字节跳动在 k8s 的性能优化实践

陈逸翔

陈逸翔

就职于字节跳动基础架构团队，从16年开始参与PaaS平台的起步建设工作。

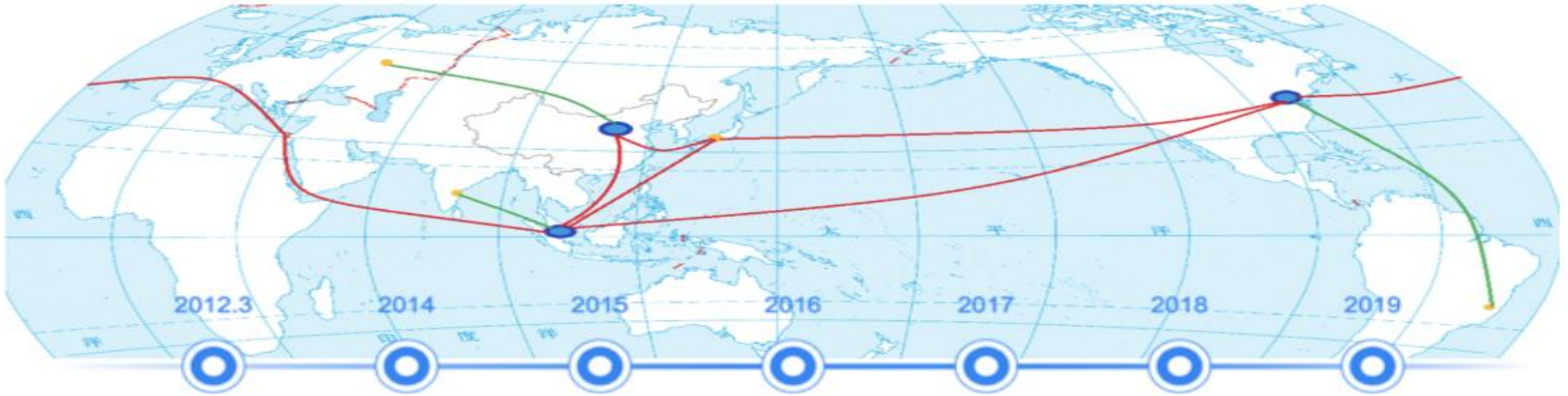
主要负责 Kubernetes 及其周边开发工作，包括但不限于 kubernetes、监控、容器等工作，支持字节跳动内部 PaaS 平台。

目前专注于内部 Kubernetes 集群的大规模场景上的性能优化，包括APIServer、scheduer 和 controller-manager，支持业务的快速发展带来的规模性的挑战。

