Kubernetes the CoreOS way

Luca Bruno

@lucabruno | luca.bruno@coreos.com | github.com/lucab

$ whoami

“CoreOS Engineer, Debian Developer and enthusiast FLOSS supporter”

- OS Software Engineer
- CoreOS, Berlin office
- PoliTo graduate
- Previously: security researcher/engineer

CoreOS Inc.

“Infrastructure with autopilot: improve reliability and security regarding critical updates, machine failures, networking outages”

- YC’13 startup
- FLOSS-centric company
- Offices in SF, NYC, Berlin

Technologies

Base ingredients of a modern stack:

- Linux
- Systemd
- Go
- Etcd
- Docker

Kubernetes

Container-based cluster orchestration system.

- Originally by Google, now under CNCF
- Higher level primitives (pods, services, deployments)
- Current stable release: 1.8

Tectonic

Commercial Kubernetes offering by CoreOS.

- Focused on automated-operations and self-updates
- Upstream kubernetes with addons, no forked components
- Our opinionated approach to infrastructure
- Source of technical materials for this talk :)

OS components

ContainerLinux

Compact Linux distribution without package management.

- Immutable OS (read-only /usr)
- Autoupdates with A/B partition-scheme
- 3 release channels:
  - Stable - 8 weeks
  - Beta - 4 weeks
  - Alpha - 2 weeks

Systemd

Dependency-based Linux init system and userspace framework.

- Originally from RedHat
- Based on unit files and dependency graph
- Extensive DBus interface
- Umbrella project for several userspace blocks:
  - udev
  - networkd
  - journald

https://github.com/systemd/systemd

Update-engine + Locksmith

Autoupdate service and updates management.

- Based on Omaha protocol (HTTPS + XML + sigs)
- Originally from Google ChromeOS
- Updates flushed to passive partition
- Updates applied via reboot
- Cluster-wide maintenance scheduling
  - Either via locksmith (bare) or CLUO (kubernetes)

https://github.com/coreos/locksmith

Ignition

First-boot machine provisioning tool.

- Userdata-based node setup
- Runs early in initrd
  - Can partition disks, mount volumes, etc
- Typed JSON as input (with external YAML transpiler)
- Support fetching remote assets

https://github.com/coreos/ignition

Torcx

Addon system for ContainerLinux.

- (Unconventional) systemd generator
- Complementary to Ignition
- Applies ephemeral changes on every boot
- Allows selecting addons version at runtime

https://github.com/coreos/torcx

Docker

Container runtime and tooling.

- Single-stop solution for anything containerized
- Originally from Docker Inc., some parts under CNCF
- Centralized, daemon-based design
- Easy, high-level CLI to cgroups and namespaces

rkt

An alternative, opinionated container runtime.

- Daemon-less, single process tree
- Native systemd integration
- Pods as execution boundaries
- Based on open specs (AppC / CNI), modularity and decentralization

https://github.com/rkt/rkt

Cluster components

Terraform

Infrastructure as code.

- Declarative language to provision cloud infrastructure
- Takes care of node first configuration (via ignition)
- Multiple plugins and modules to talk to cloud APIs

HashiCorp
Terraform

Kubelet

Kubernetes agent running on each node.

- Deployed as a systemd+rkt service unit
- Drives container execution via CRI (GRPC)
- Multiple backends/shims:
  docker, containerd, rkt, cri-o, hyper

etcd

Distributed key-value store.

- Consistent database (in CAP terms)
  - Based on the Raft consensus algorithm
- Currently at major version 3
- GRPC interface (HTTP2 + protobuf)
- Deployed as a systemd+rkt service unit

https://github.com/coreos/etcd

CNI

Spec and plugins for container networking.

- Originally from rkt, now under CNCF
- Decouple networking from container runtimes
- Plugin system, customizable for your network setup
- Simple spec plus some core plugins
- Linear chaining with JSON I/O

CNI

Flannel

CNI implementation of overlay networking.

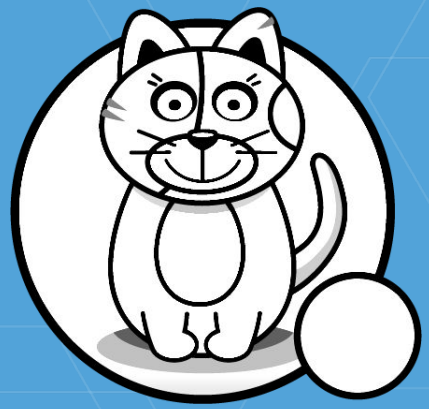
- Originally from CoreOS, currently by Tigera
- Provide a flat network for pods
- Overlay based on VXLAN encapsulation

https://github.com/coreos/flannel

Calico

Pod network policy.

- Originally from Tigera
- Provide network segmentation for pods
- Compatible with CNI
- Large configuration matrix, including:
  - Iptables on top of Flannel overlay (canal)
  - Iptables on top of BGP (calico)



Prometheus

Large scale monitoring.

- Open source clone of Google's Borgmon
- Scalable time series storage and alerting
- Poll-based design
- Protobuf exposed over HTTP endpoint

https://github.com/prometheus/prometheus

Kubernetes operators

Automated pilots for kubernetes apps.

- Putting operation knowledge into software
- Originally taking care of CoreOS components
  - Etcd, ContainerLinux, Prometheus
- Larger family by now
  - Vault, Rook, Habitat and more

CLUO

ContainerLinux Update Operator

- K8s-aware replacement for Locksmith, two pieces:
  - Cluster-wide operator
  - Node-local agent
- Allows for cluster-coordinated node upgrades, managed inside kubernetes
- Pre-reboot and post-reboot hooks

https://github.com/coreos/container-linux-update-operator

Demo time
+
Questions

We’re hiring in all departments!

Email: careers@coreos.com Positions: coreos.com/ careers