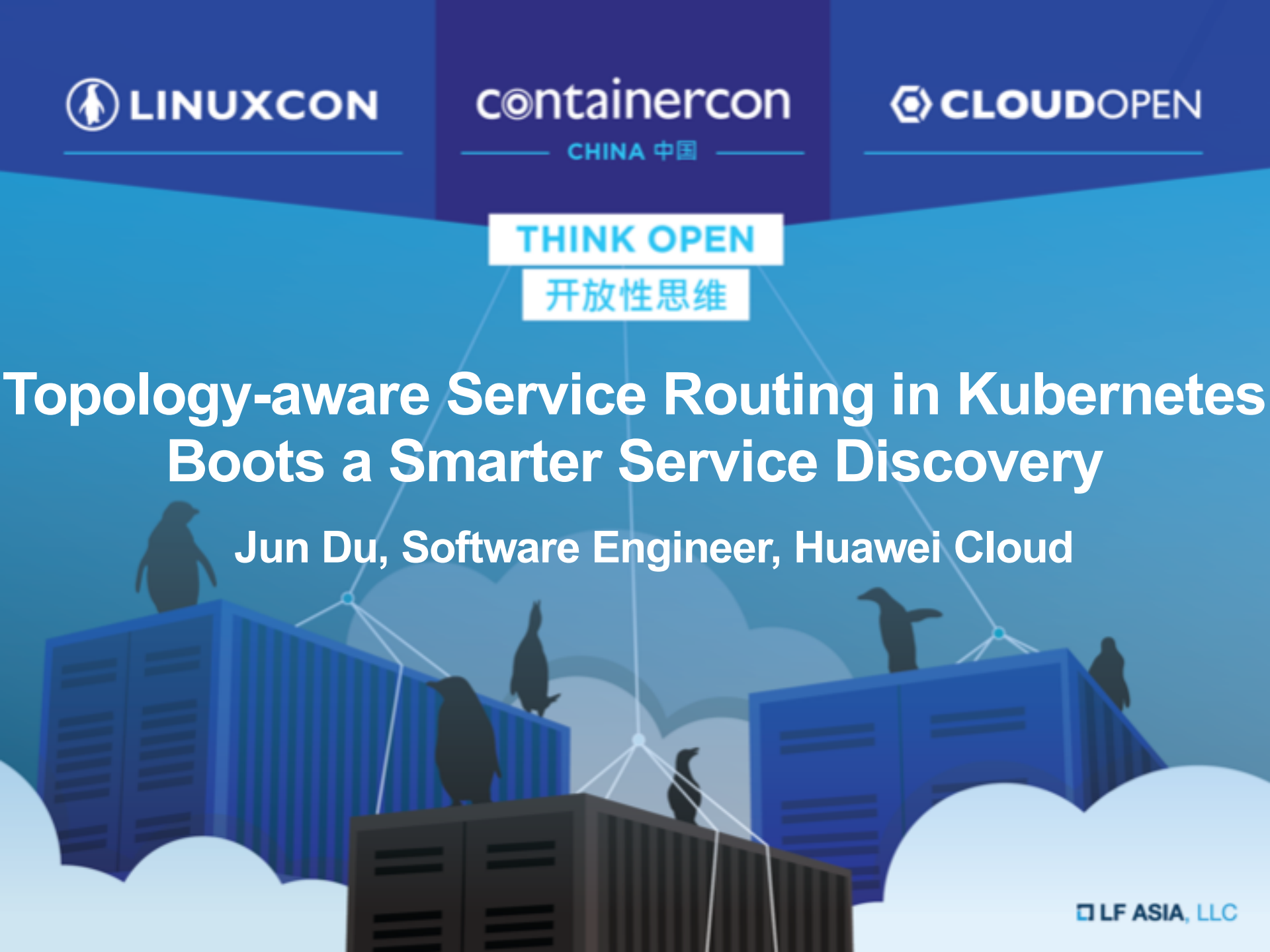


Topology-aware Service Routing in Kubernetes Boots a Smarter Service Discovery

Jun Du, Software Engineer, Huawei Cloud

LF ASIA, LLC

Agenda

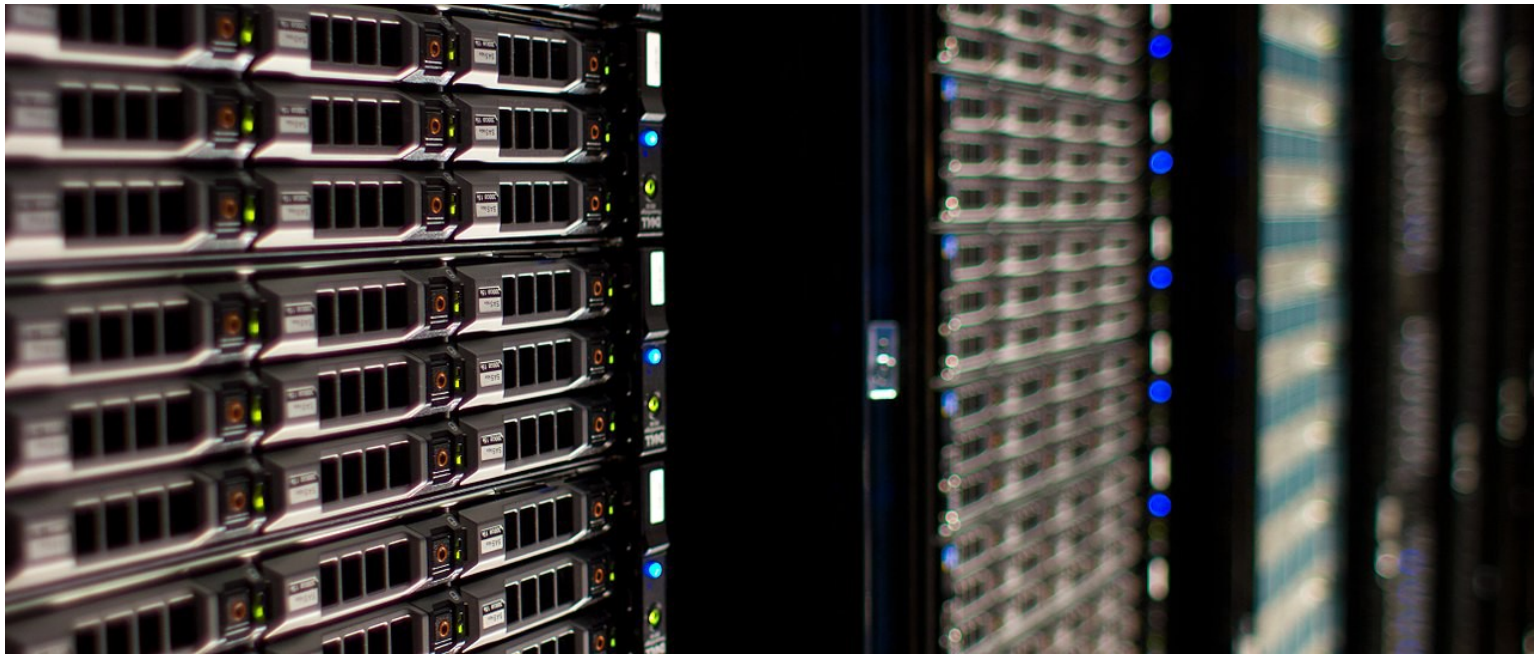
- Topologies in Kubernetes
- Topology-aware service routing
- Solutions and prototypes
- Q&A

Topology is Arbitrary

- AZ
- Region
- Rack
- Host
- Generator
- Anything you like…

Topology in Kubernets scheduler

Where should I run this Pod?



Scheduling is about finding hardware to run your code.

Node Affinity