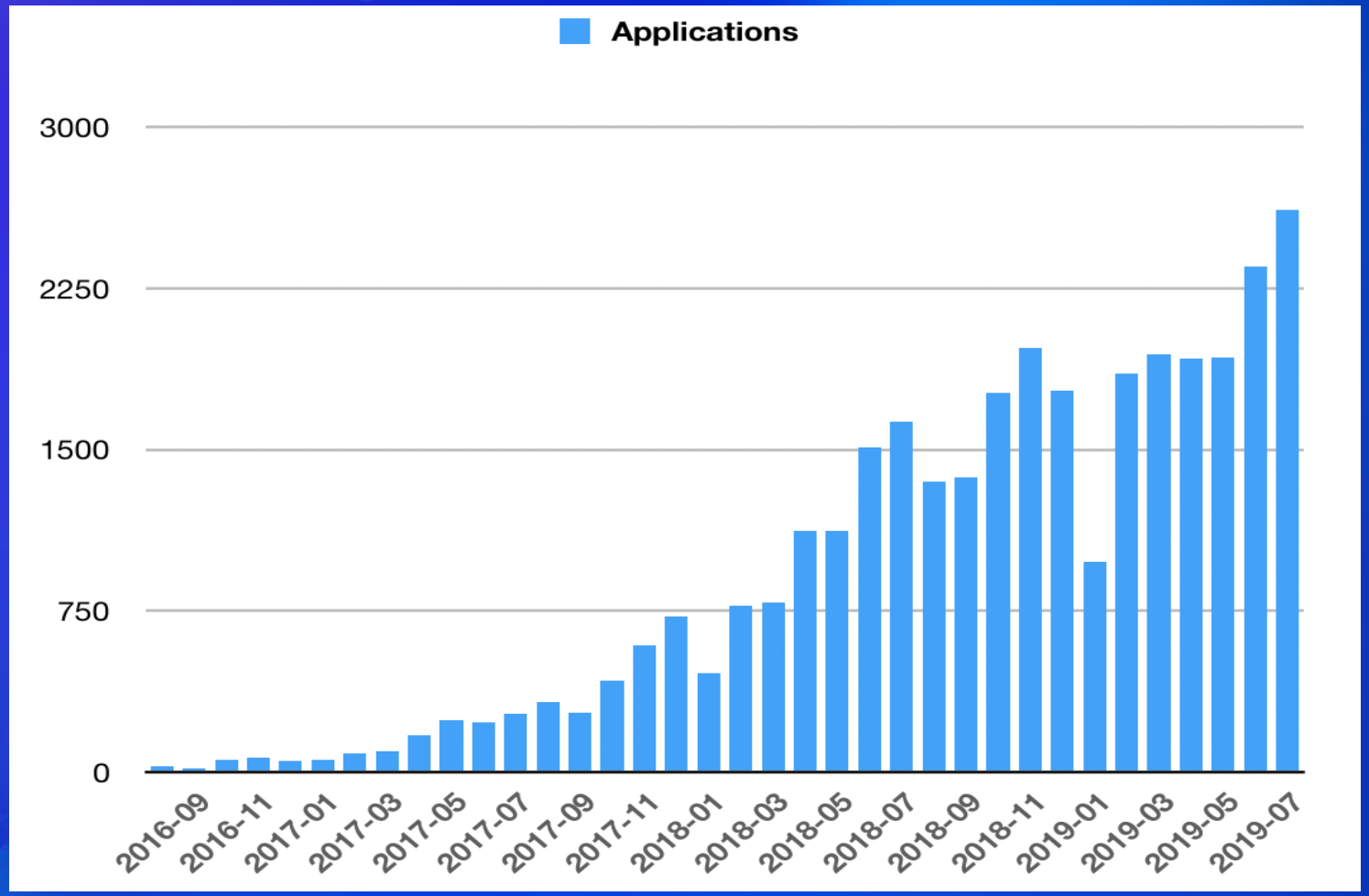
目录

Introduction

国际化业务


2012.3
2014
2015
2016
2017
2018
2019
ByteDance founded, Toutiao launched
Launched first video platform, which was later rebranded
Launched first overseas product and started building content creator community
Launched UGC video platforms
Acquired musical.ly and News Republic, launched TikTok
International expansion accelerated, first education product
Helo gained traction, ByteDance moved into enterprise space
ByteDance
头条

快速增长的微服务



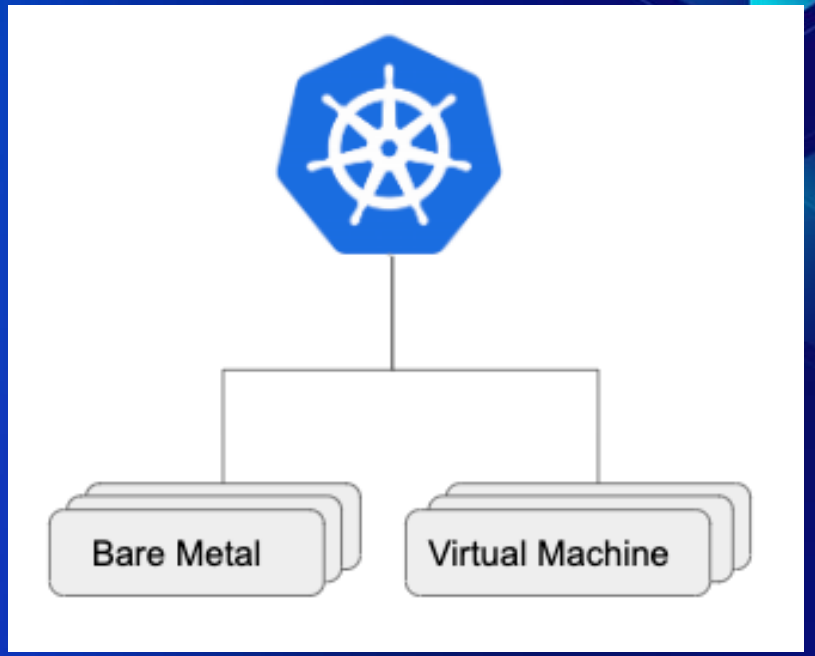
- 超过 20k 的微服务
- 每月新增 2k 的微服务
- 运行容器 1 million+

