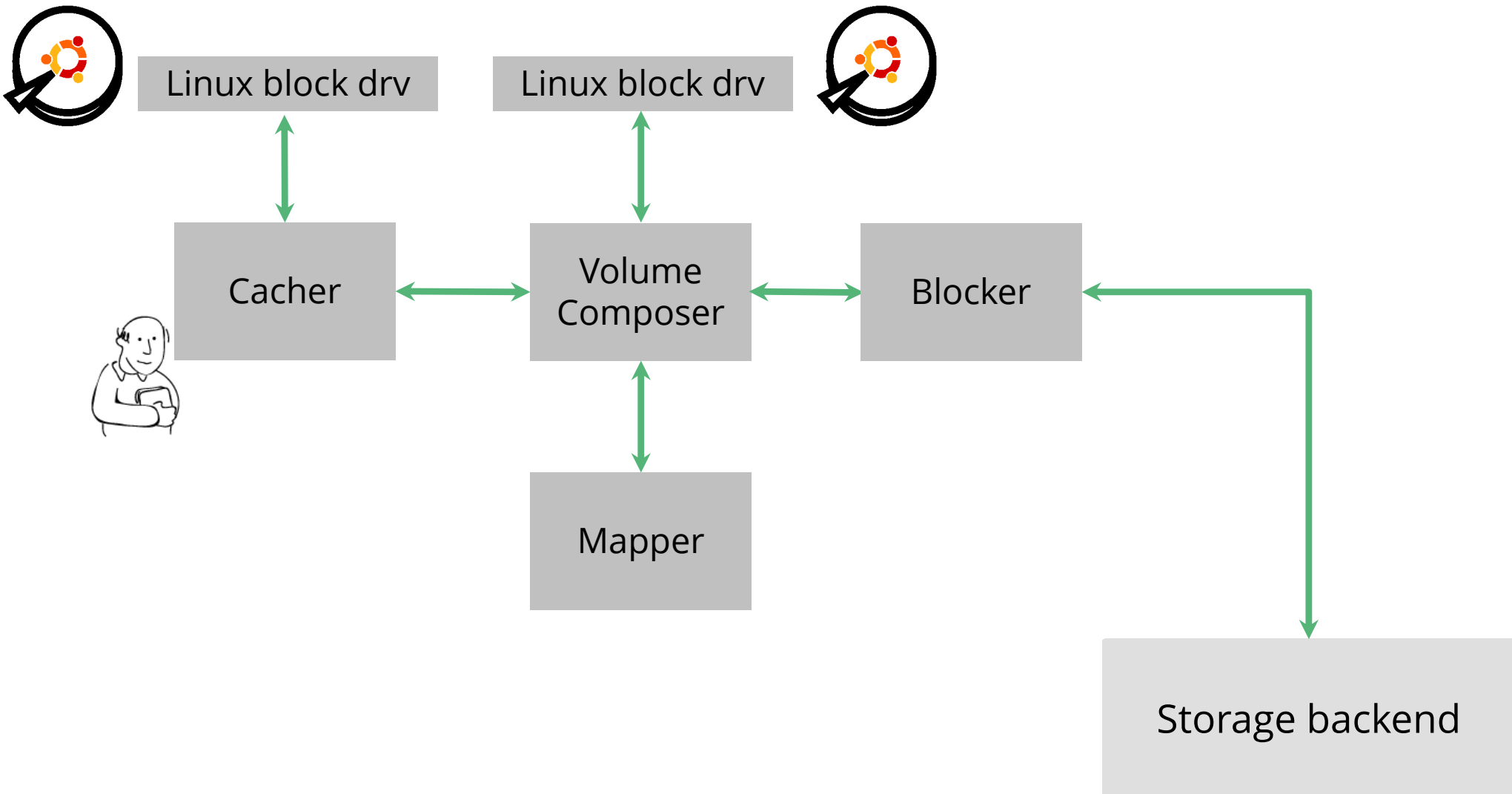
Flexible I/O pipeline

VHPC'13
vkoukis@grnet.gr



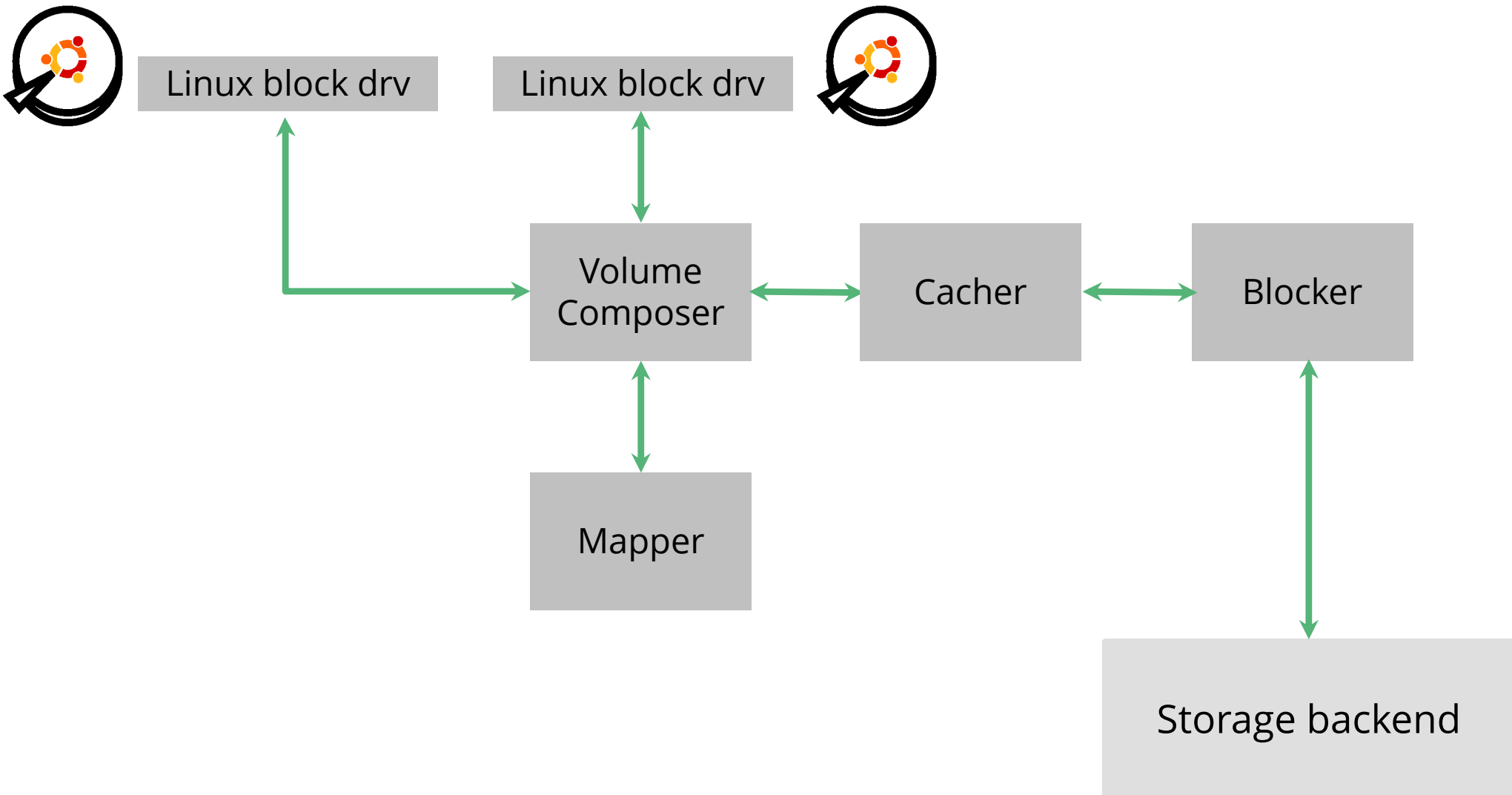
Flexible I/O pipeline

VHPC'13
vkoukis@grnet.gr



Flexible I/O pipeline

VHPC'13
vkoukis@grnet.gr



Archipelago overview

vkoukis@grnet.gr

Distributed Storage System

- Powering storage in clouds

Decouples storage **resources** from storage **backends**

- Files / Images / Volumes / Snapshots

Unified way to provision, handle, and present resources

Decouples **logic** from actual physical **storage**

