Edge Kubernetes - 构建边缘云计算基础设施

高步双（花名：江博）

2019-03-16

扫码关注公众号,获取 316 北京站 PPT

About Me

阿里云
奥运会全球指定云服务商

高步双（花名：江博），阿里云容器服务技术专家，2017年加入阿里，主要从事 Kubernetes 相关产品的设计研发工作，先后负责或参与资源调度系统稳定性分析引擎、Kubernetes&YARN 混部、EDAS Kubernetes版本和 Serverless 版本、Edge Kubernetes、安全容器 等相关产品、项目的设计和研发工作。

1 边缘云计算

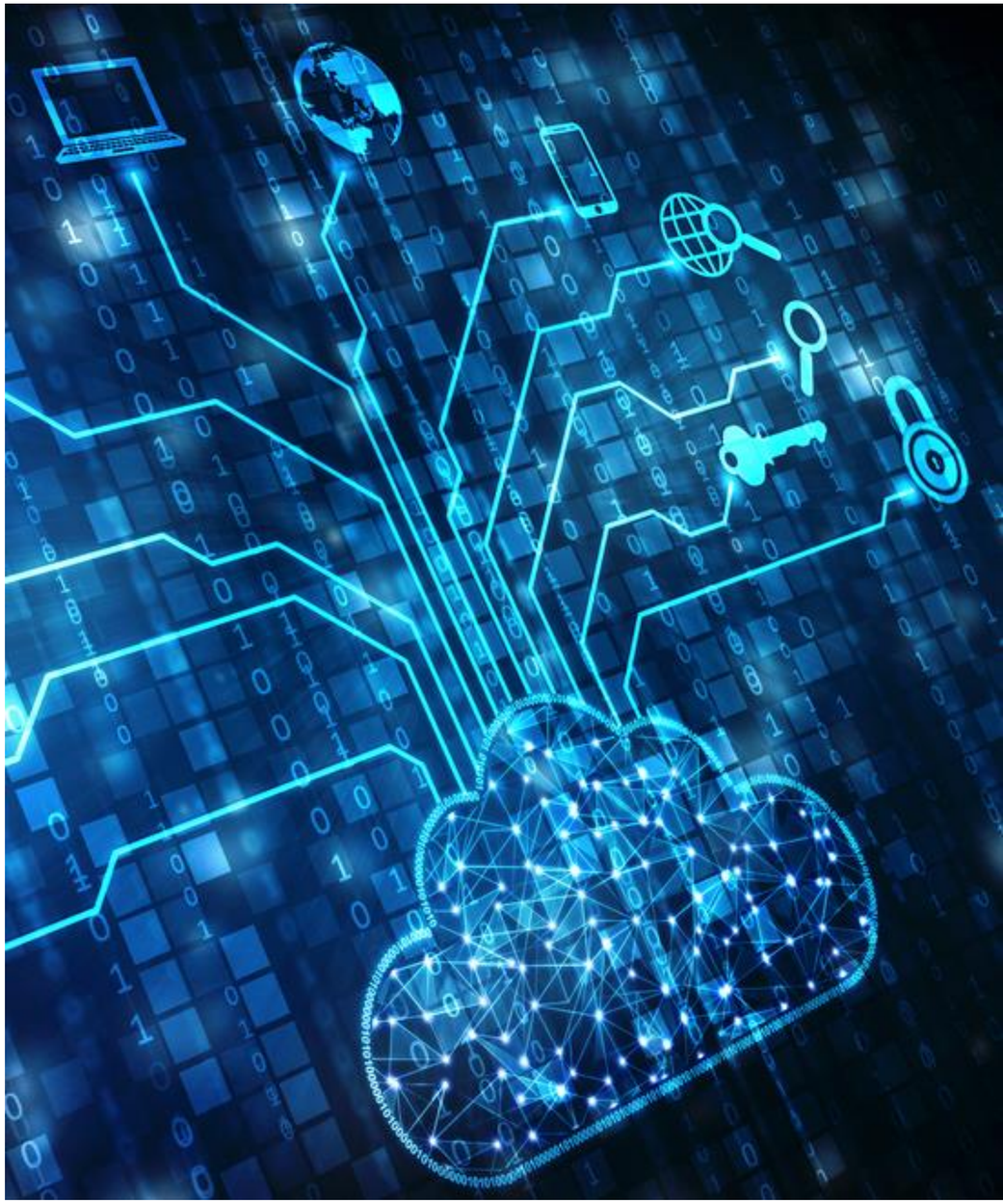
云计算

阿里云 奥运会全球指定云服务商

云计算，是一种将可伸缩、弹性、共享的物理和虚拟资源池以按需自服务的方式供应和管理，并提供网络访问的模式。

特点：

1. 计算资源集中、规模庞大。
2. 高可用性：基于分布式服务器集群设计。
3. 虚拟化：计算、存储、网络资源虚拟池化共享。
4. 可扩展性：根据用户需求灵活动态分配或释放资源。



边缘计算

阿里云 奥运会全球指定云服务商

边缘计算是一种将主要处理和数据存储放在网络的边缘节点的分布式计算形式。- ISO/IEC JTC1/SC38

“边缘”相对概念，指从数据源到云计算中心路径之间的任意计算、存储和网络资源。

相对传统云计算：数据上传到云端，通过云端的资源完成存储和计算，但随着万物互联时代的到来，智慧城市、智能制造、智能交通、智能家居，5G时代、宽带提速、IPv6的不断普及，导致数百亿的设备接入网络，在网络边缘产生ZB级数据，传统云计算难以满足物联网时代大带宽、低时延、大连接的诉求。

