Kubernetes

Peng Yanghua

Agenda

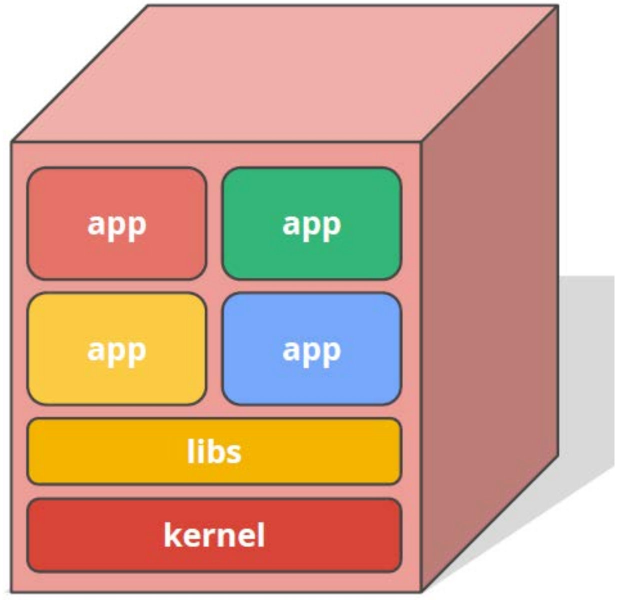
- Background
- Architecture
- Core Concepts
- Kubernetes for distributed machine learning

Background

Old way: applications on host

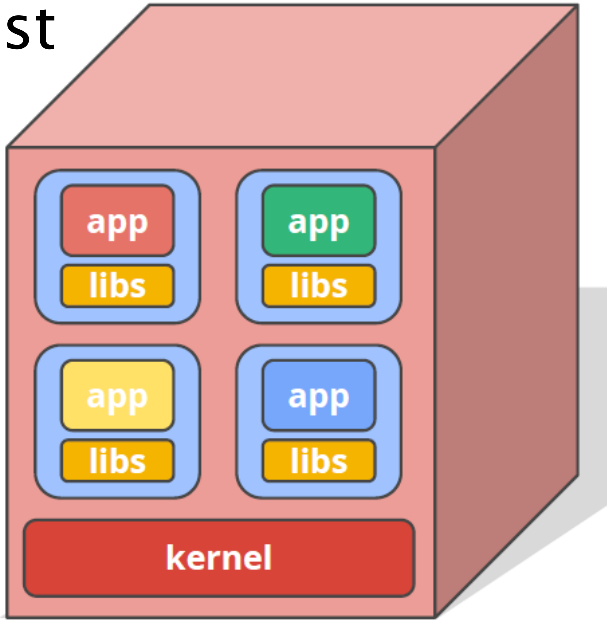
- Applications and OS share filesystem.
  - Libraries, configurations, resources entangled with each other and the host

- VM
  - Heavy-weight



New way: applications in container

- Container
  - OS-level virtualizations
  - Isolated from each other and the host
  - Small and fast
  - 1 app to 1 image



