Kubernetes OS in eBay

Liu Yingke

Agenda

Node OS evaluation and selection

OS management and upgrade

Host network introduction

Node OS Evaluation and Selection

Ecosystem and Maturity

Docker storage backend

Package Management and OS upgrades

Security

Node OS Evaluation and Selection

Ecosystem and Maturity

	CentOS/ Fedora	Ubuntu	CoreOS	Atomic*	Snappy	RancherOS
Ecosystem & Maturity	General-Purpose		Container-Optimized			
	Mature		Earliest container-optimized distribution. But the company is small	Backed by Red Hat	Backed by Canonical. The OS is designed for mobile originally	Most New. Everything in RancherOS is a docker container

The benefits of container-optimized distribution.

Small sized

OS **atomic updates** and **atomic rollback**.

Node OS Evaluation and Selection

Docker storage backend

- Device mapper is the our current option as it is the most production ready docker storage driver

Better support from RedHat

- Specify partitions into PV
- Auto grow
- Percentage usage
- Root size

Node OS Evaluation and Selection

Package Management

	CentOS 7	Ubuntu 15.10	CoreOS 835.8.0	Atomic F23 *	Snappy Ubuntu 15.04	RancherOS 0.4.1
Package Manager	rpm	dpkg	update_egine	rpm-ostree	snappy	docker

- CoreOS, its OS upgrade system CoreUpdate is not free thus we can't setup local repo in production environment.
- Snappy, whole base OS is a snappy package, i.e. ubuntu-core, but no instructions to build it by ourselves, and user can't setup the private snappy store.
- **Atomic**, rpm-ostree is opensource and easy to setup local repo in production.

Node OS Evaluation and Selection

Security

	CentOS 7	Ubuntu 15.10	CoreOS 835.8.0	Atomic F23 *	Snappy Ubuntu 15.04	RancherOS 0.4.1
Security	SELinux	AppArmor	SELinux	SELinux	AppArmor	SELinux

Node OS Evaluation and Selection

Why Atomic?

- Container-optimized
- Backed by Red Hat, which has most amount of contributors to linux kernel
- Multiple distributions options
  - Fedora: Comes with fairly **newer versions** of kernel and docker.
  - CentOS: Production ready and free. With better Same code build with RHEL.
  - RHEL: Comes with Enterprise support.
- Package management, **rpm-ostree** is opensource and easy to setup local repo in production.

Atomic Deep Dive

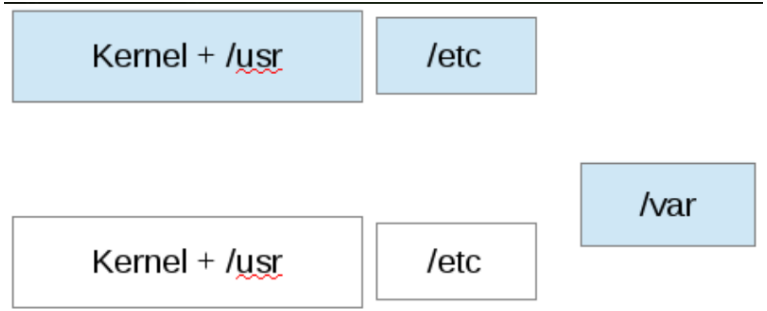
OSTree File System Model

- Easy to rollback, two versions of OS snapshots are in the system.
- More secure.

  /usr is read-only bind mount.

  /etc is ‘rebased’ on upgrades

  /var is untouched



Rpm-ostree Management is easy.

