UNIFIED CLOUD STORAGE WITH SYNNEFO + GANETI + ARCHIPELAGO + CEPH

VANGELIS KOUKIS, TECHNICAL LEAD, SYNNEFO
Running a public cloud: ~okeanos

History
- Design started late 2010
- Production since July 2011

Numbers
- Users: > 5000
- VMs: > 7000 currently active
- More than 250k VMs spawned so far, more than 70k networks
Running a public cloud: ~okeanos

Our choices

- Build own AWS-like service (Compute, Network, Storage)
- Persistent VMs
- Everything open source
- Production-quality IaaS
- Super-simple UI

How?
Running a public cloud: ~okeanos

The tough stuff

- Stability
- Persistent VMs: VMs are not cattle, they are pets
- Commodity hardware
- Scalability
- Manageability: Gradual rollout of upgrades and new features
Running a public cloud: ~okeanos

Our approach

- Synnefo
- Google Ganeti
- DRBD
- Archipelago
- Ceph
- OpenStack APIs
Architecture

 synnefo

FOSDEM'14
vkoukis@grnet.gr
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
View/create/destroy virtual networks
Interconnect VMs, with NIC hotplugging
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 Image
...from the command-line, and in Python scripts
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!
Identity: Astakos

Identity Management, Resource Accounting and SSO
- Platform-wide service
- Simple service- (Cyclades, Pithos) and user-facing APIs
- Multiple authentication methods per user
- Fine-grained per-user, per-resource quota

A single dashboard for users
- View/modify profile information and active authentication methods
- Easy, integrated reporting of per-resource quotas
- Project management: View/Join/Leave projects
- Manage API access and retrieve authentication tokens
Identity: Astakos

Supported 3rd-party providers
- Shibboleth / AAI Federation
- Google
- Twitter
- LinkedIn
Thin Compute layer over Ganeti

- Python/Django
- Supports *multiple* Ganeti clusters, for scaling
- OpenStack APIs

Networking

- No restrictions on deployment – it’s the *Ganeti* side
- IPv4/IPv6 public networks, complete isolation among VMs
- Thousands of private networks, private L2 segments over single VLAN
- Software-Defined Networking, pilots with VXLAN integration
Compute/Network/Image/Volume: Cyclades
Interaction with Ganeti

Support for all Ganeti storage templates

External Storage Interface for SAN/NAS support

Networking = gnt-network +
  snf-network (KVM ifup scripts) +
  nfhdhcpd (custom NFQUEUE-based DHCP server)

Asynchronous operation
- **Effect path:** Receive API requests, enqueue requests over RAPI
- **Update path:** Receive asynchronous notifications, update DB
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
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
Clone

Ubuntu + user data

Snapshot
The big picture

Storage backend 1 (e.g., Ceph cluster 1)
Storage backend 2 (e.g., Ceph cluster 2)
Storage backend 3 (e.g., NFS over NAS)

Archipelago Core
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

SNAPSHOT FILE \(\xrightarrow{\text{CLONE}}\) VOLUME \(\xrightarrow{\text{LIVE VOLUME}}\) SNAPSHOT FILE

ARCHIPELAGO

MAP1

RO \(\xrightarrow{\text{MAP2}}\) RW

MAP2

RO \(\xrightarrow{\text{MAP3}}\) RW

MAP3

STORAGE

[Diagram of resource composition with various components and connections]
Archipelago interfaces
Running Archipelago
Flexible I/O pipeline

- Linux block drv
- Cacher
- Volume Composer
- Cacher
- Mapper
- Storage backend
Experience

Operations

- Rolling hardware and software upgrades
  - kernel, Ganeti, RADOS, Synnefo
  - with no VM downtime
- Node evacuations with live VM migrations
- Cross-datacenter move, Intel → AMD, no VM downtime
- On-the-fly migration from NFS-backed storage to RADOS
- IP renumbering of all VMs
Experience

Scalability
- From few physical hosts to multiple racks
  - dynamic addition of Ganeti clusters

Diverse workloads
- Different network and storage backends
- Choice exposed to the user
Try it out!

http://www.synnefo.org
Thank you!