Need a container-centric platform

- Management unit: host -> container

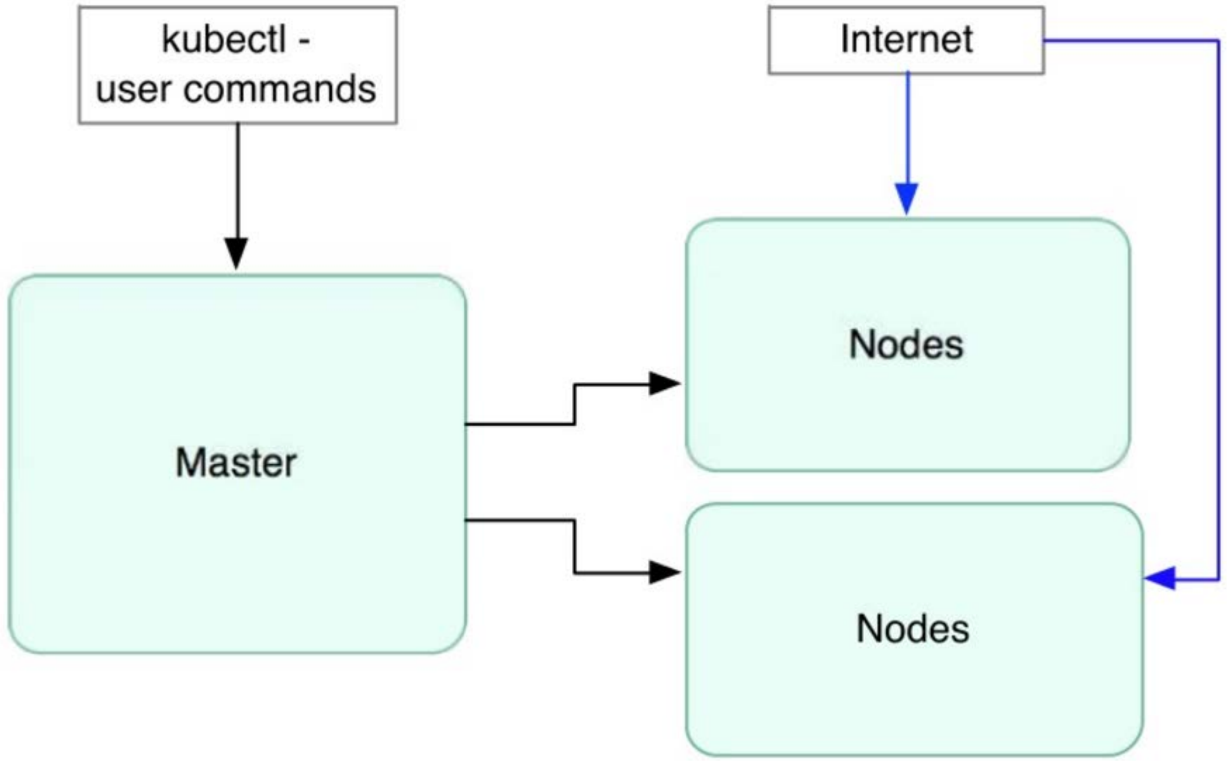
- Automate orchestration for scale, just like Openstack for virtual machines.

Kubernetes

- A platform for automating container deployment, scaling and operations.
- Based on Google's 15-year experience on Borg.
- Open source, written in Go language.