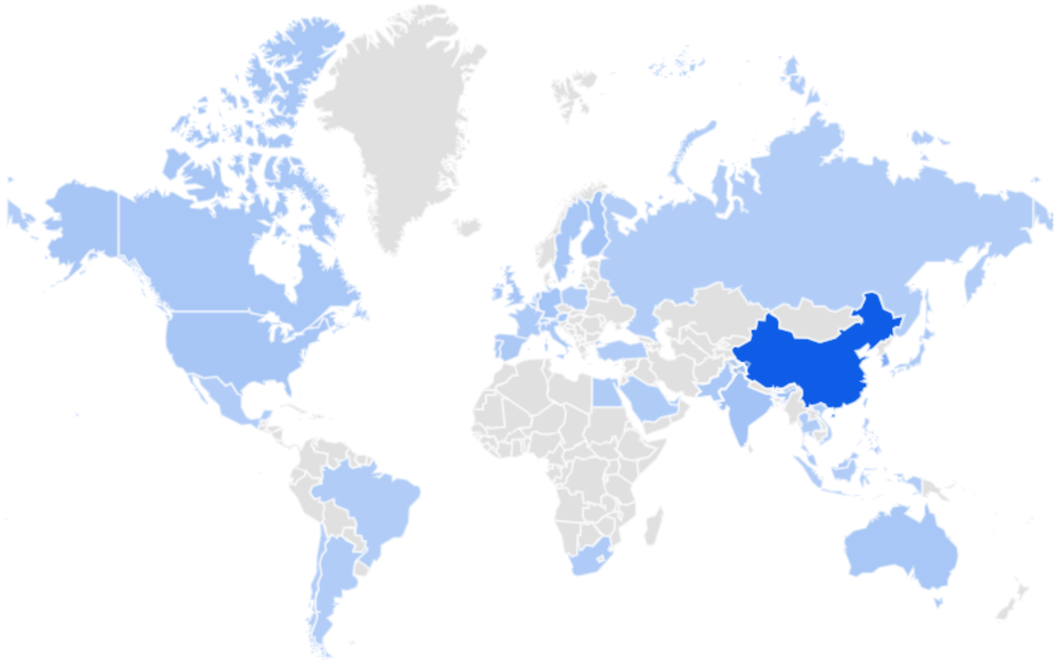
边缘计算关注度(Google Trends)

关键字: edge computing

关键字: 边缘计算

边缘计算关注度-热力图

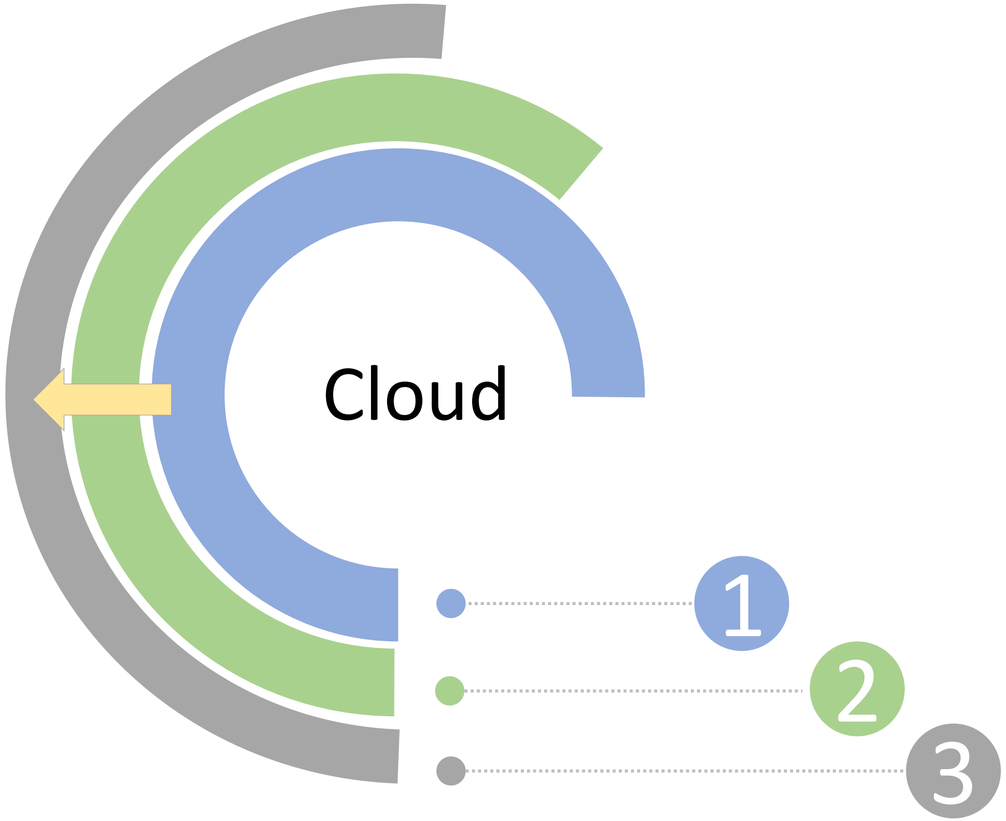
阿里云 奥运会全球指定云服务商



1	中国	100
2	韩国	31
3	台湾	20
4	新加坡	18
5	香港	12

边缘云计算 – “云边端三体协同”

阿里云 奥运会全球指定云服务商



“大带宽、低时延、大连接”

概念

边缘云计算是基于云计算技术的核心和边缘计算的能力，构筑在边缘基础设施之上的云计算平台。形成边缘位置的计算、网络、存储、安全等能力全面的弹性云平台，并与中心云和物联网 终端形成“云边端三体协同” 的端到端的技术架构，通过将网络转发、存储、 计算，智能化数据分析等工作放在边缘处理，降低响应时延、减轻云端压力、降低带宽成本，并提供全网调度、算力分发等云服务。(引自: 边缘云计算技术及标准化 白皮书(2018))

① **“云”**

中心云，又称传统云计算。管控端，全网算力统一管理与调度，与边缘协同。

② **“边”**

“边”位于边缘云计算中的边侧。靠近设备和数据源的计算资源，用于部署边缘侧应用。可能是云厂商也可能是用户自己的边缘节点。

③ **“端”**

“端”位于边缘云计算中的端侧，即设备端。跟据 Gartner 的报告，到2020年全球连接到网络的设备将达到约208亿台。

边缘云五要素

阿里云 奥运会全球指定云服务商

1. 低时延

边缘云计算可就近提供网络覆盖、数据存储和计算能力，设备数据的产生和计算可低时延就近处理.