Kubernetes in ByteDance

集群现状

生产环境

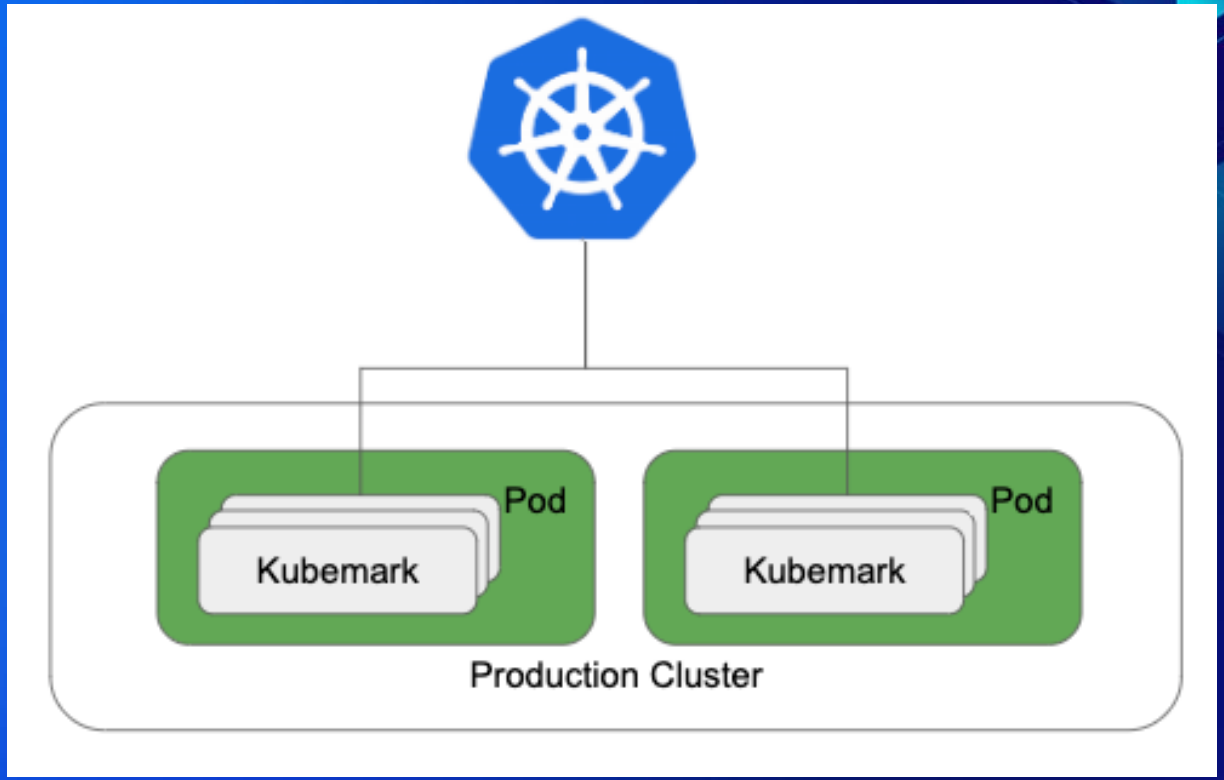
- 1,000,000+ Pod
- 60,000+ Node
- 70+ 集群
- 6,000+ Node/largest cluster