- Software-Defined Storage

Archipelago logic

VHPC'13
vkoukis@grnet.gr

Thin provisioning, with **clones** and **snapshots**

- Independent from the underlying storage technology

Hash-based data deduplication

Pluggable architecture

- Multiple endpoint (northbound) drivers
- Multiple backend (southbound) drivers

Multiple storage backends

- Unified management
- with storage migrations

synnefo

grnet

Integration with Synnefo

vkoukis@grnet.gr

IaaS open source cloud software

Identity, Storage, Compute/Network/Image/Volume services

Production since June 2011

- Powering the ~okeanos public cloud

Written in Python/Django, exposes OpenStack APIs

Uses Google Ganeti at the backend

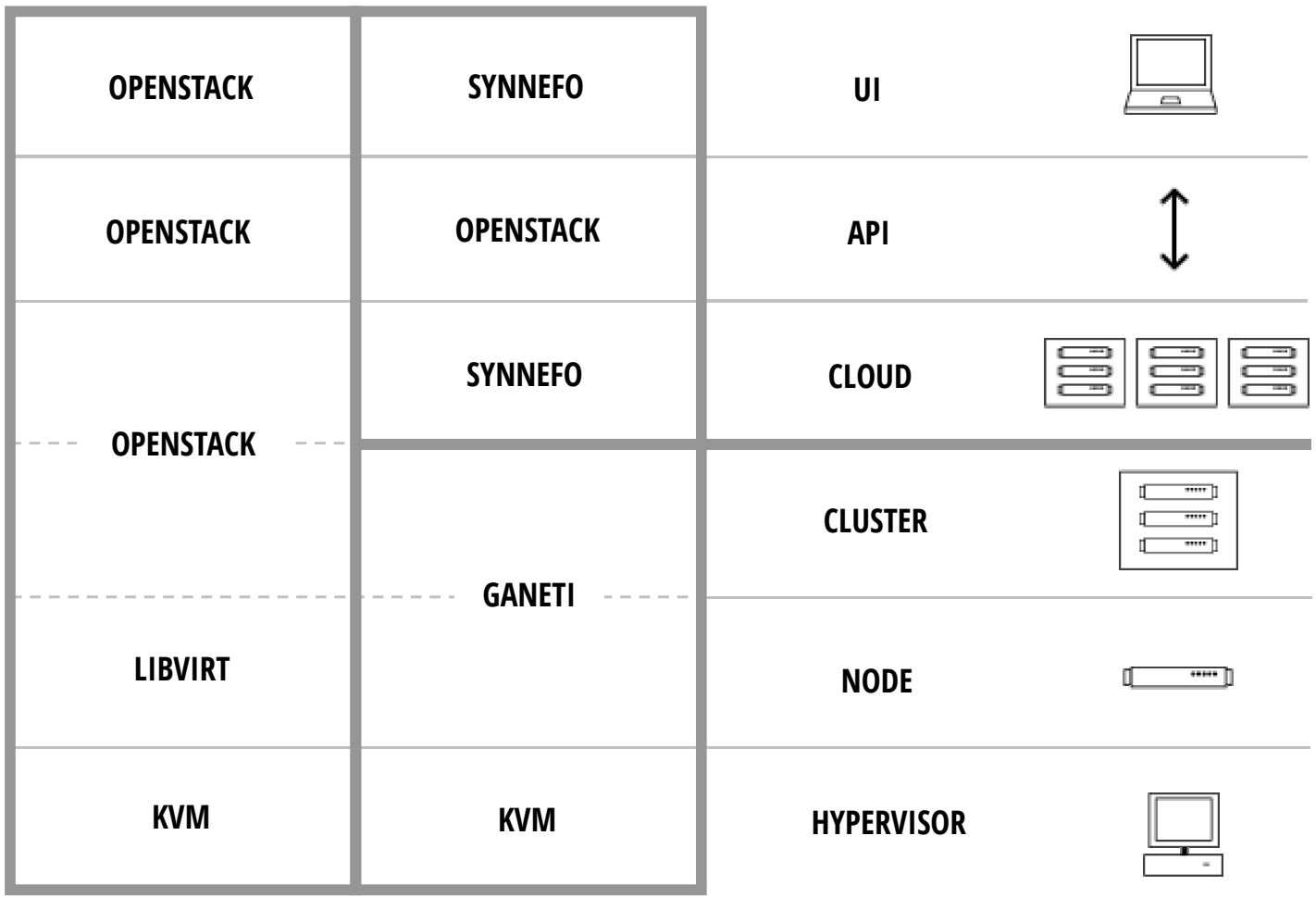
BSD licensed

synnefo

grnet

Cluster vs Cloud

VHPC'13
vkoukis@grnet.gr



Google Ganeti

VHPC'13
vkoukis@grnet.gr

Mature, production-ready VM cluster management

- used for Google's corporate infrastructure

Multiple storage backends out of the box

- LVM, DRBD
- Files on local or shared directory
- RBD (Ceph/RADOS)