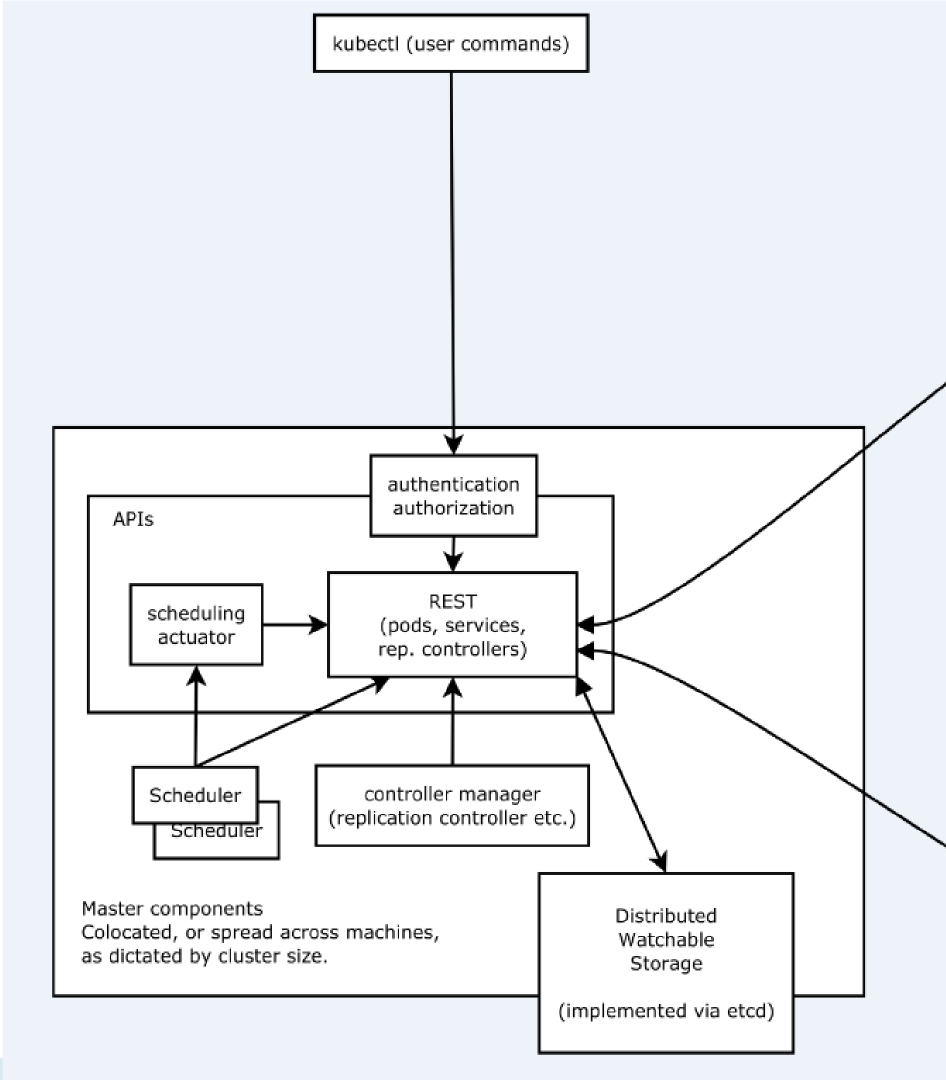
Architecture

High-level architecture



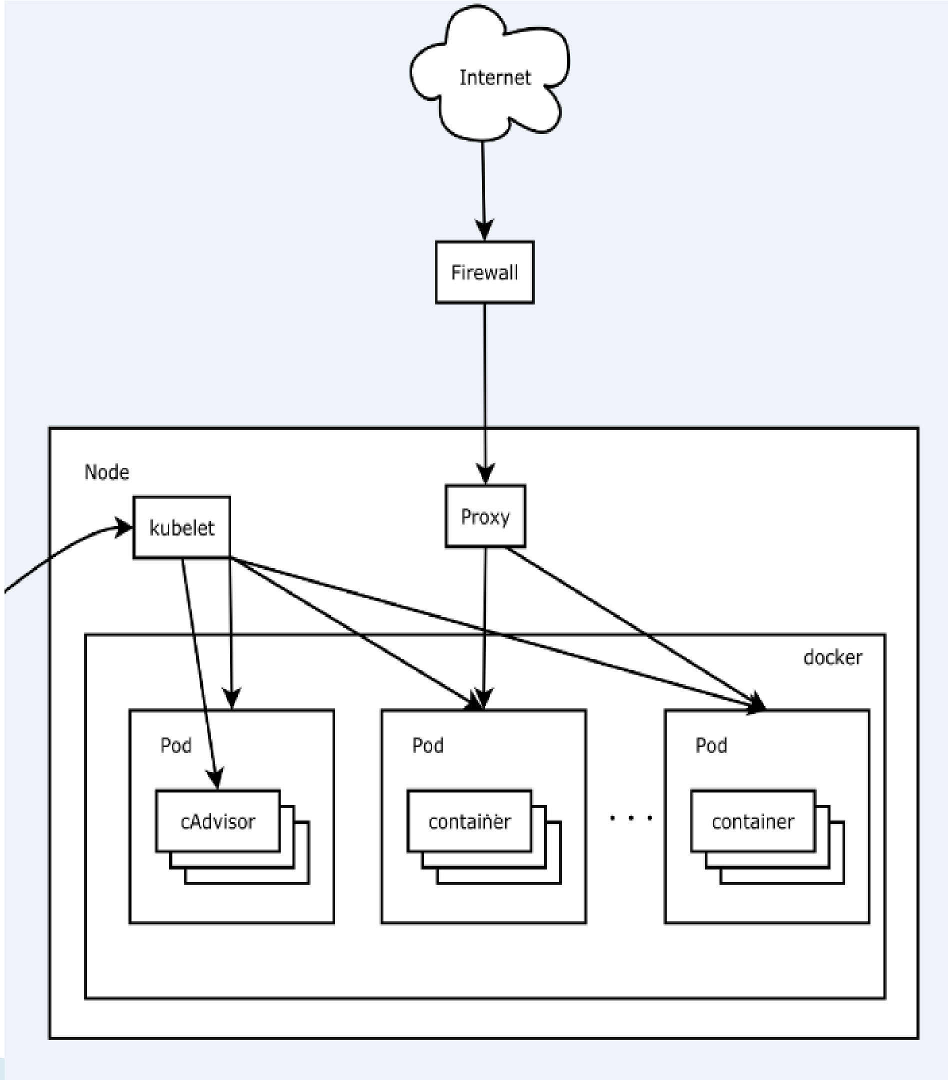
Master

- REST API
  - object operations
  - authentication, authorization
- Scheduler
  - random
  - round-robin
- Controller manager
  - node controller
  - replication controller
- Storage
  - etcd, distributed reliable