2. 边缘自治

在中心云端管控网络断连的情况下，边缘节点可实现自治理，保证边缘侧应用的正常运行。

3. 资源管理

云端管控可灵活调度边缘应用运行在指定边缘侧。

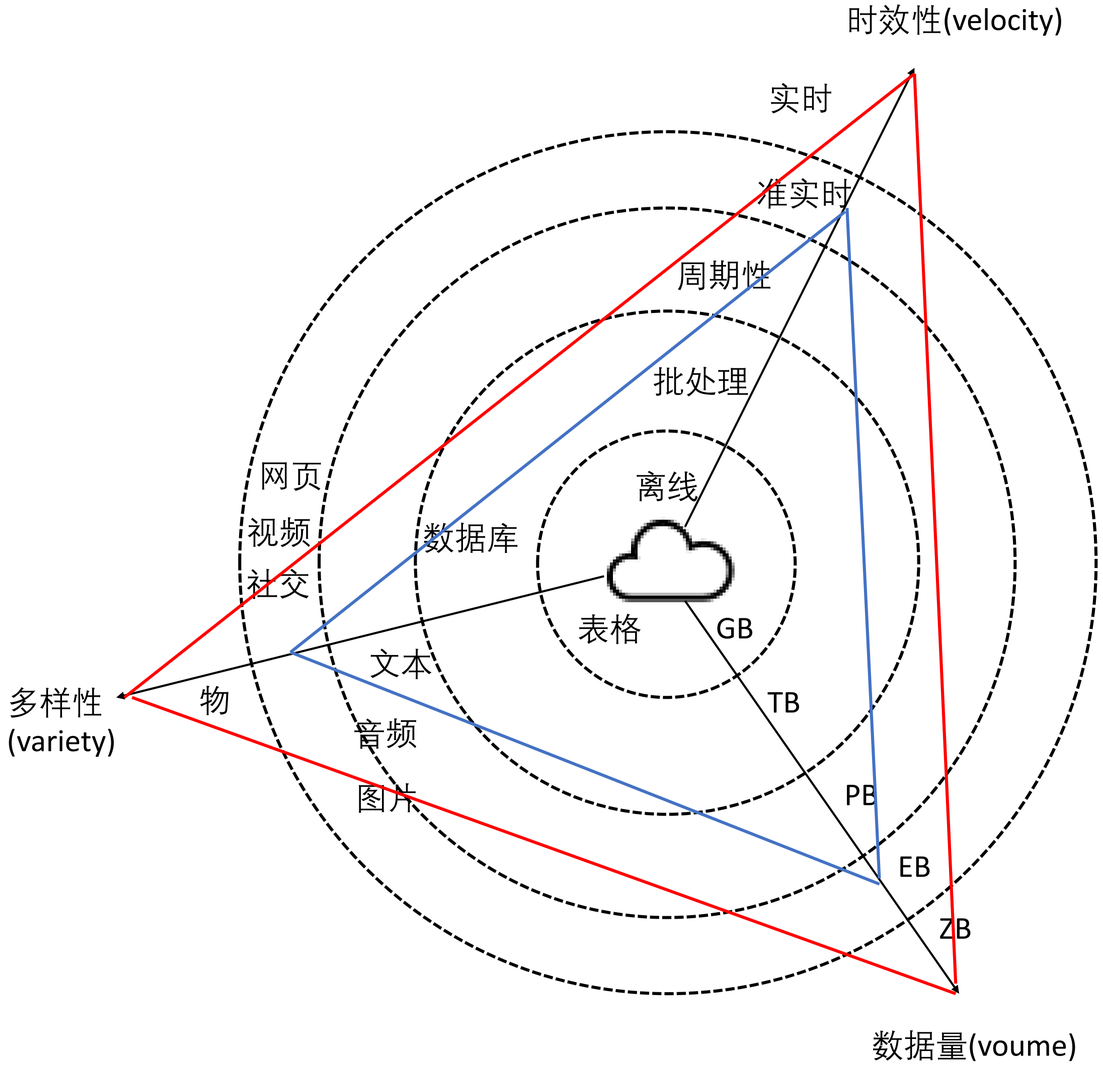
4. 安全性

网络安全、数据安全、租户隔离、安全容器。

5. 标准化

统一且开放的服务接口。

机遇&挑战：中心云 vs 边缘云



中心云
1. 数据类型：文本、音视频、图片和结构化数据库等。
2. 量级：PB。
3. 实时性：要求低，多以离线为主。

边缘云
1. 数据类型：+百亿计联网设备产生的海量数据。
2. 量级：ZB。
3. 实时性：要求高

场景

阿里云 奥运会全球指定云服务商

01

IoT

边缘云计算把数据、数据处理和应用程序集中在边缘设备中，通过中心云端实现智能的流动和分配，赋予广泛的边缘设备直接存储数据和处理任务的能力，让物联网成为真正的智能物联网。

02

自动驾驶

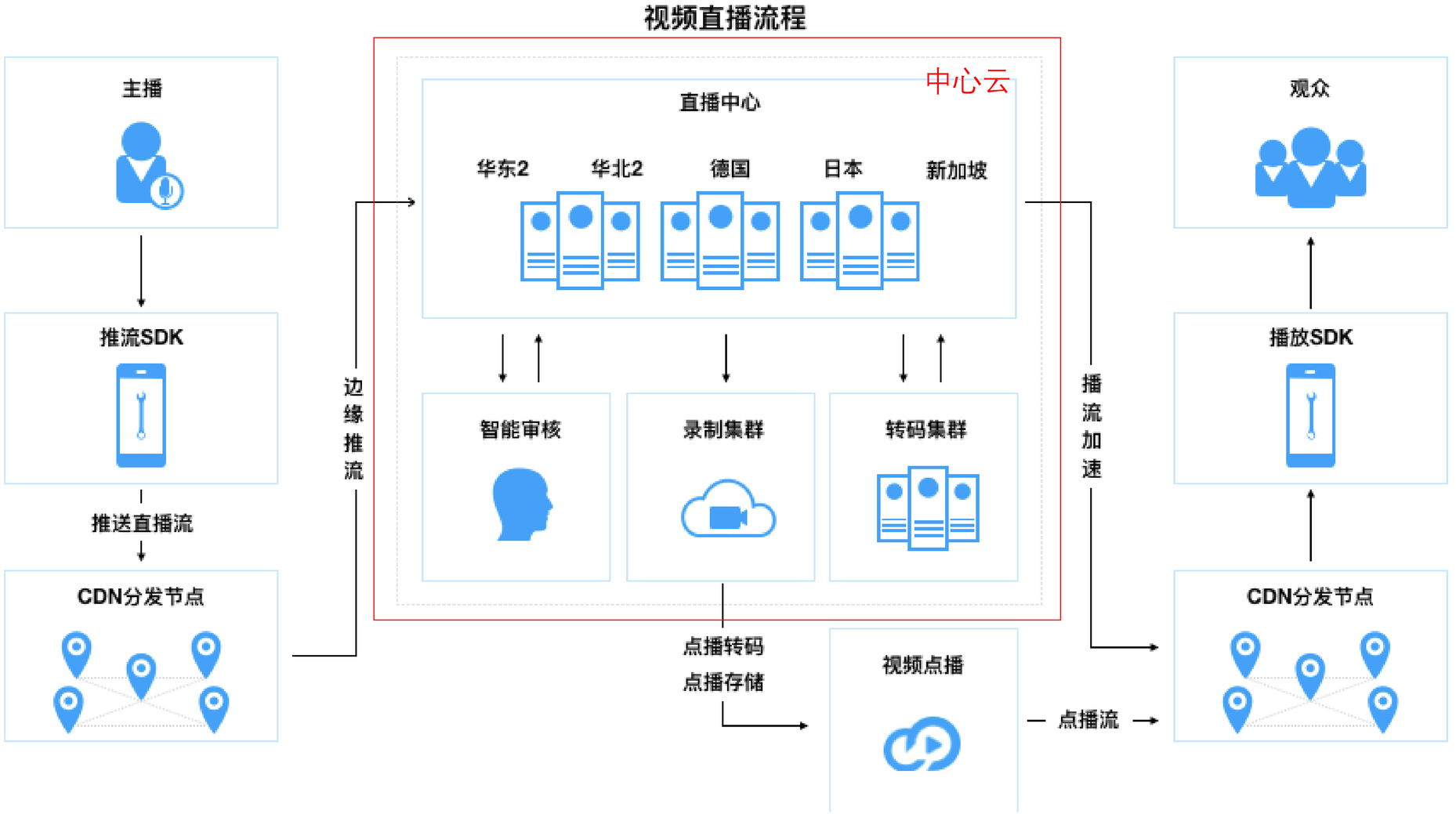
自动驾驶汽车可以自我学习，通过传感器、视频采集数据并处理，并可以把诸如交通状况信息通过边缘节点共享给其他汽车。