Should I run my Pod on this Node?

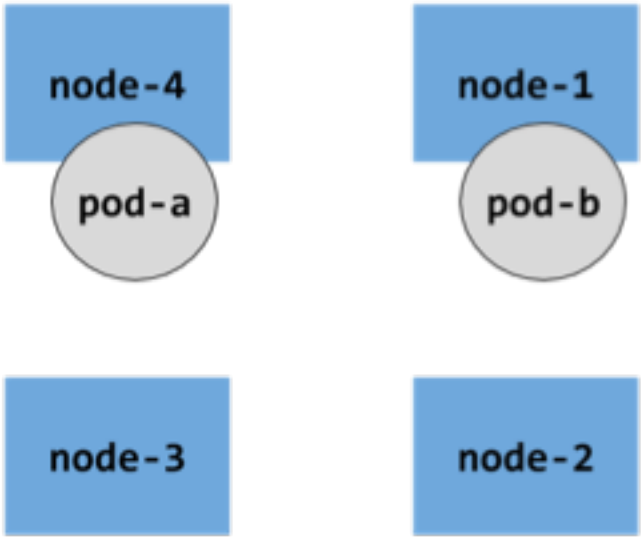
```
# provided by Kubernetes:
k8s.io/hostname
failure-domain.beta.k8s.io/zone
failure-domain.beta.k8s.io/region
beta.k8s.io/instance-type
beta.k8s.io/os
beta.k8s.io/arch

# user-defined (cluster admin, cloud provider, etc):
rack, disktype, ...

pod:
  name: postgres-primary
  ...
  affinity:
  - node: failure-domain.beta.k8s.io/zone=us-east-1a
---
pod:
  name: postgres-standby
  ...
  affinity:
  - node: failure-domain.beta.k8s.io/zone=us-east-1b
```