Docker storage driver

- Device mapper, direct-lvm mode

OS management

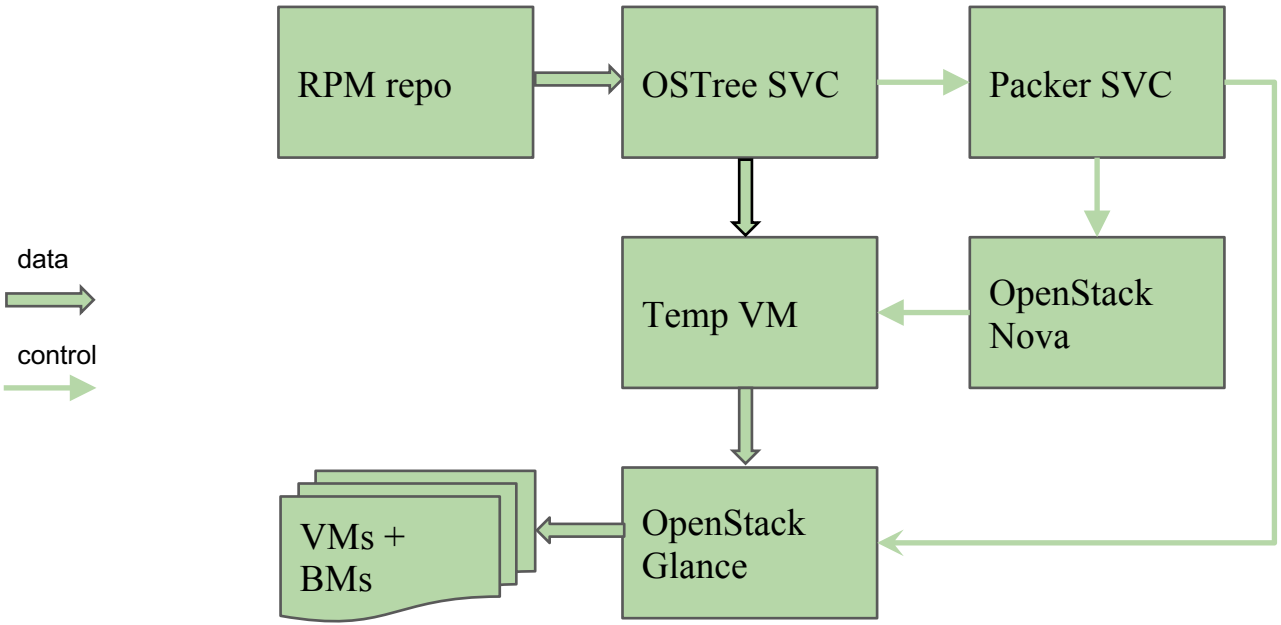
OS image generating

OS upgrade

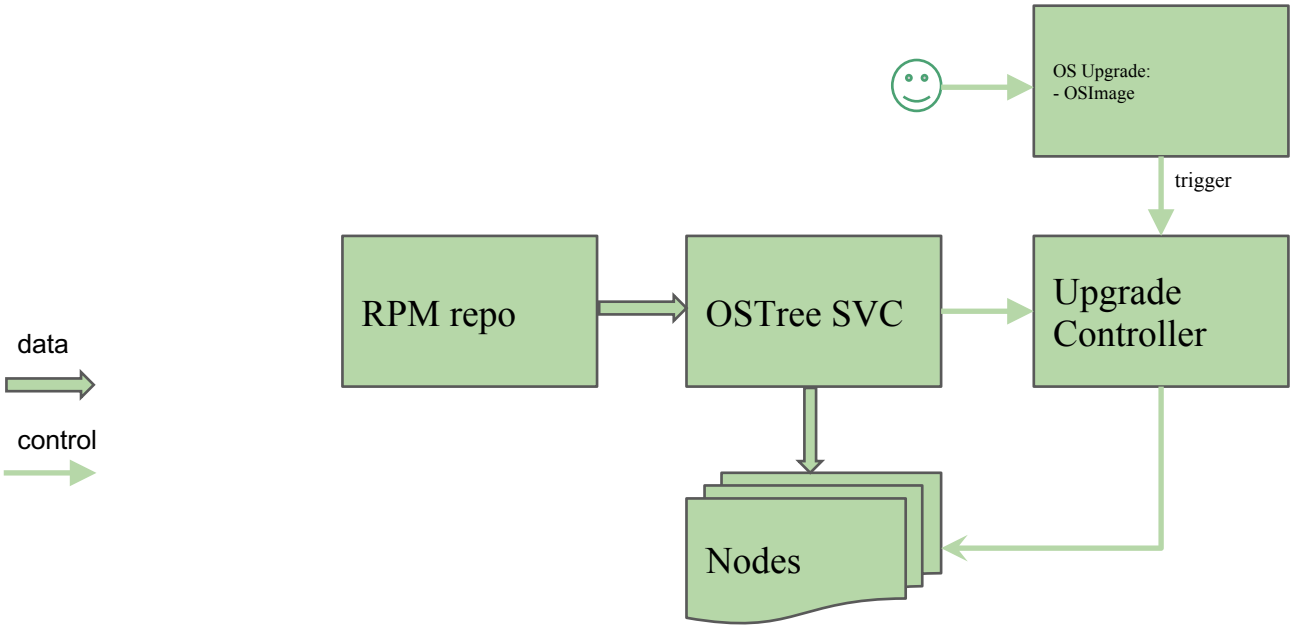
OS configuration

OS monitoring

Image generating



OS Upgrade


OS Upgrade:
- OSImage
trigger
RPM repo
OSTree SVC
Upgrade Controller
Nodes
data
control

OS configurations

All packages are installed in a read-only partition

Parameter configuration:

- A salt module downloaded to salt-master by controller
- Apply nodes with salt

Host monitoring

Kubelet collects most of metrics (CPU/memory/disk)