03

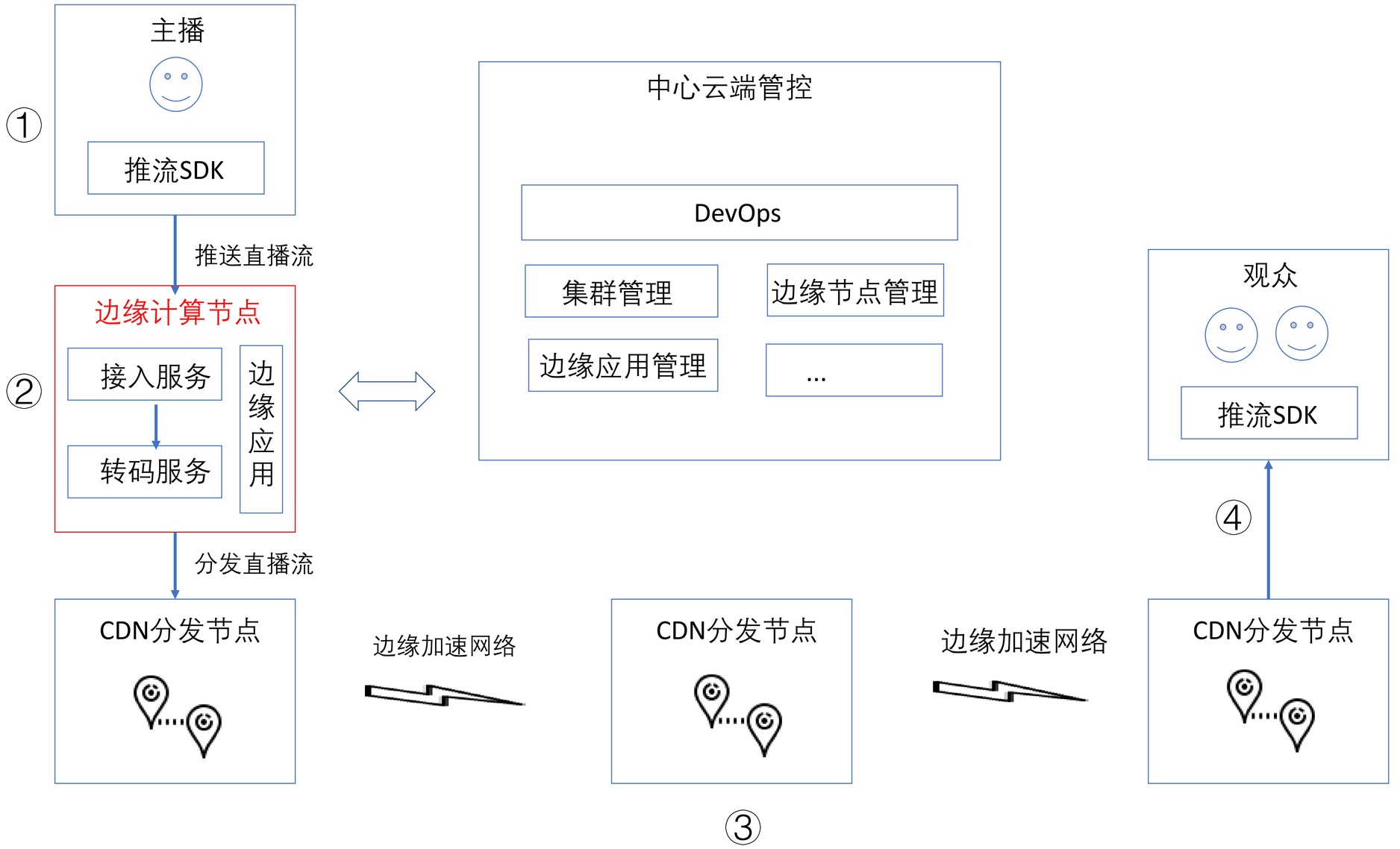
视频直播

主播媒体流直接在边缘节点完成转码，转码后并分发到CDN边缘节点。在主播直播推流时，实现就近节点进行转码和分发，同时支 持高并发实时弹幕的边缘分发，减少了对中心的压力。

场景案例：视频直播互动传统解决方案



视频直播互动场景-边缘云计算解决方案



大带宽、低时延、大连接

1. 在主播和CDN之间加入边缘计算节点。
2. 直接在就近边缘节点完成直播流接入、转码以及就近CDN节点分发。
3. 媒体流不再路经中心云端，无中心瓶颈，时效性提高，带宽和存储成本降低。
4. 管控端：边缘节点和边缘应用的管理等。

2 基于 Kubernetes 打造“云边端”一体化边缘云基础设施

服务治理发展史

阿里云 奥运会全球指定云服务商

传统 PaaS

“原始社会”...

容器时代

- 颠覆传统PaaS
- 微服务
- ...