Pod Affinity/Anti-affinity

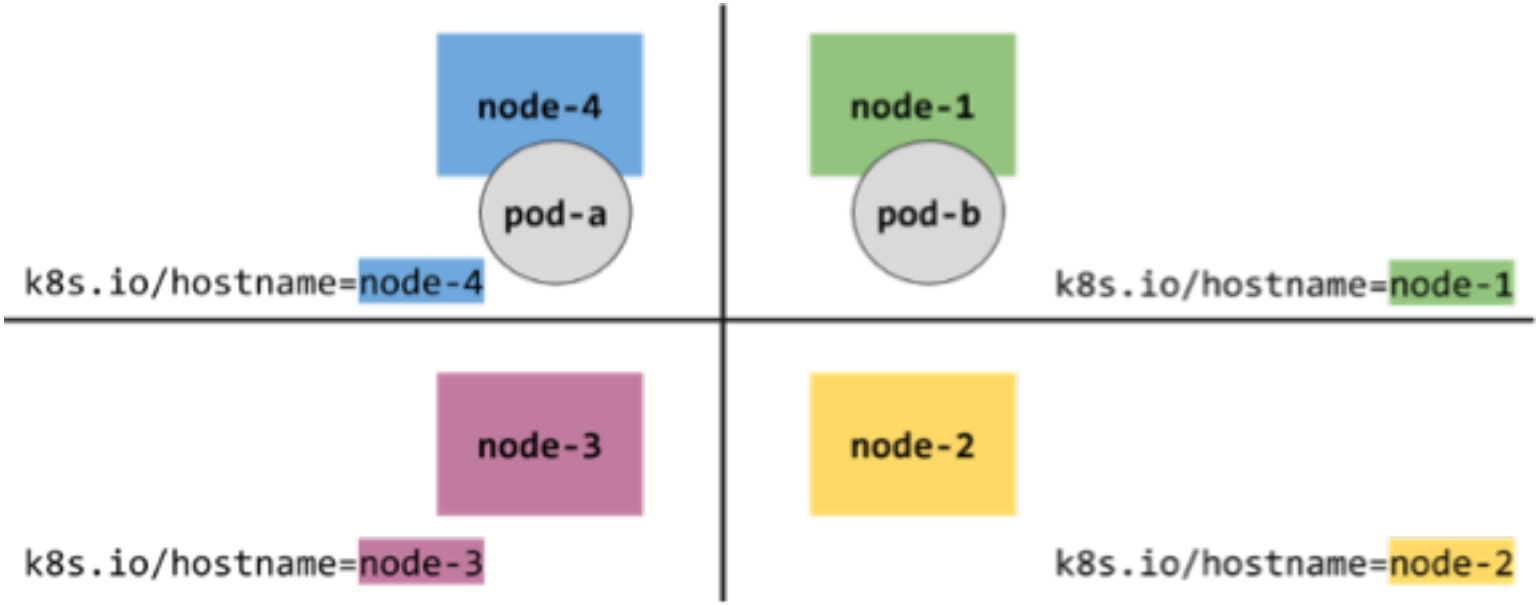
- Labels identify topologies
- topologyKey is the key of Node Labels