集群现状

压测环境

- 150,000+ Pod
- 10,000+ Node
- < 3s latency



组件性能和稳定性优化

遇到的问题

- APIServer/ETCD
  - Master 重启时 etcd slow range query 过多
  - 弱网环境优化

遇到的问题

- Scheduler/Controller Manager
  - 调度器调度慢
  - 大服务上线集群变慢

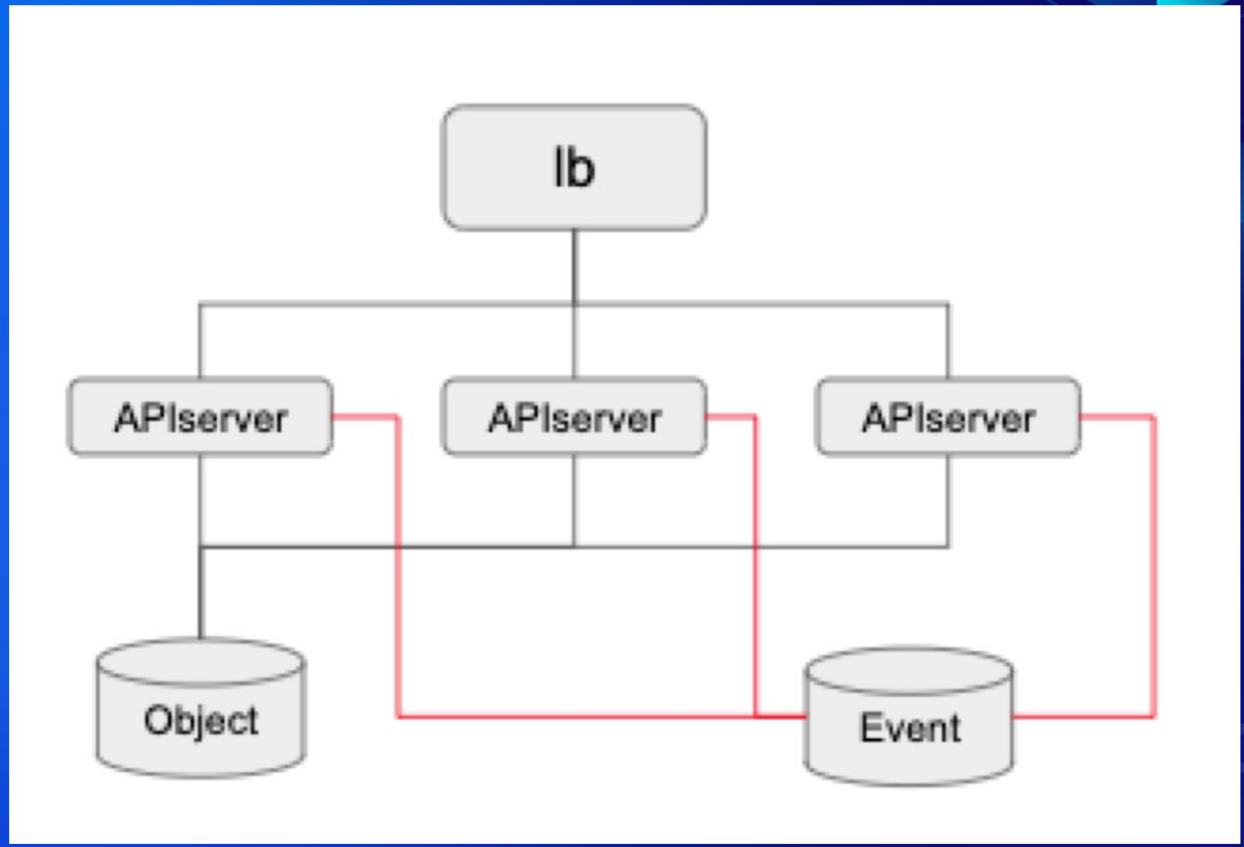
遇到的问题

- Kubelet
  - 单机压力过大
  - Host 网络模式端口冲突

ETCD

Object 拆分

- SSD 部署
- 200k+ Events
- Lease 问题



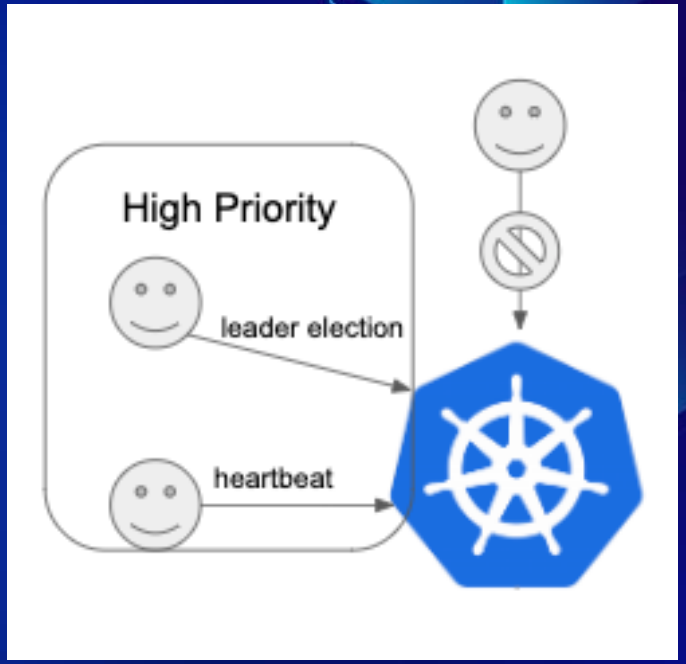
ETCD

range query

- APIServer 对 ETCD 的 List 请求进行分页请求
- 原生的 range API 设计实现比较低效
- 增加了 no total 的参数避免 range 所有 key
- limit range 1k in 2m kv: 7min->7s