kubectl (user commands)
APIs
authentication authorization
scheduling actuator
REST (pods, services, rep. controllers)
Scheduler
Scheduler
controller manager (replication controller etc.)
Master components Colocated, or spread across machines, as dictated by cluster size.
Distributed Watchable Storage (implemented via etcd)

Node

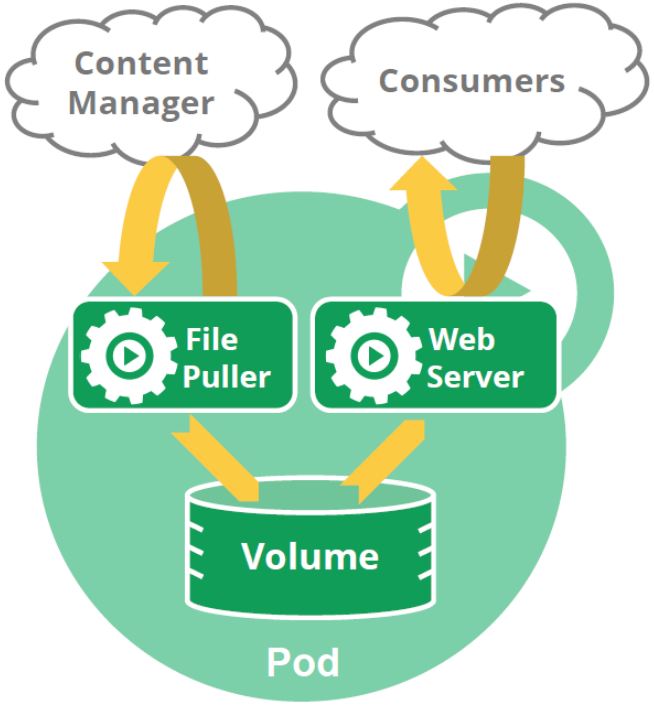
- kubelet
  - node agent
- kube proxy
  - user requests
- Docker
  - container runtime
- cAdvisor
  - container monitoring
- Flannel
  - inter-node overlay



Core concepts

Pods

- A group of containers sharing common resources
  - e.g., IP, filesystem.
  - usually different components of an app.
- Pod-level scheduling
  - mostly one container one pod

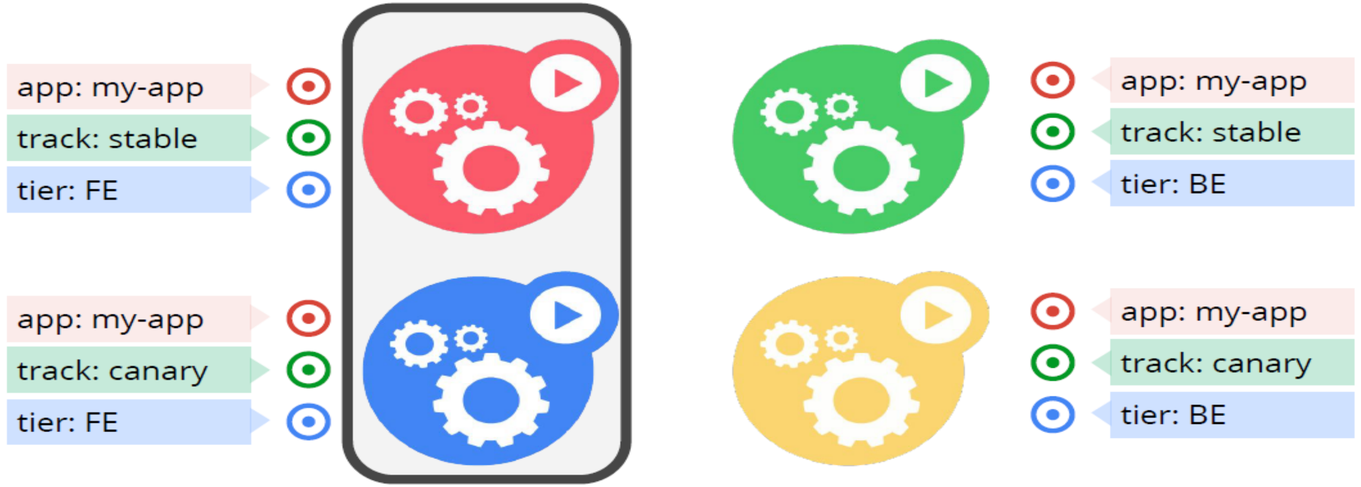


Volumes

- Pod-level shared storage
  - Any containers within the pod can access it.
  - Communication