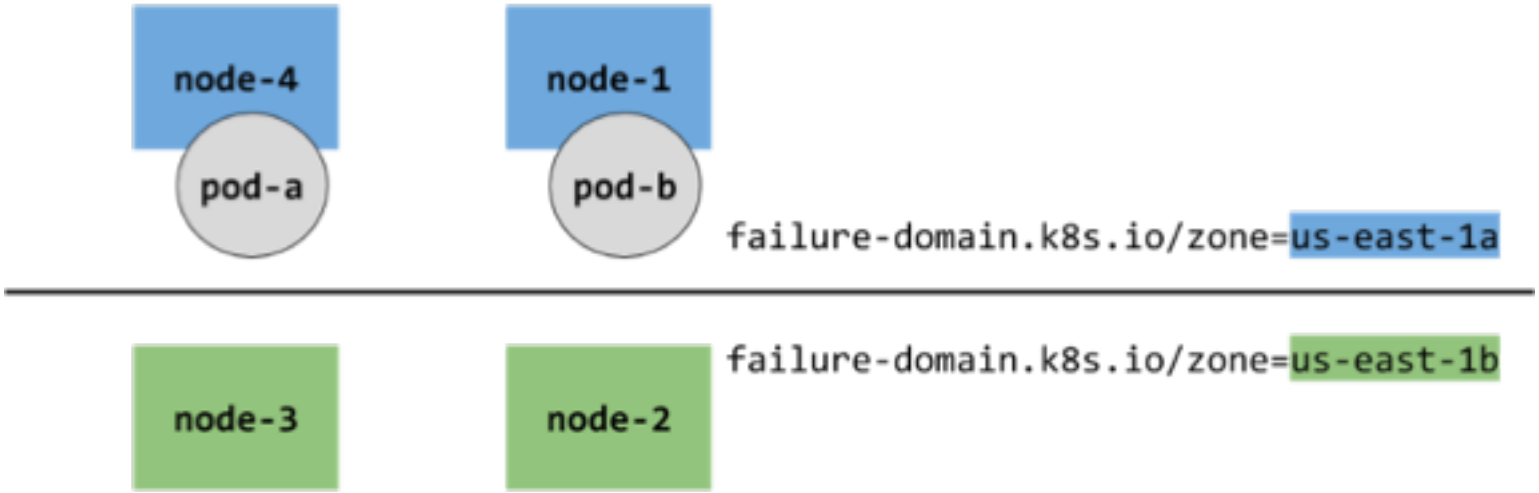


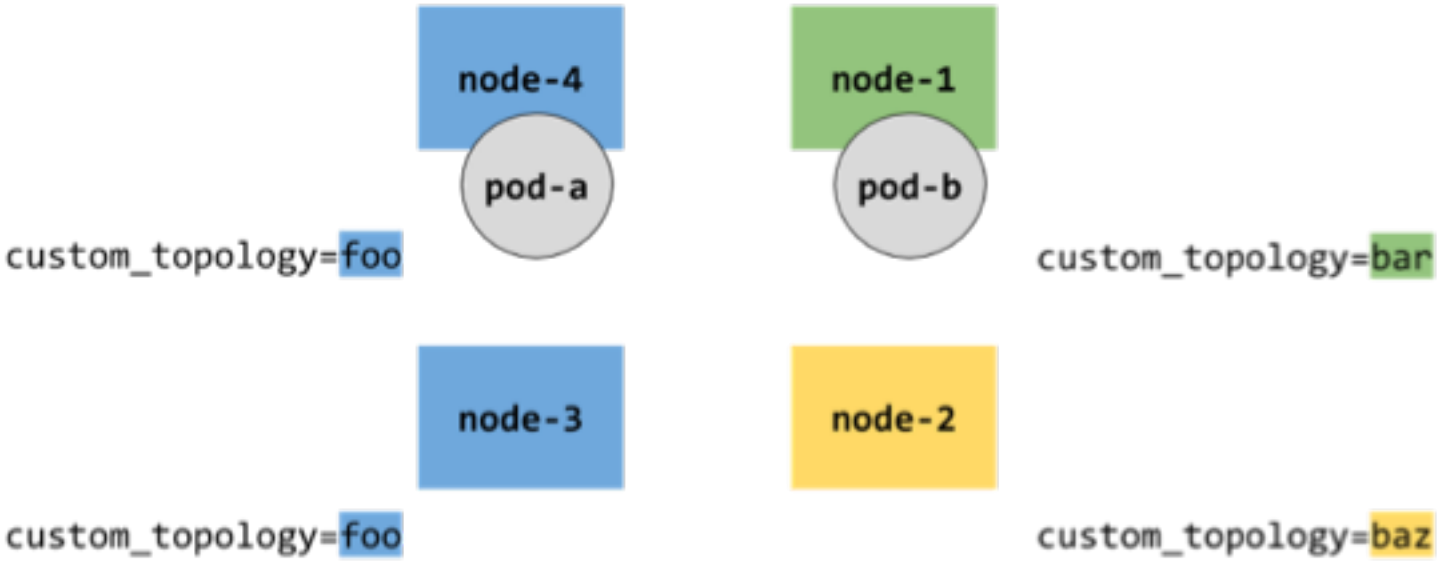

```
podAffinity:
    requiredDuringSchedulingIgnoredDuringExecution:
        - labelSelector:
            matchExpressions:
              - key: app
                operator: In
                values:
                - web-frontend
topologyKey: "kubernetes.io/hostname"
podAntiAffinity:
...
```

Should I run my Pod in the same hostname as a web-frontend Pod?