External Storage Interface for SAN/NAS support

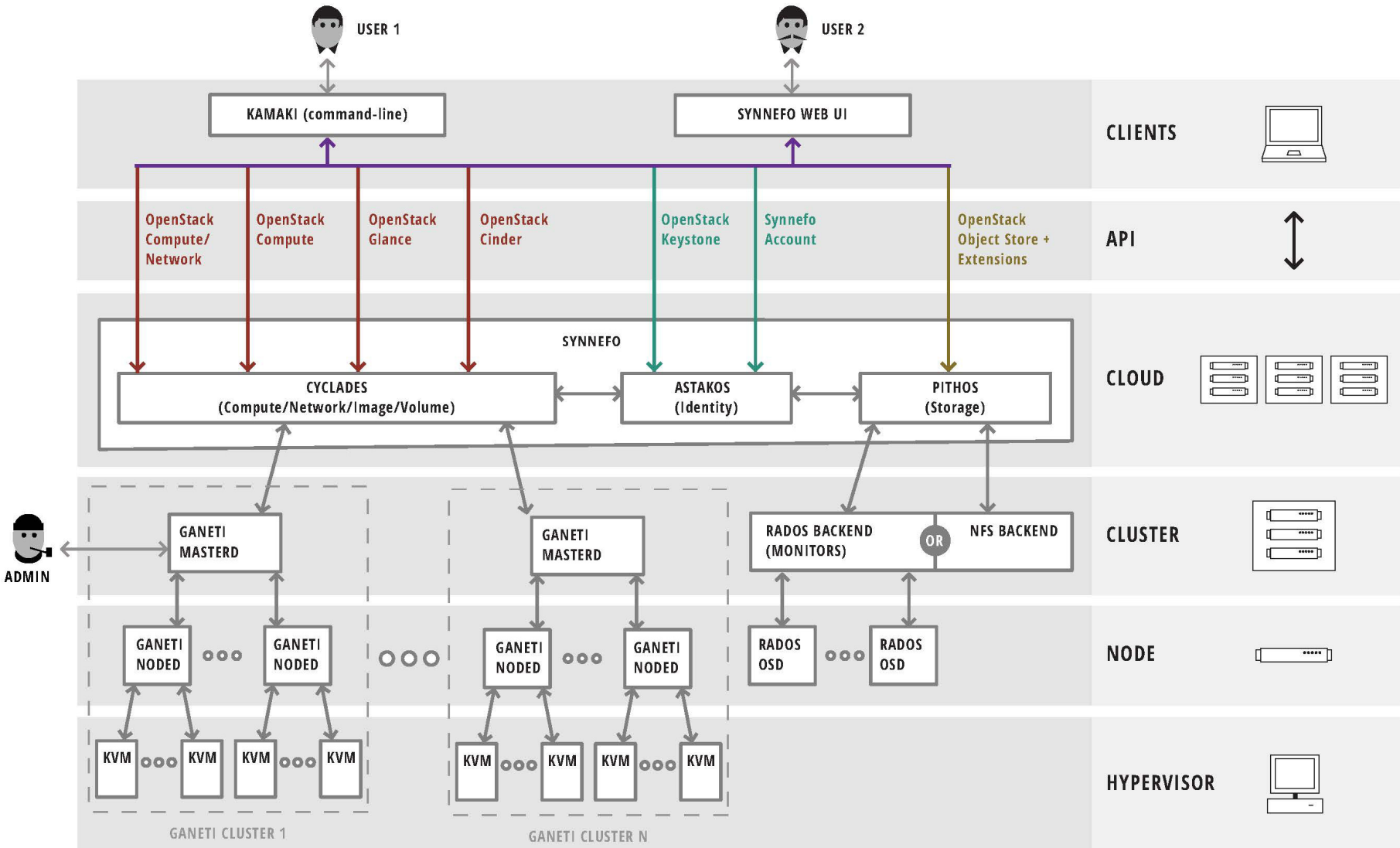
Ganeti cluster = *masterd* on master, *noded* on nodes

Easy to integrate into existing infrastructure

- Remote API over HTTP, pre/post hooks for every action!