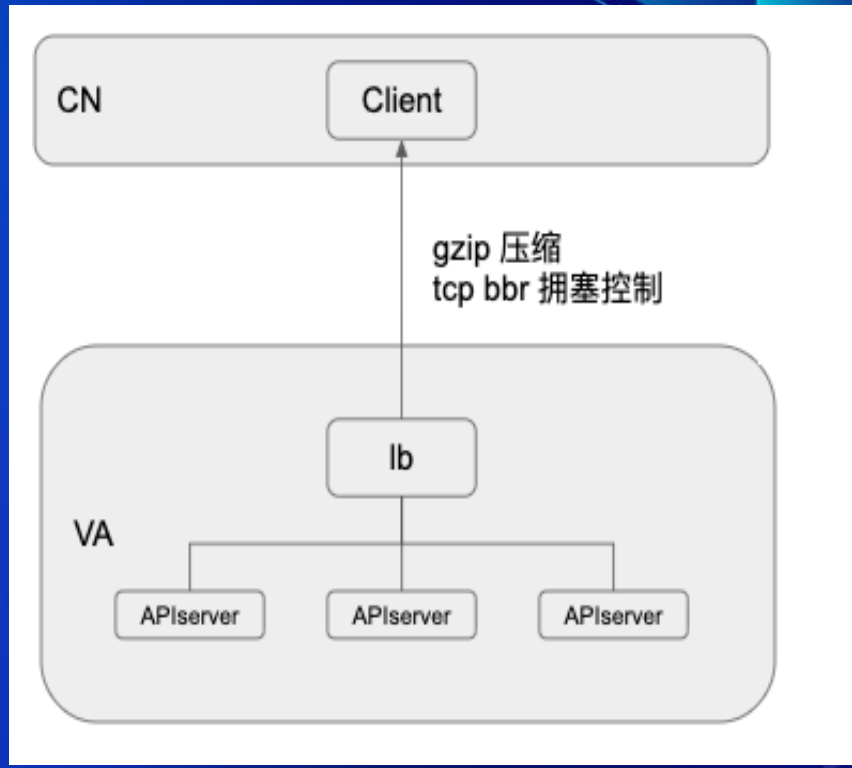
APIServer

- **request priority**
  - max-requests-inflight && max-mutating-requests-inflight
  - Too many request 后拒绝请求
  - kubelet 心跳断开失联 & 组件 leader lease lost
  - 定义高优请求豁免限制



APIServer

- **国际化**
  - 跨地区网络硬件限制，List 超时
  - Gzip 压缩PodList后，跨地区传输数据 400 MB -> 40 MB
  - bbr 优化后传输时间减少一半



Controller Manager(KCM)