Labels

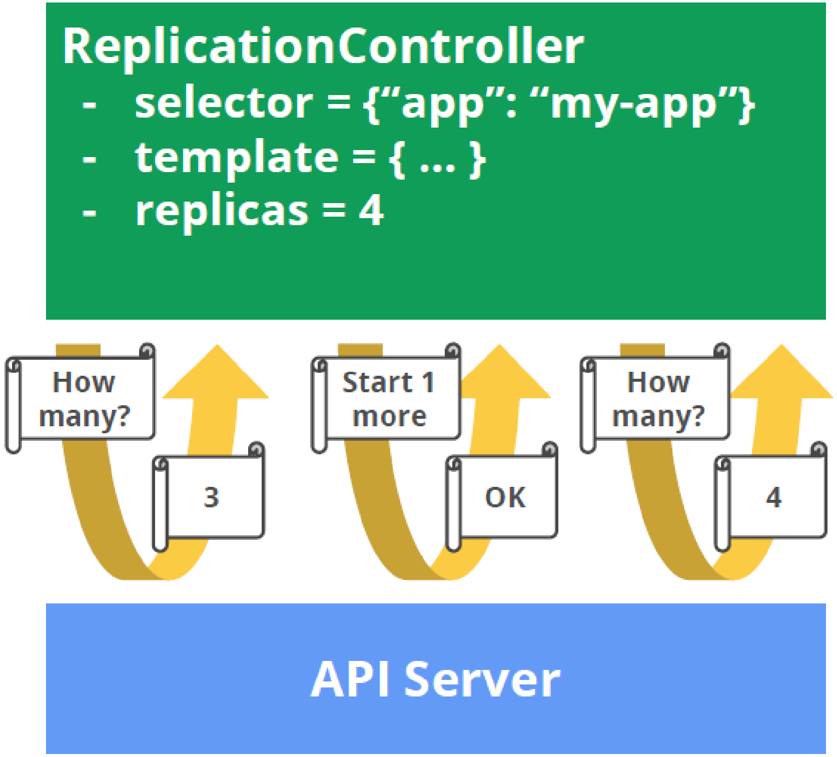
- A key-value pair attached to an object
  - query-able by selectors
  - the only grouping mechanism



app = my-app, tier = FE

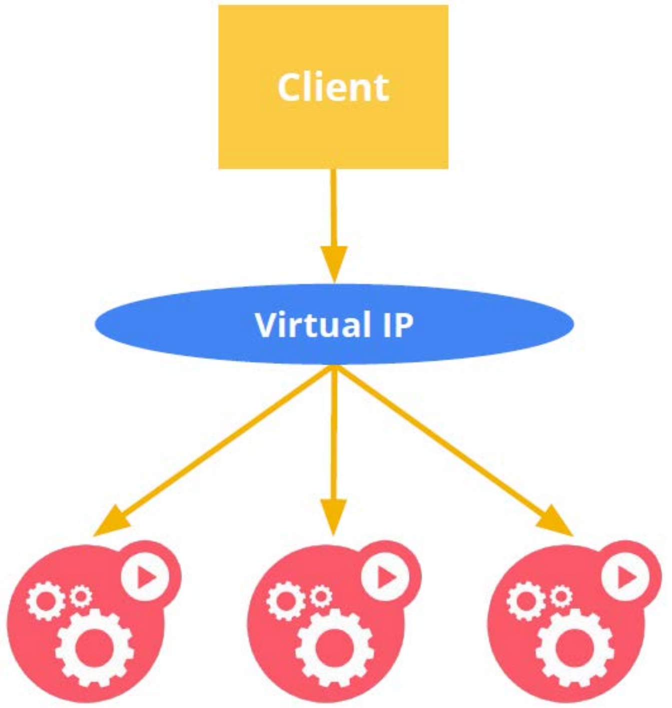
Replications

- Ensure N copies of a pod
  - or auto-scaling
  - grouped by a label selector
- Rolling updates
  - +1/-1
  - without downtime