Topologies in Pod (Anti-)Affinity




node-4
pod-a
node-1
pod-b
failure-domain.k8s.io/zone=us-east-1a
node-3
node-2
failure-domain.k8s.io/zone=us-east-1b


node-4
pod-a
node-1
pod-b
custom_topology=foo
custom_topology=bar
node-3
node-2
custom_topology=foo
custom_topology=baz

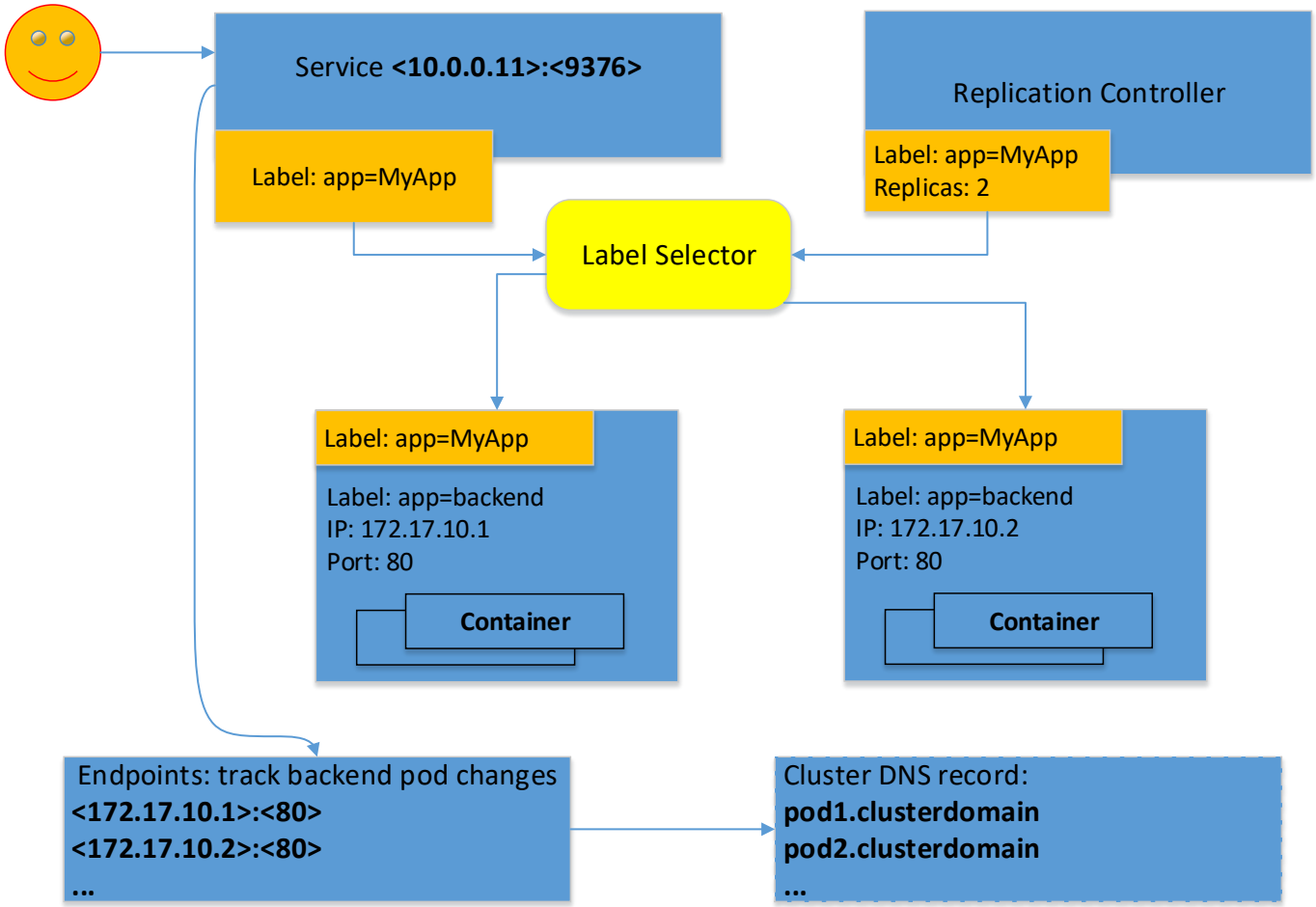
Supported topology-aware features in Kubernetes