Node problem detector as privileged container on each host

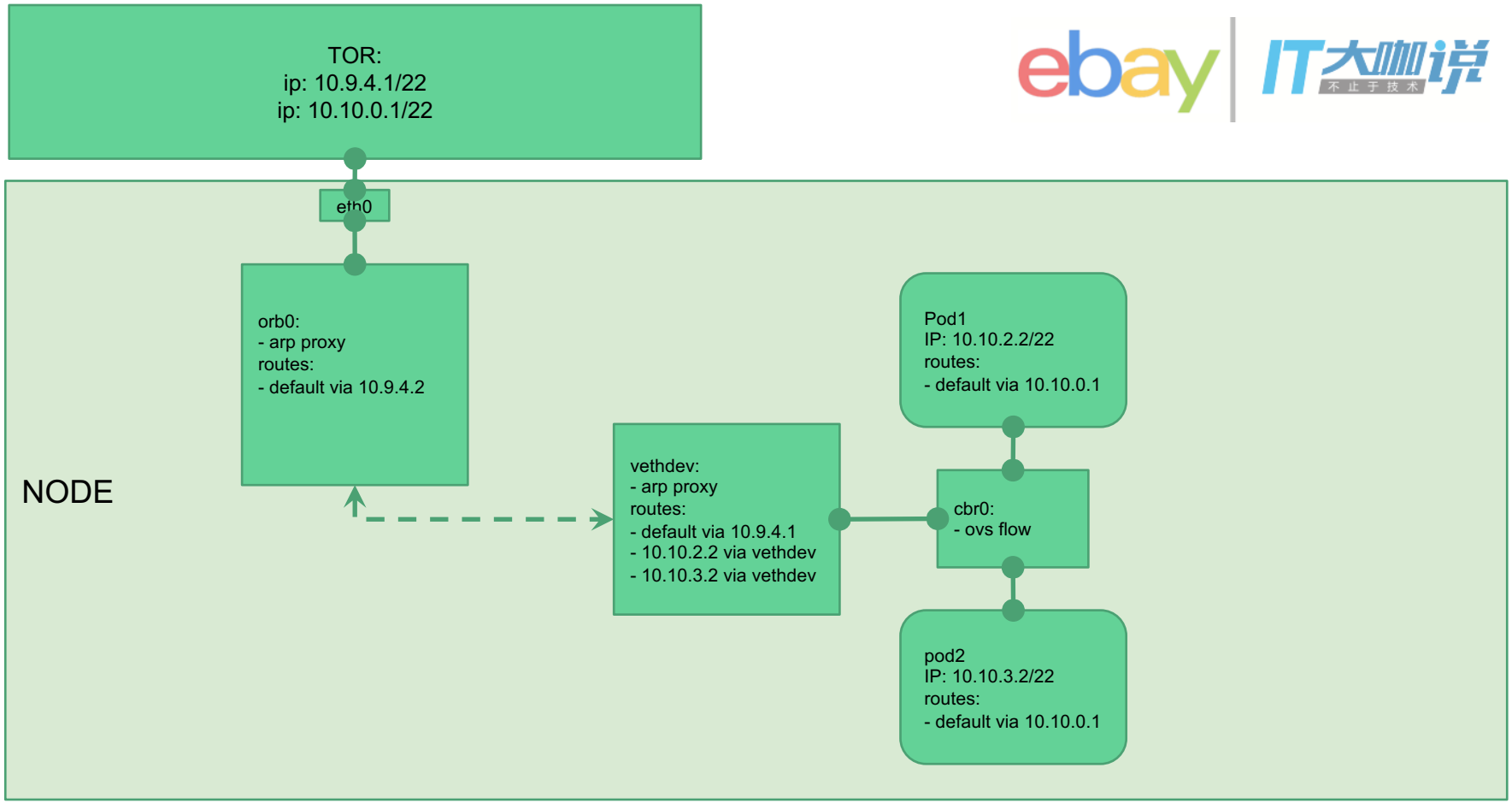
Host Network Introduction

Requirements

- Pod to Pod inter-communication
- Pod to Node in cluster inter-communication
- Pod to other out of cluster inter-communication

Limitations

- We use a big subnet to TOR, cannot do L3 to host


TOR:
ip: 10.9.4.1/22
ip: 10.10.0.1/22
eth0
orb0:
- arp proxy
routes:
- default via 10.9.4.2
NODE
vethdev:
- arp proxy
routes:
- default via 10.9.4.1
- 10.10.2.2 via vethdev
- 10.10.3.2 via vethdev
Pod1
IP: 10.10.2.2/22
routes:
- default via 10.10.0.1
cbr0:
- ovs flow
pod2
IP: 10.10.3.2/22
routes:
- default via 10.10.0.1

Questions