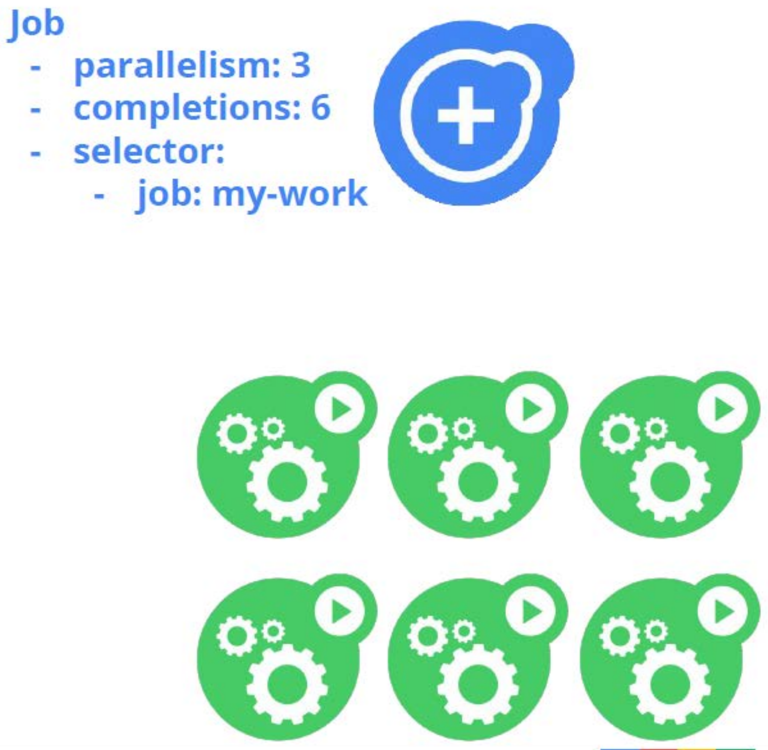
Services

- A group of same pods that work together
  - grouped by a label selector
  - a virtual IP for client access
  - discover via DNS



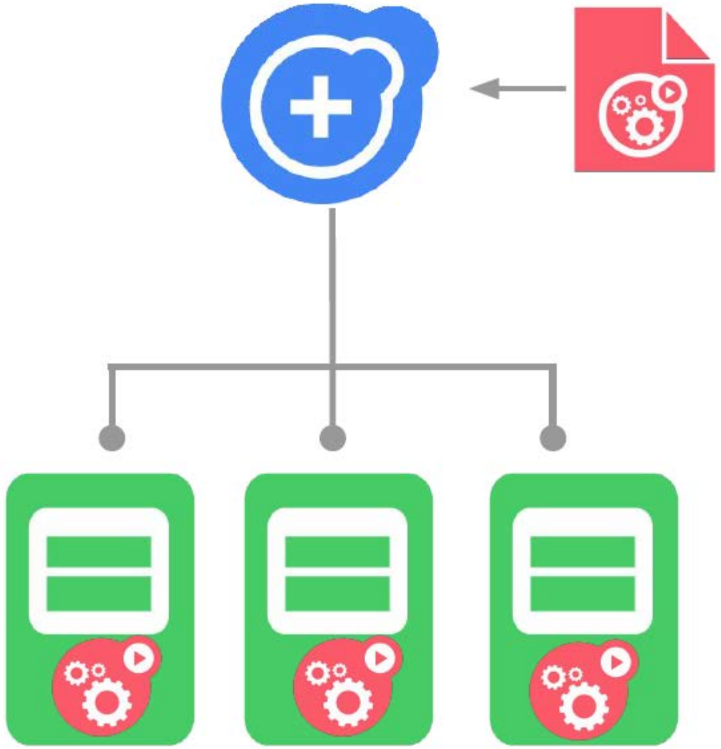
Jobs

- A group of pods that run to completion
  - pods do not always restart on failures
  - usually no external IP for access