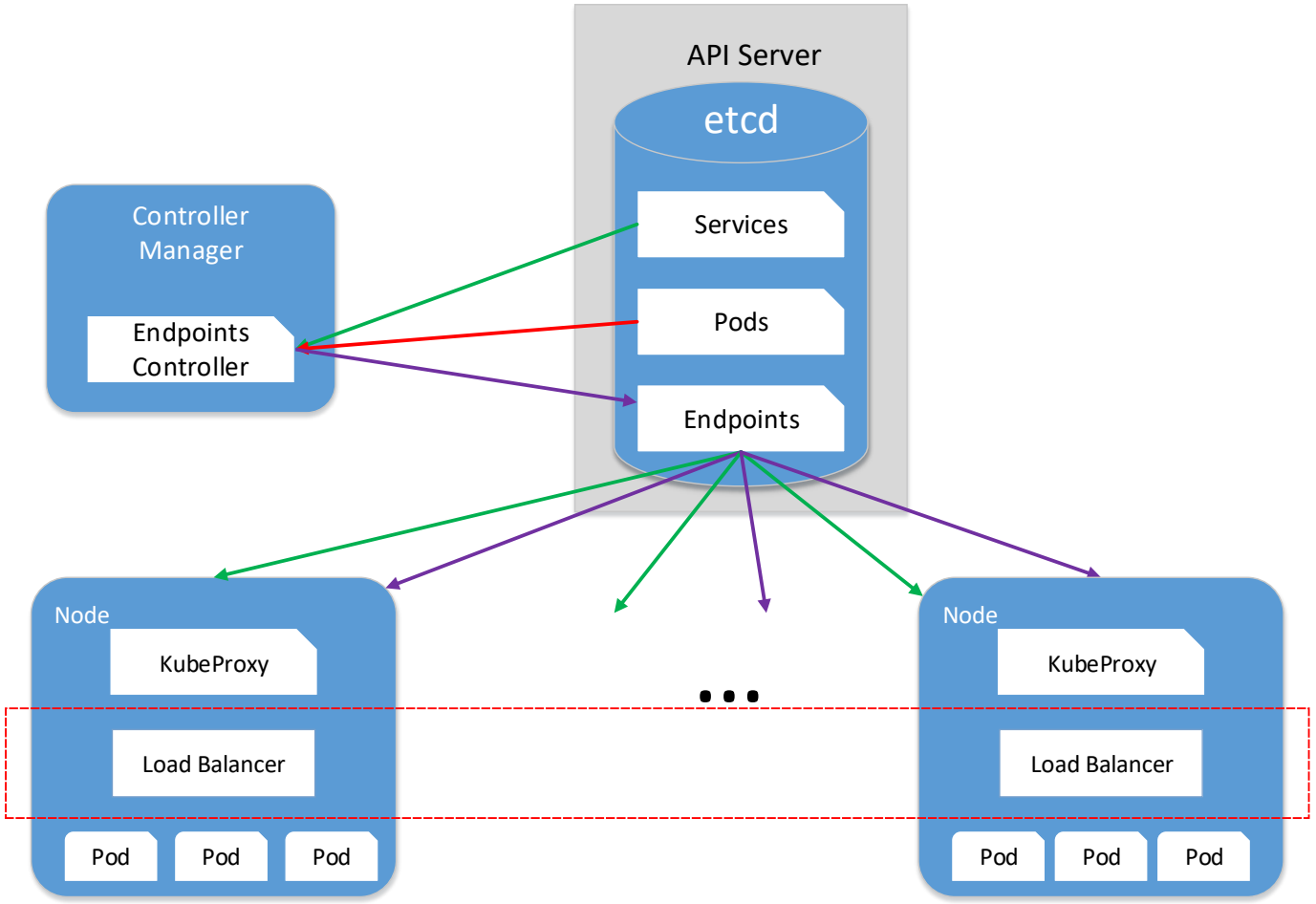
- Node level
  - Workloads
  - Volumes
- Within a node
  - Hardware

Agenda

- Topologies in Kubernetes
- Requests of topology-aware service routing
- Solutions and prototypes
- Q&A

Kubernetes Service & Endpoints




API Server
etcd
Services
Pods
Endpoints
Controller Manager
Endpoints Controller
Node
KubeProxy
Load Balancer
Pod
Pod
Pod
Node
KubeProxy
Load Balancer
Pod
Pod
Pod
iptables
IPVS