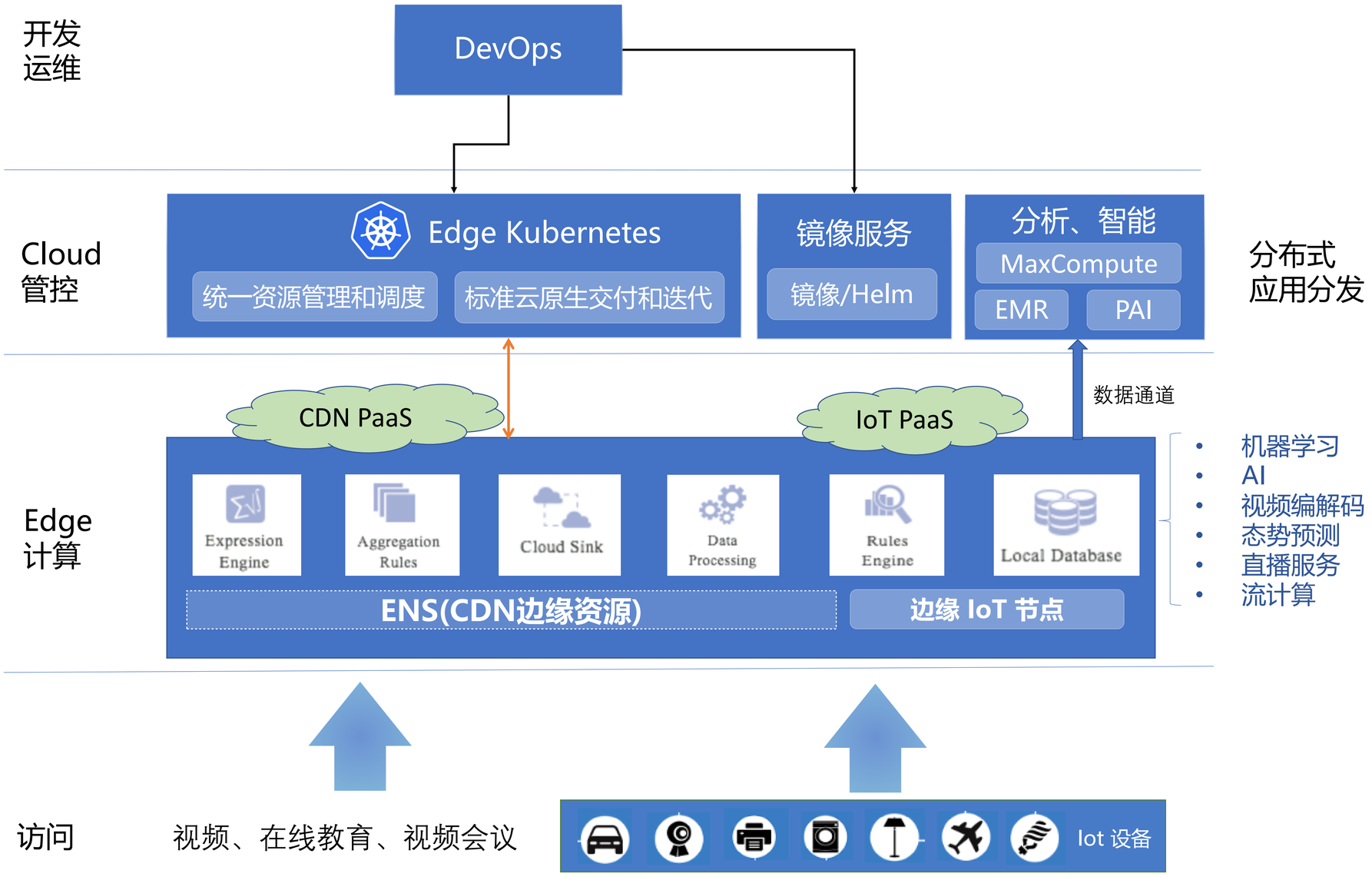
K8s时代

- 探索
- 用起来
- 生产检验
- 生态
- ...

后 K8s 时代

- 服务网格
- Serverless
- 边缘计算
- 基因计算
- HPC
- AI
- ...

边缘容器打造云原生云-边一体



- 价值
  - 贴近数据源，降低数据生产和决策之间的延迟；
  - 降低集中计算的成本；
  - 一站式边缘容器方案，提高边缘应用的可运维性；
  - 提供弱网环境下的边缘自治能力
- 形态
  - 作为构建IOT PaaS、CDN PaaS的底座；
  - 面向EndUser的edge-k8s 托管服务；
  - 二方云产品服务接入、能力下沉通道；
- 功能
  - 支持ENS资源调度；
  - 延续容器托管服务形式，以云原生的方式打造“云-边”一体；
  - 支持边缘计算力的快速接入；
  - 在离线或间歇性连接状态下边缘节点的自治；
  - 支持边缘容器网络、网络策略；

集群架构

阿里云 奥运会全球指定云服务商

云端
(管控端)

K8s Master

K8s Master

K8s Master

边侧
(边缘节点)

边缘区域1

节点1

Kubelet

节点2

Kubelet

边缘区域2

节点3

Kubelet

节点4

Kubelet

端侧
(设备)


camera


sensor


car


ship


express


oil


hospital


factory

...