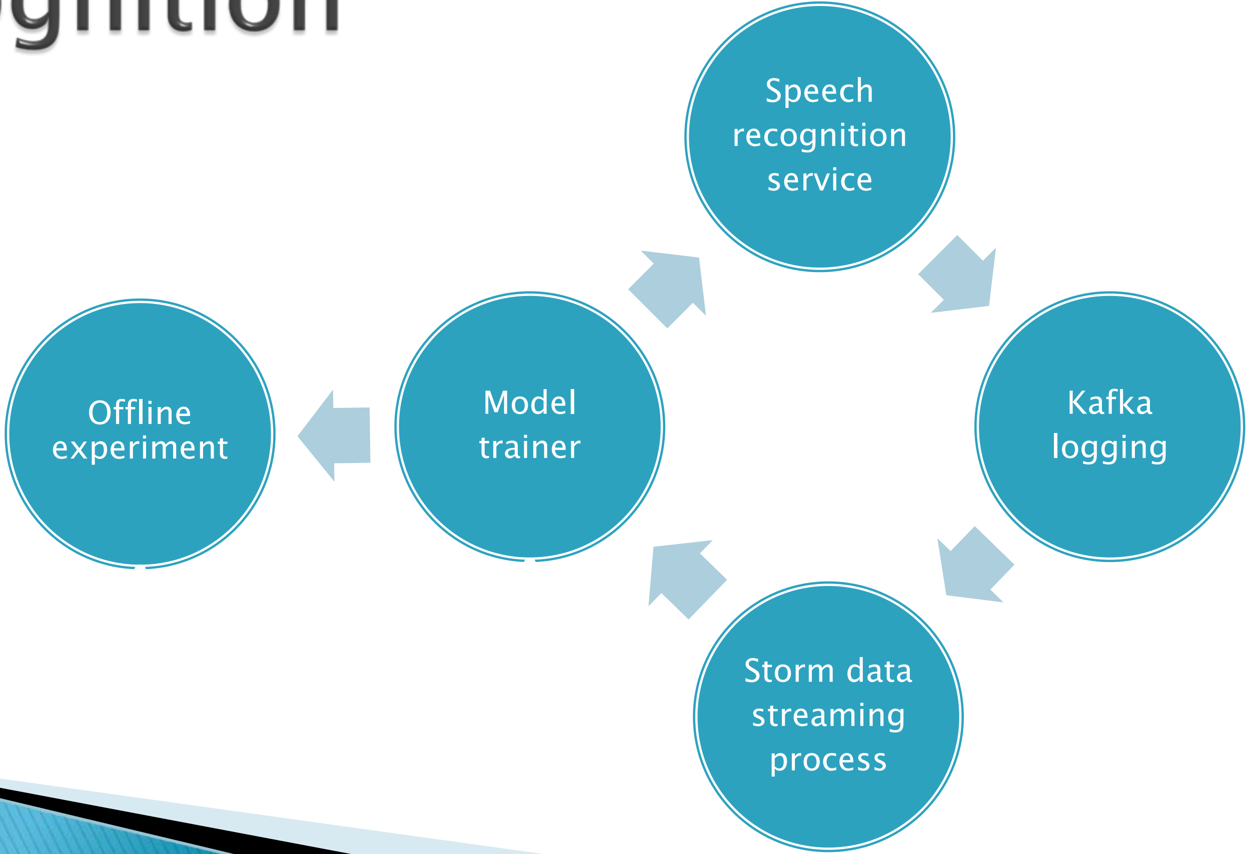
DaemonSets

- Run a pod on every node
  - created as nodes come and go
  - useful for running cluster-wide services, e.g., logging, storage.



Kubernetes for Distributed Machine Learning

Workload example: speech recognition



Why kubernetes?

- A platform
  - Deep learning system
  - A lot of others, such as web server, the logger, data processor etc.

- Online and offline jobs
  - Online: serve production traffic
  - Offline: experiments

kubernetes: running different kinds of workloads efficiently in a cluster.