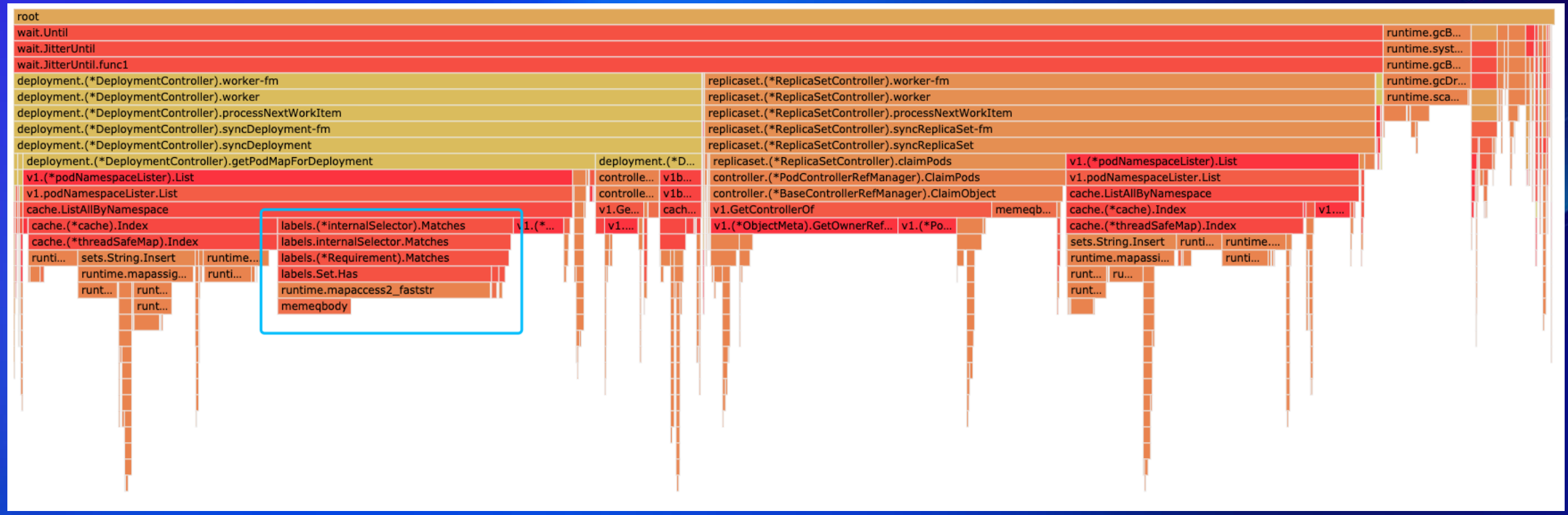
•横向拆分

- 单个 controller 负载过高影响核心 controller
- 核心 controller 拆分单独部署
- 支持配置 leader-election 对象，实现多个 controller 高可用

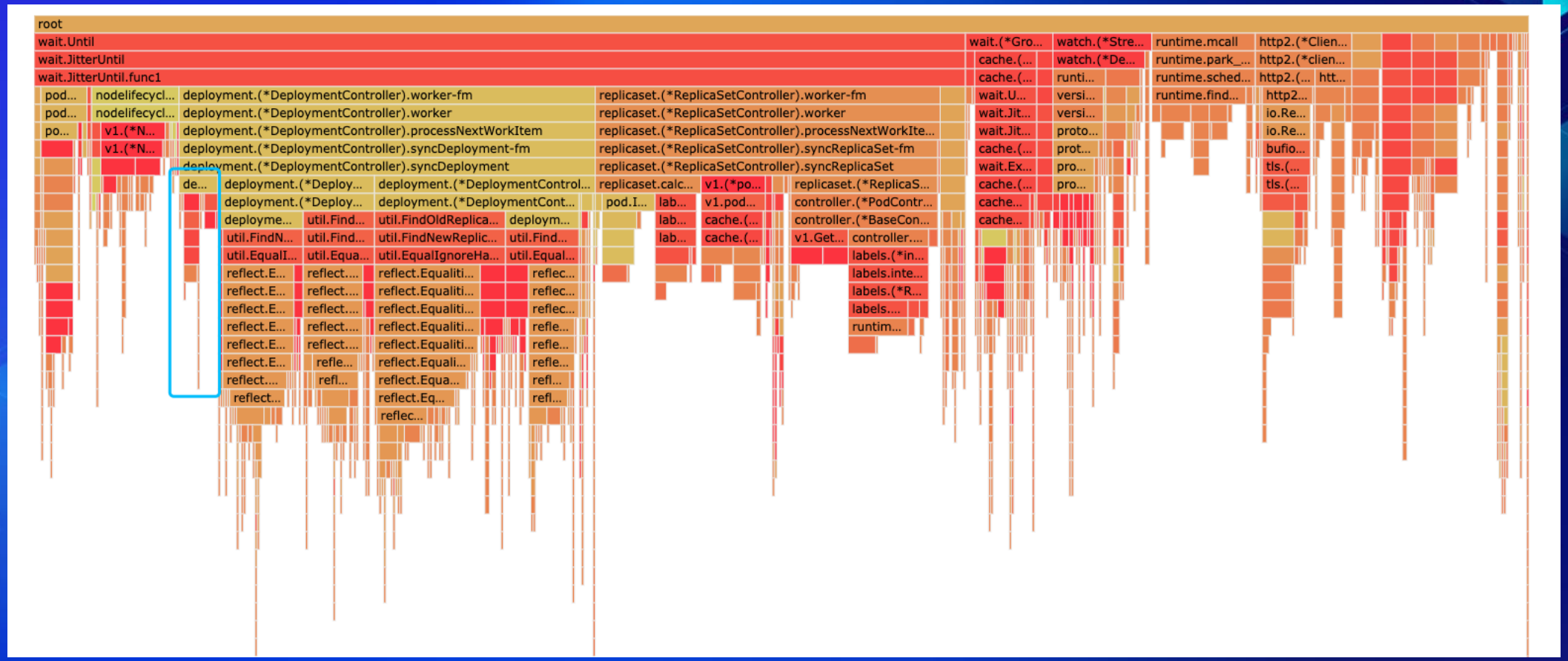
Controller Manager(KCM)