1. 标准化
2. 统一管控
3. 低时延
4. 安全性
5. 边缘自治

仍要解决的问题

阿里云 奥运会全球指定云服务商

节点自治

网络长期断连。
节点POD自动迁移。
节点异常或重启，内存数据丢失。

多租隔离

共享集群的模式下，保证不同租户之间资源的有效隔离。

网络自治

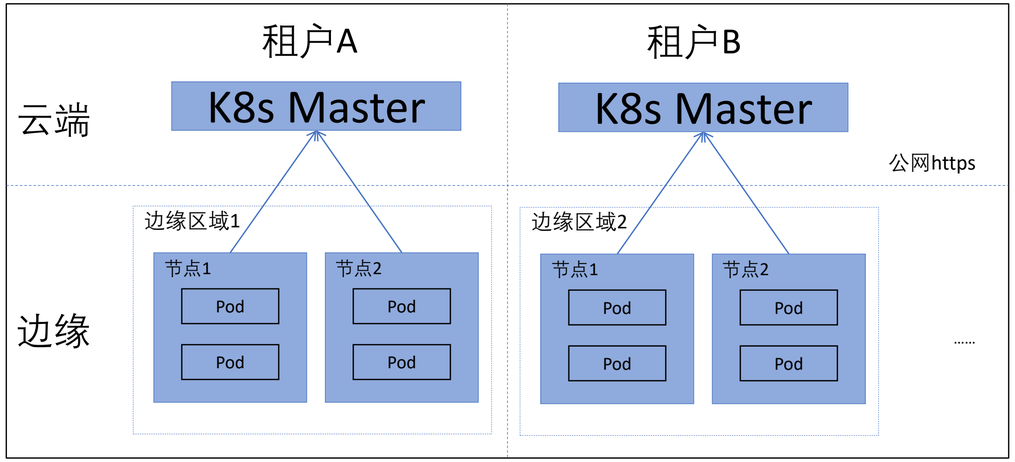
管控断网后边缘侧应用、服务和组件的影响。

标准 K8s API 交付

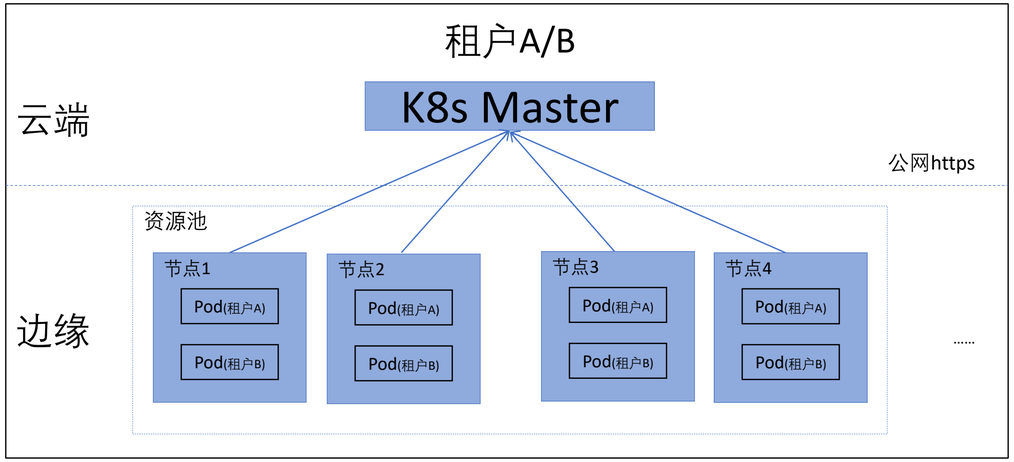
多租隔离 - 场景模型

阿里云 奥运会全球指定云服务商

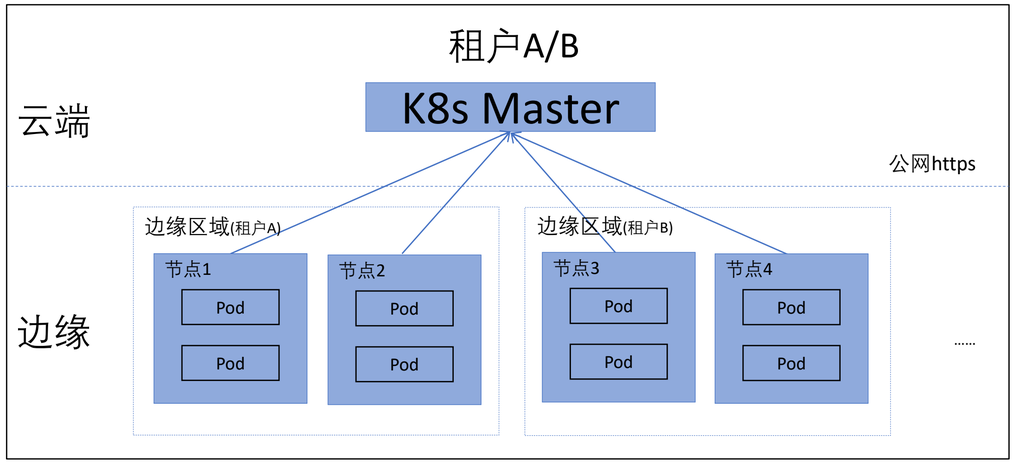
① 独享型



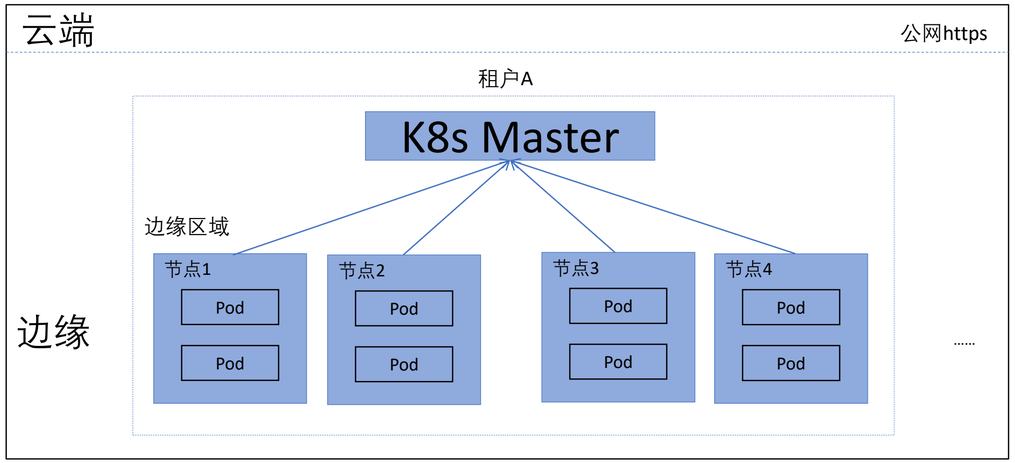
② 开放共享型



③ 半开放共享型



④ 边缘独享型

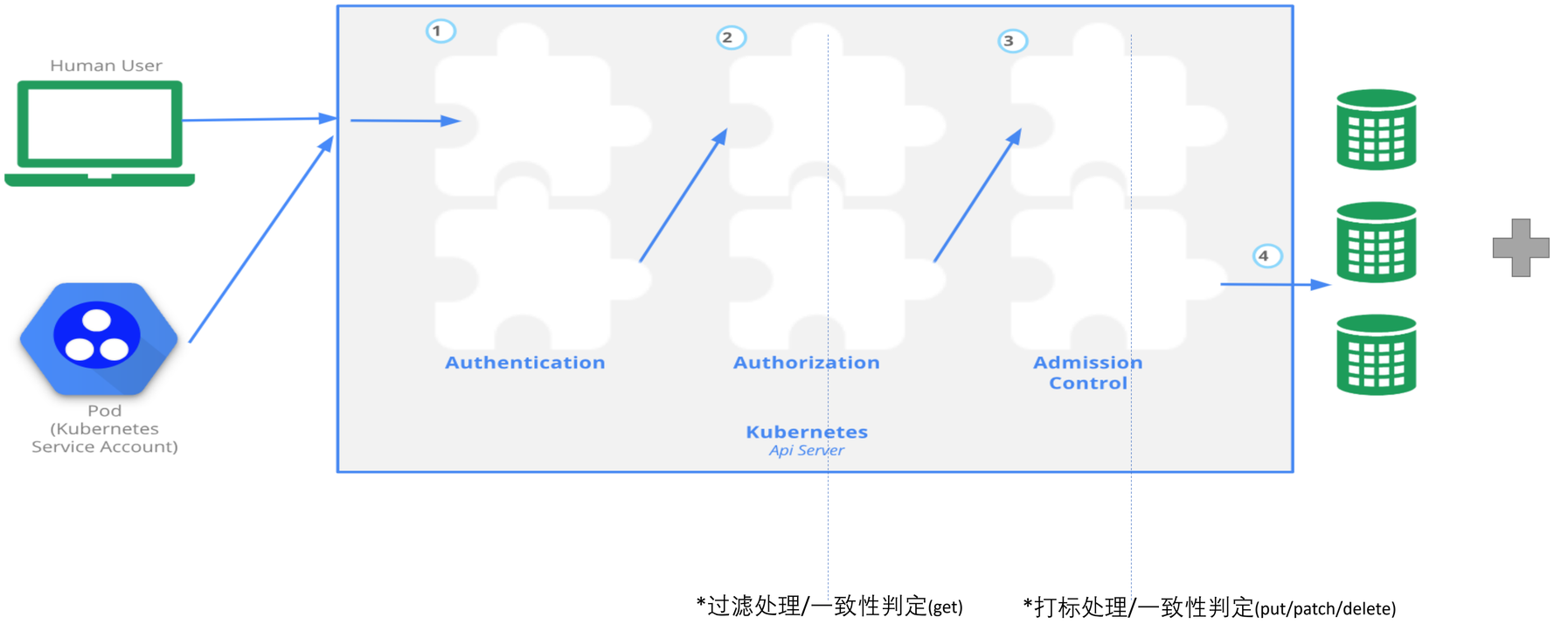


多租隔离 - API资源隔离诉求

阿里云 奥运会全球指定云服务商

编号	请求类型	请求资源类型	资源名称	租户名称	用例	多租隔离诉求
1	GET	Cluster资源(如Nodes)	否	-	/api/v1/nodes	只返回用户相关的Cluster资源
2	GET	Namespace资源(如Pods)	否	否	/api/v1/pods	只返回用户相关的Namespace资源
3	GET	Namespace资源(如Pods)	否	是	/api/v1/namespaces/{ns-name}/pods	是否访问用户所属的namespace
4	GET	Cluster资源(如Nodes)	是	-	/api/v1/nodes/{node-name}	是否访问用户所属的Cluster资源