Architecture

VHPC'13
vkoukis@grnet.gr



synnefo

grnet

Storage service: Pithos

vkoukis@grnet.gr

Exposes the OpenStack Object Storage (Swift) API

- plus extensions, for sharing and syncing

Rich sharing, with fine-grained Access Control Lists

Content-based addressing for blocks

Partial file transfers, deduplication, efficient syncing

Backed by Archipelago

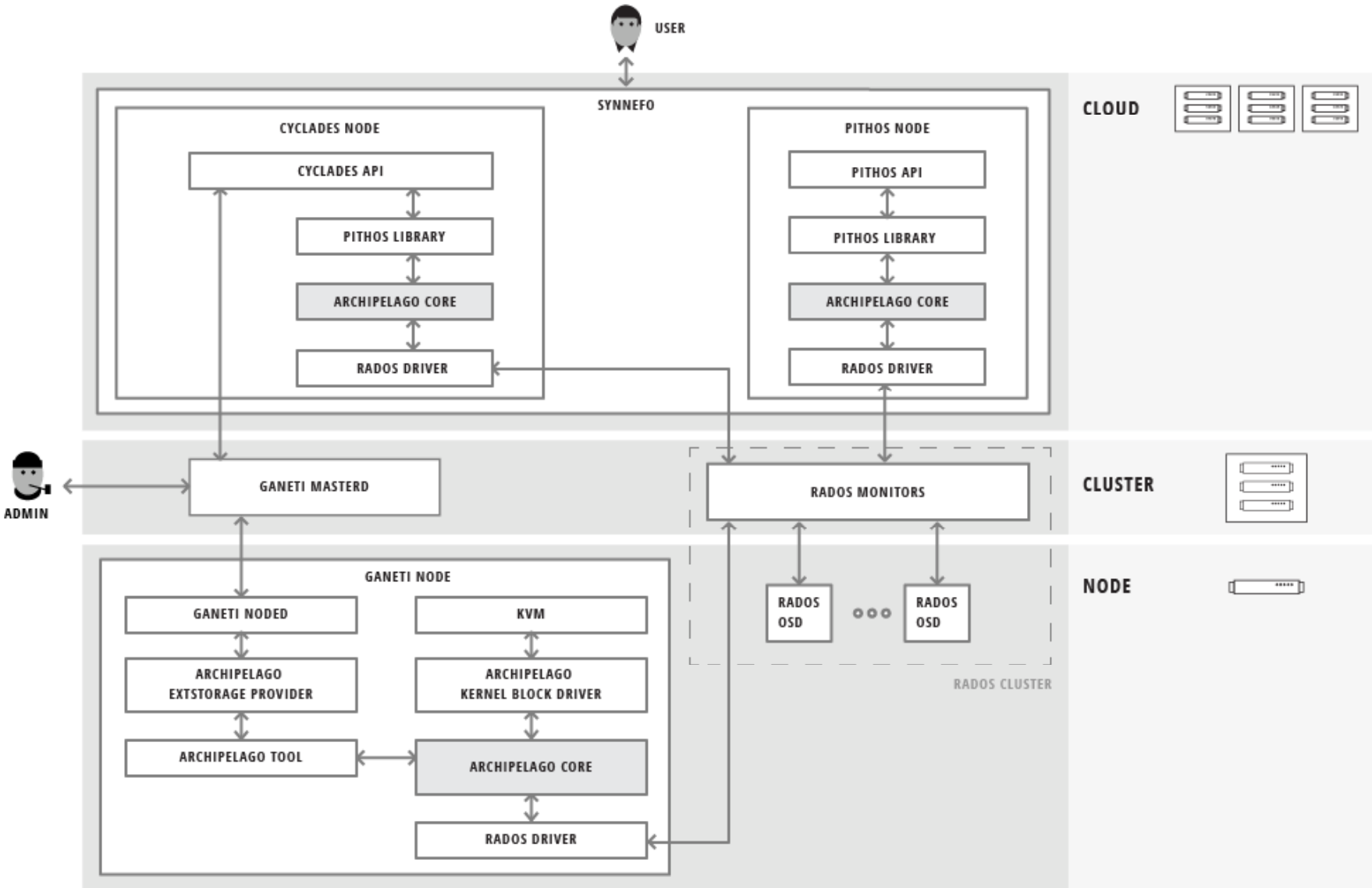
- Provides a northbound endpoint for Archipelago
- Implements the HTTP gateway
- Exposes the Swift API to end users

synnefo

grnet

Integration with Synnefo

VHPC'13
vkoukis@grnet.gr



Future Directions

vkoukis@grnet.gr

I/O Flow identification

Per-user, per-volume policy enforcement for QoS

Archipelago XSEG over RDMA

- Archipelago I/O pipelines across physical nodes
- 1-sided operations for remote communication

Expose the Archipelago data path inside VMs

- Direct VM userspace access to storage
- for ultra-low latency access to storage

Improved handling of multiple backends - Tiered Storage

- Volumes living across backends, automated migrations

Try it out!

VHPC'13
vkoukis@grnet.gr

http://www.synnefo.org

Support: synnefo@googlegroups.com