FLEXIBLE STORAGE FOR HPC CLOUDS WITH ARCHIPELAGO AND CEPH. VANGELIS KOUKIS
TECHNICAL LEAD, SYNNEFO
Outline

Archipelago overview
Storage resources, clones/snapshots
HPC workflow with Archipelago
Resource Composition
Archipelago Implementation
Flexible I/O pipelines
Integration with Synnefo
Future directions
Archipelago overview

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

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
Unified view of resources

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
Spawn

my own Ubuntu

Freeze

my own Ubuntu
The big picture

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)
golden Debian
End-to-end workflow with unified storage
Live demo!

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
Linux block driver

Northbound interface

Volume Composer

Mapper

Archipelago Core

Ceph/RADOS driver

Southbound interface

RADOS

Monitor nodes

Object Storage nodes

block I/O

object I/O
Resource composition
Archipelago interfaces
Running Archipelago
Implementation Internals
Flexible I/O pipeline

- Linux block drv
- Linux block drv
- Volume Composer
- Blocker
- Mapper
- Storage backend
Flexible I/O pipeline

- Linux block drv
- Linux block drv
- Cacher
- Volume Composer
- Blocker
- Mapper
- Storage backend
Flexible I/O pipeline

- Linux block drv
- Linux block drv
- Volume Composer
- Mapper
- Cacher
- Blocker
- Storage backend
Archipelago overview

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

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
Integration with Synnefo

IaaS open source cloud software
Identity, Storage, Compute/Network/Image/Volume services
Production since June 2011
  - Powering the ~oceanos public cloud
Written in Python/Django, exposes OpenStack APIs
Uses Google Ganeti at the backend
BSD licensed
Cluster vs Cloud

<table>
<thead>
<tr>
<th>OPENSTACK</th>
<th>SYNNEFO</th>
<th>UI</th>
</tr>
</thead>
<tbody>
<tr>
<td>OPENSTACK</td>
<td>OPENSTACK</td>
<td>API</td>
</tr>
<tr>
<td>OPENSTACK</td>
<td>SYNNEFO</td>
<td>CLOUD</td>
</tr>
<tr>
<td>OPENSTACK</td>
<td>GANETI</td>
<td>CLUSTER</td>
</tr>
<tr>
<td>LIBVIRT</td>
<td>NODE</td>
<td></td>
</tr>
<tr>
<td>KVM</td>
<td>KVM</td>
<td>HYPervisor</td>
</tr>
</tbody>
</table>
Google Ganeti

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!
Storage service: Pithos

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
Integration with Synnefo

VHPC'13
vkoukis@ernet.gr
Future Directions

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!

http://www.synnefo.org

Support: synnefo@googlegroups.com