5	GET	Namespace资源(如Pods)	是	-	/api/v1/namespaces/{ns-name}/pods/{name}	是否访问用户所属的namespace
6	POST	Cluster资源(如Nodes)	-	-	/api/v1/nodes	无
7	POST	Namespace资源(如Pods)	-	-	/api/v1/namespaces/{ns-name}/pods	否访问用户所属的namespace
8	PUT/PATCH/DELETE	Cluster资源(如Nodes)	是	-	/api/v1/nodes/{node-name}	是否访问用户所属的Cluster资源
9	DELETE	Cluster资源(如Nodes)	否	-	/api/v1/nodes	只删除用户所属的Cluster资源
10	PUT/PATCH/DELETE	Namespace资源(如Pods)	-	-	/api/v1/namespaces/{ns-name}/pods/{name}	否访问用户所属的namespace
11	GET/POST/PUT/PATCH/DELETE	公共租户资源(kube-public/kube-system/default)	-	-	/api/v1/namespaces/default/pods	GET请求放行，其他请求拒绝

多租隔离 - API认证鉴权模型



措施

1. 过滤处理
2. 一致性判定
3. 打标处理

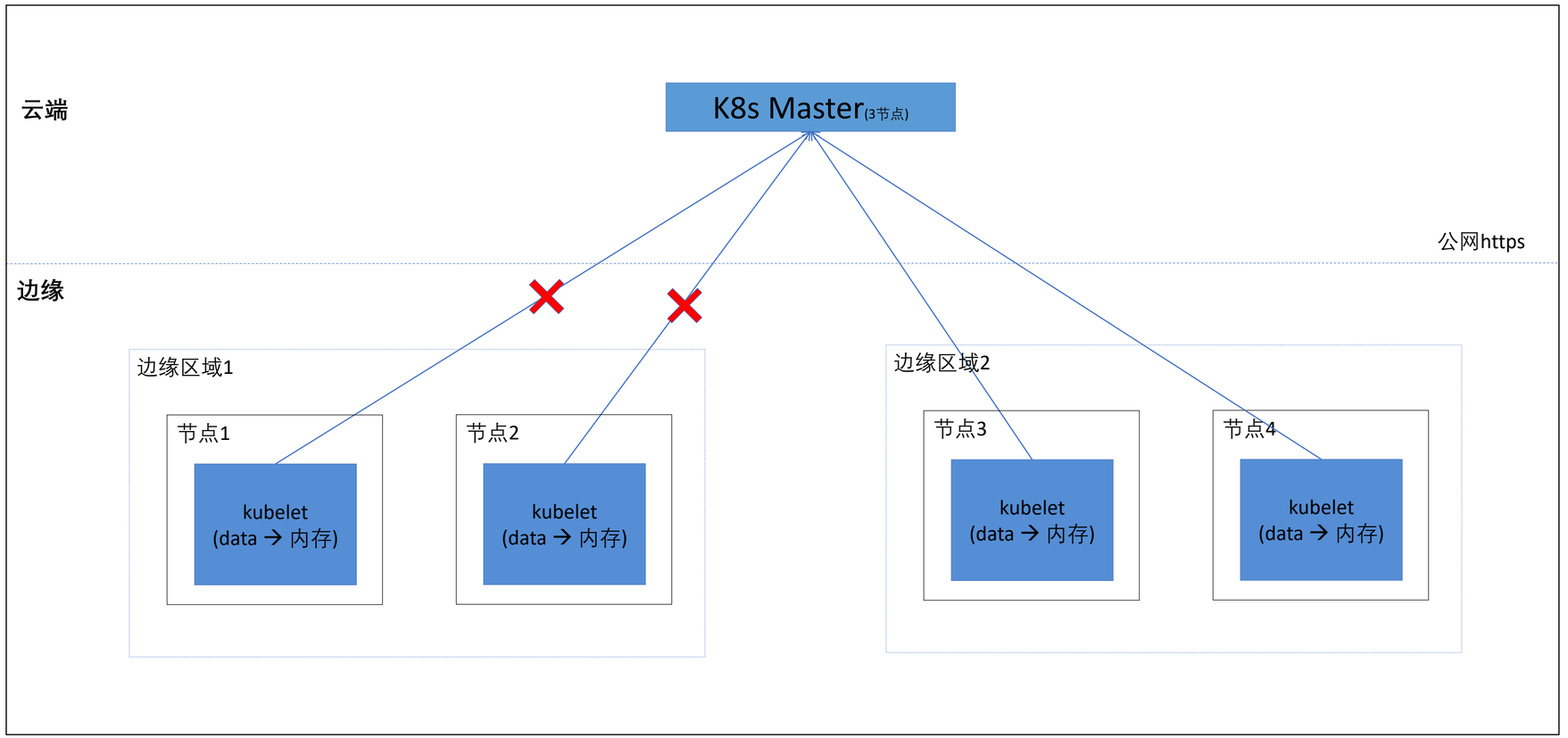
图片出处: https://kubernetes.io/docs/reference/access-authn-authz/controlling-access/