Topology-aware service routing: user stories

- Clear demand for node-local
  - per-node services: fluentd, aws-es-proxy
  - secure
- “Find zone-local backends for service X”?
  - data costs
  - performance
- Extend: “locality” means same topological level
  - select a subset of endpoints based on topology

Topology-aware service routing: problem statements

- Hard requests or soft requests?
  - try local, then go wider?
  - always want that one?
- How hard to try?
  - weight per topo
- What if multiple backends satisfy?
  - probabilities

Agenda

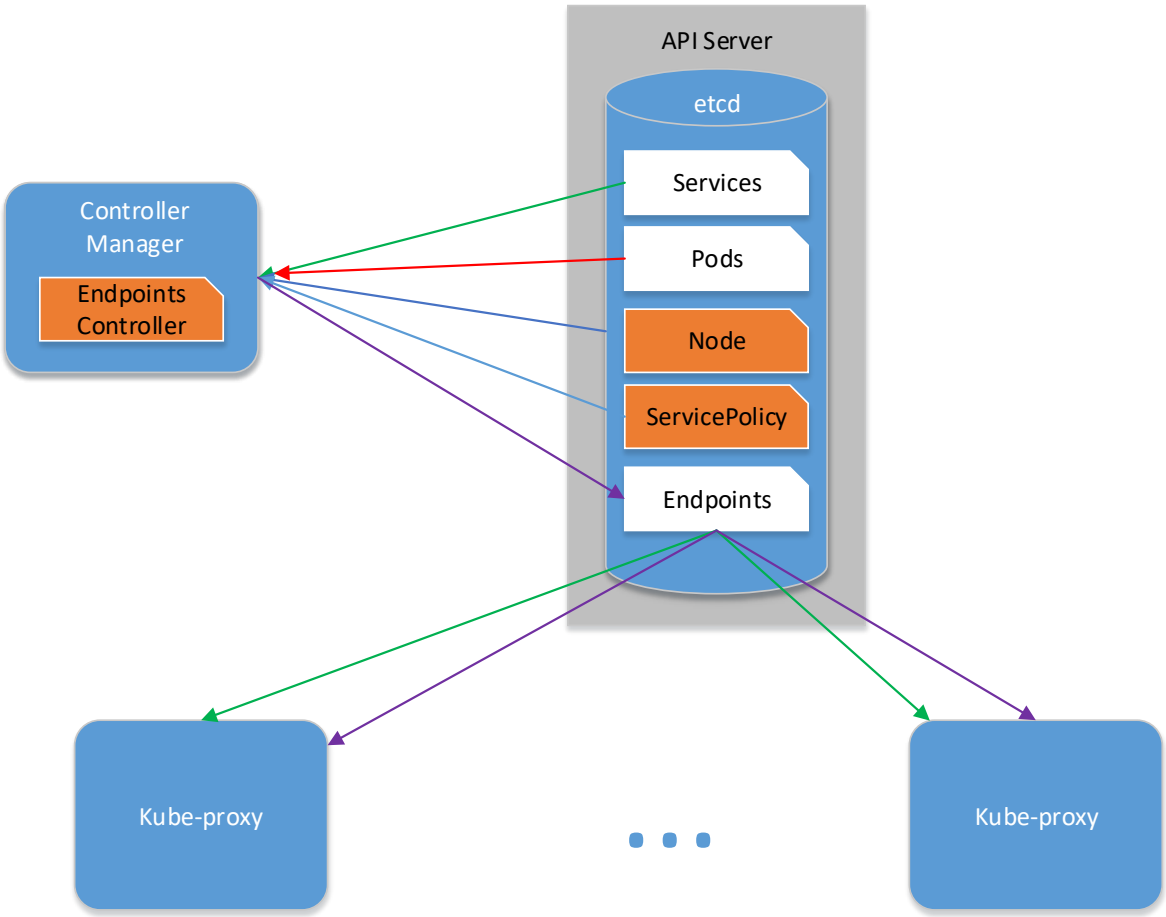
- Topologies in Kubernetes
- Requests of topology-aware service routing
- Solutions and prototypes
- Q&A

Solutions and prototypes

API Changes:

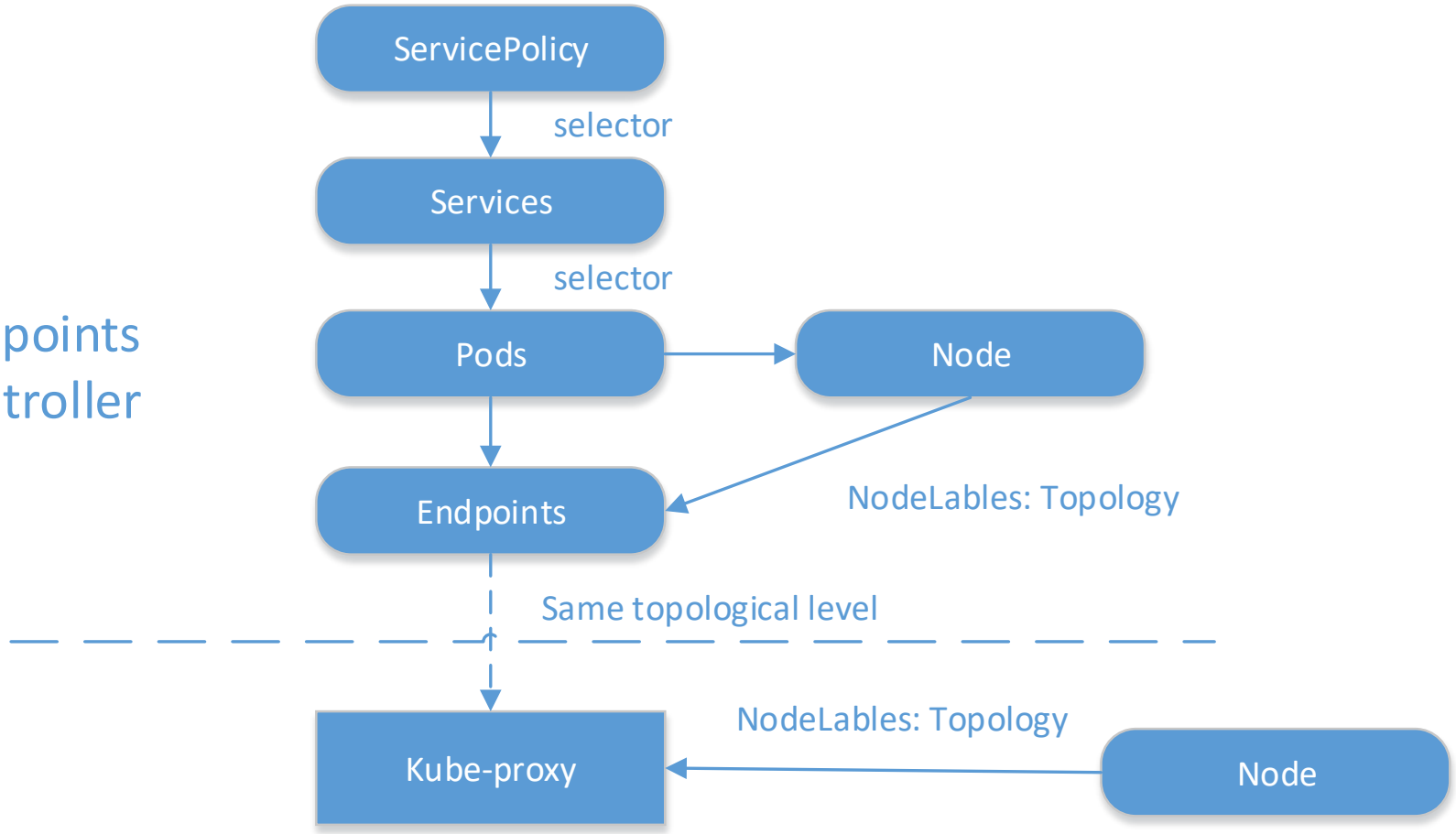
```
kind: ServicePolicy
metadata:
  name: service-policy-example
  namespace: foo
spec:
  serviceSelector:
    matchLabels:
      app: bar
  topology:
    key: kubernetes.io/hostname # Any topology key you want
    mode: required/preferred/ignored
---
# Endpoints API changes
type EndpointAddress struct {
  // labels of node hosting the endpoint
  Topology map[string]string
}
```

Architecture



Data Flow

Endpoints Controller



Topology-aware service routing

- Running well in Huawei Cloud CCE

- Happy to open source the implementation

  - Proposal: https://github.com/kubernetes/community/pull/1551

Contact US!


LINUXCON
containercon
CLOUDOPEN
CHINA 中国
THINK OPEN
开放性思维
LF ASIA, LLC