- **Informer cache indexer**
  - Deployment & RS controller 经常看起来不工作
  - Deployment & RS controller sync 的时候需要 list namespace 的 pod 后 match label 导致过载
  - 创建的 label 中增加 name={{metadata.name}}，在 informer 中增加该 label 的 index
  - sync duration pct99: 100ms -> 10ms

优化前： Deployment Controller 48%，主要消耗在 getPodMapForDeployment



优化后： Deployment Controller 27%



Scheduler

Hostuniq

服务单机最多部署X个实例

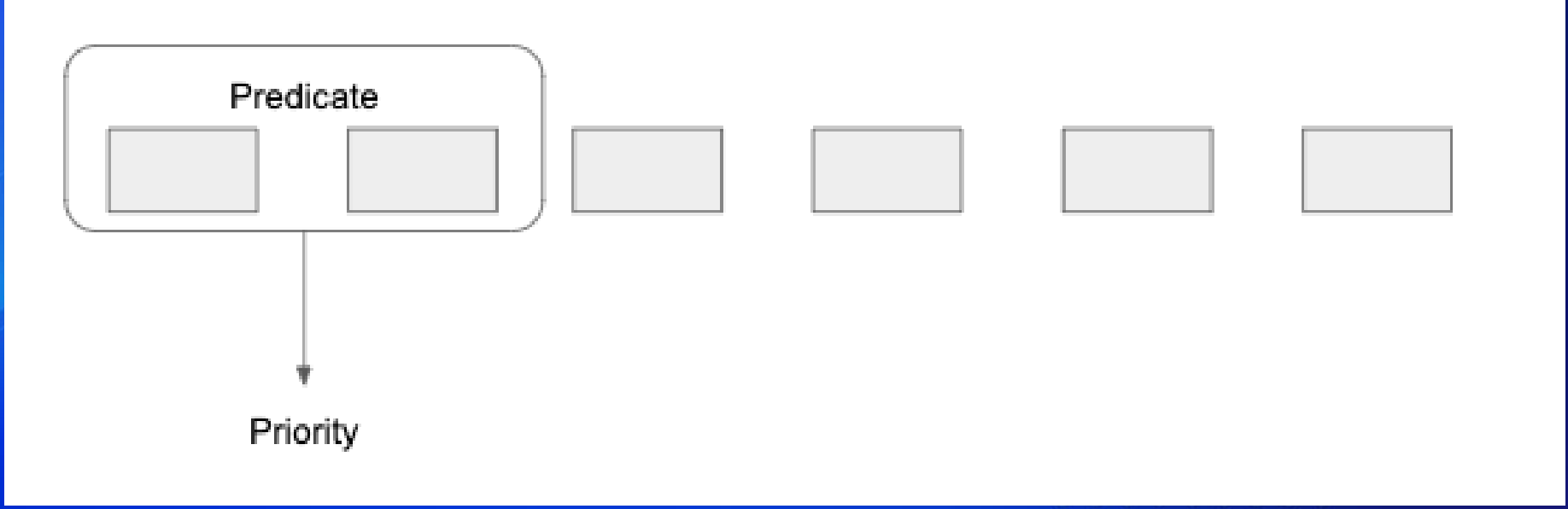
self anti affinity 效率太低，且没有soft threshold