But kubernetes is not so perfect

- Resource allocation is static.
  - Manually change resource requirements if having more resources
- Deep learning job configuration
  - Best number of parameter servers and trainers
  - Where to deploy
- GPU support is not enough.
- **We need a job scheduler running on kubernetes in machine learning clusters for better utilization and efficiency.**

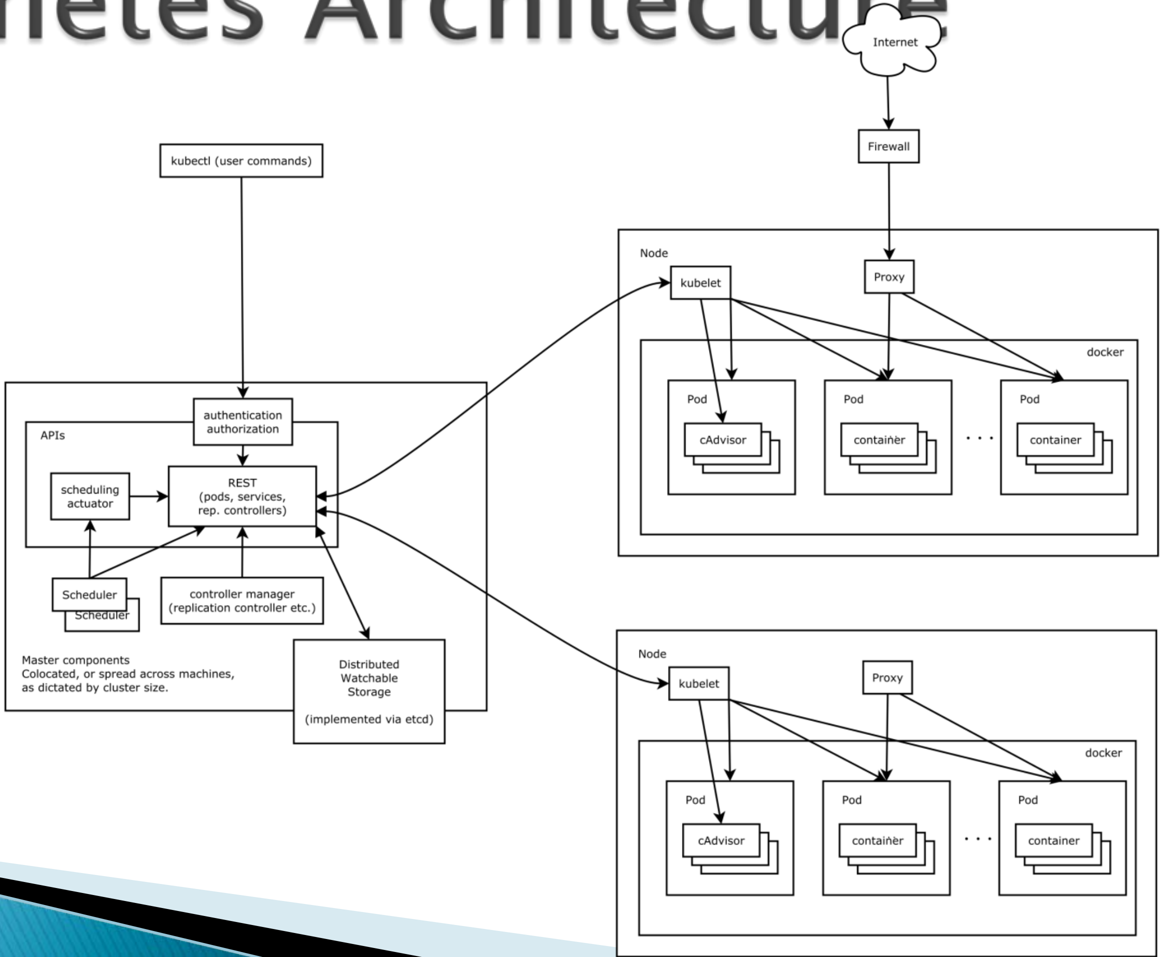
Thanks

Backup

Comparison with Mesos

- Resource isolation and sharing across distributed frameworks, e.g., spark, kubernetes.

Kubernetes Architecture


Internet
Firewall
kubectl (user commands)
authentication authorization
APIs
scheduling actuator
REST (pods, services, rep. controllers)
Scheduler
Scheduler
controller manager (replication controller etc.)
Master components Colocated, or spread across machines, as dictated by cluster size.
Distributed Watchable Storage (implemented via etcd)
Node
kubelet
Proxy
docker
Pod
cAdvisor
Pod
container
. . .
Pod
container
Node
kubelet
Proxy
docker
Pod
cAdvisor
Pod
container
. . .
Pod
container