EvenPodsSpread

Schedule latency：5s->100ms

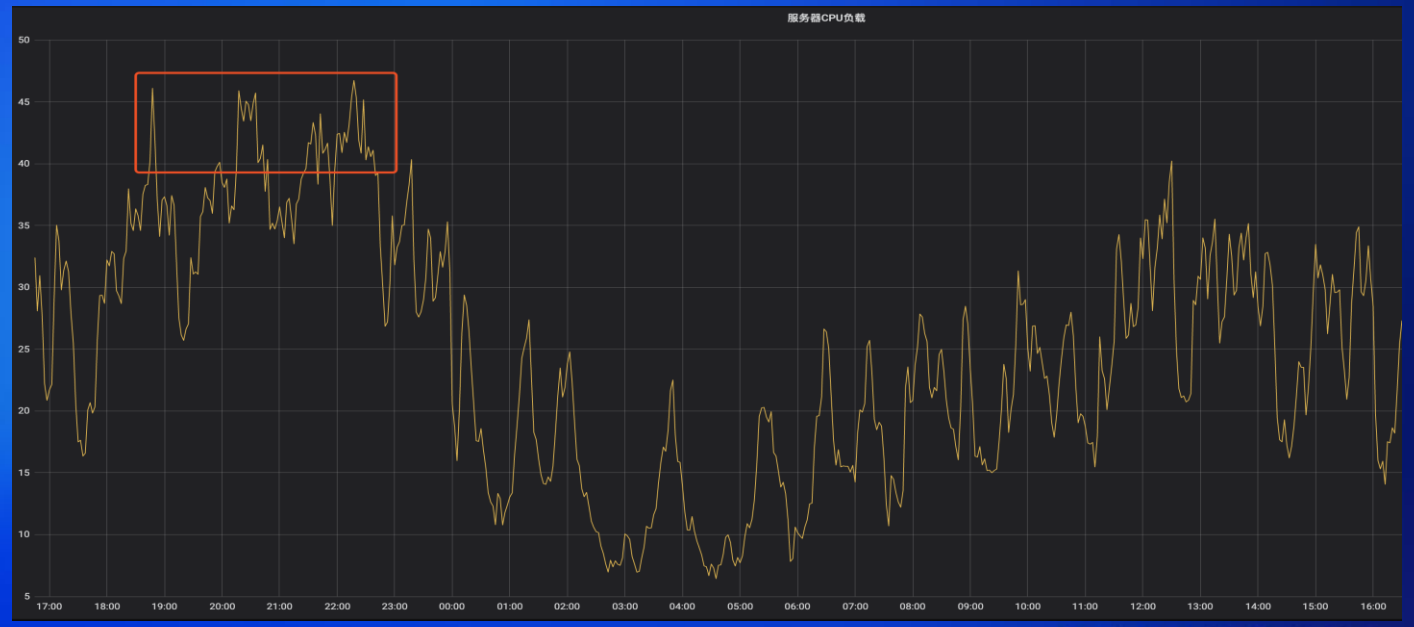
Scheduler

- 抽样调度
  - 不需要遍历所有的Node，遍历部分节点尝试调度预选，成功后再进行优选
  - 缓存调度结果，同一个 Deployment 下的 Pod 可以跳过预选



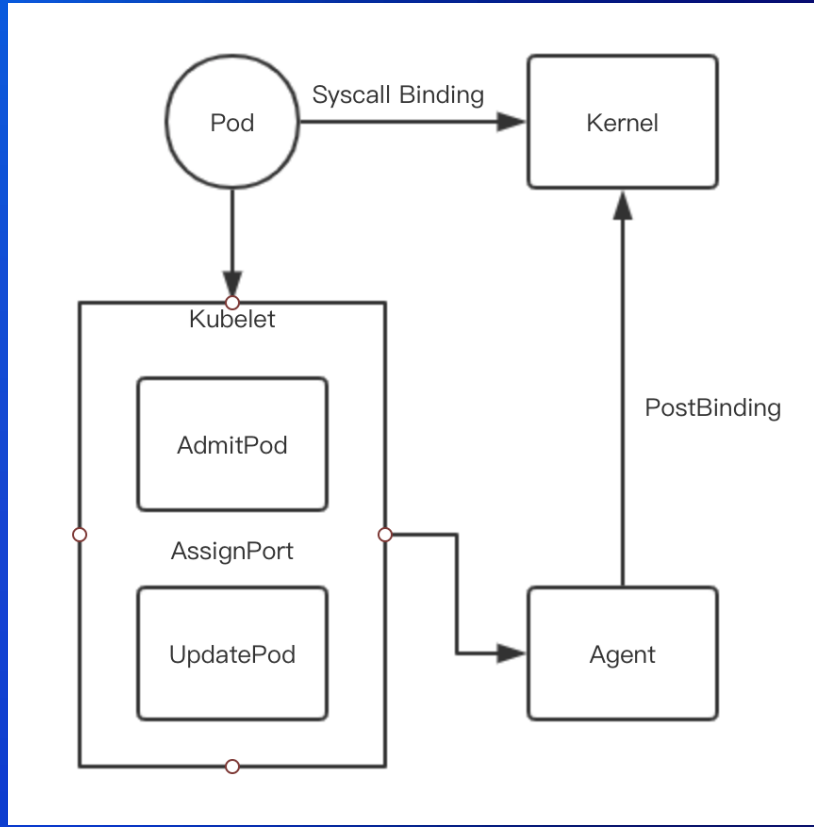
Kubelet

- Load eviction
  - Load = running + uninterruptible
  - 直观评估机器负载指标
  - Soft eviction & Hard eviction



Kubelet

- Port assign
  - Host 模式解决端口冲突
  - PodAdmit 时进行端口分配
  - Bpf post-bind hooks



Q&A

Thanks For Watching

本PPT来自2019携程技术峰会

更多信息请关注“携程技术中心”微信公众号~