多租隔离 - Hack API

阿里云 奥运会全球指定云服务商

编号	请求类型	用例	多租隔离要求	多租隔离处理	实现模块
1	GET	/api/v1/nodes	只返回用户相关的Cluster资源	过滤处理	Authenticate后增加过滤处理
2	GET	/api/v1/pods	只返回用户相关的Namespace资源	过滤处理	Authenticate后增加过滤处理
3	GET	/api/v1/namespaces/{ns-name}/pods	是否访问用户所属的namespace	一致性判定	Authenticate后增加判定处理
4	GET	/api/v1/nodes/{node-name}	是否访问用户所属的Cluster资源	一致性判定	Authenticate后增加判定处理
5	GET	/api/v1/namespaces/{ns-name}/pods/{name}	是否访问用户所属的namespace	一致性判定	Authenticate后增加判定处理
6	POST	/api/v1/nodes	无	打标处理	Admission Control
7	POST	/api/v1/namespaces/{ns-name}/pods	否访问用户所属的namespace	一致性判定/打标处理	Admission Control
8	PUT/PATCH/DELETE	/api/v1/nodes/{node-name}	是否访问用户所属的Cluster资源	一致性判定	Admission Control
9	DELETE	/api/v1/nodes	只删除用户所属的Cluster资源	过滤处理	Authenticate后增加过滤处理
10	PUT/PATCH/DELETE	/api/v1/namespaces/{ns-name}/pods/{name}	否访问用户所属的namespace	一致性判定	Authenticate后增加判定处理
11	GET/POST/PUT/PATCH/DELETE	/api/v1/namespaces/default/pods	GET请求放行，其他请求拒绝	一致性判定	Authenticate后增加判定处理

节点自治

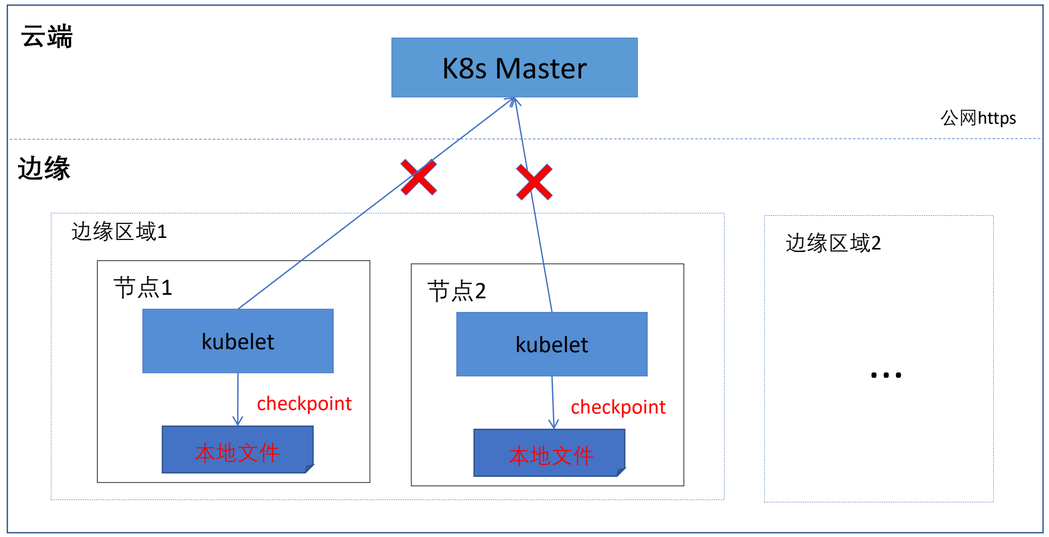


- 云端<->边缘网络错综复杂
- 稳定性差，断网数秒甚至数天之久
- 节点 POD 驱逐

节点自治 – 社区方案

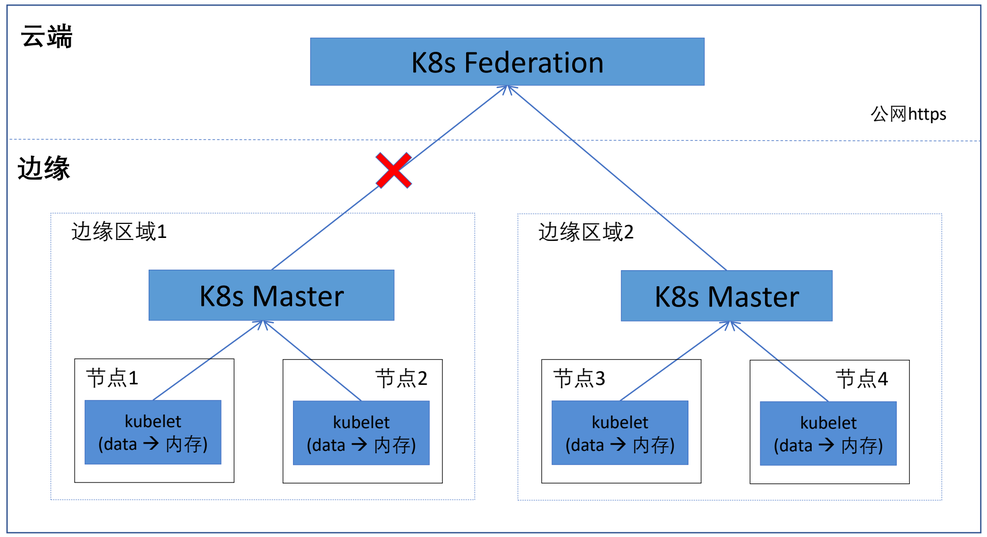
阿里云 奥运会全球指定云服务商

社区方案一



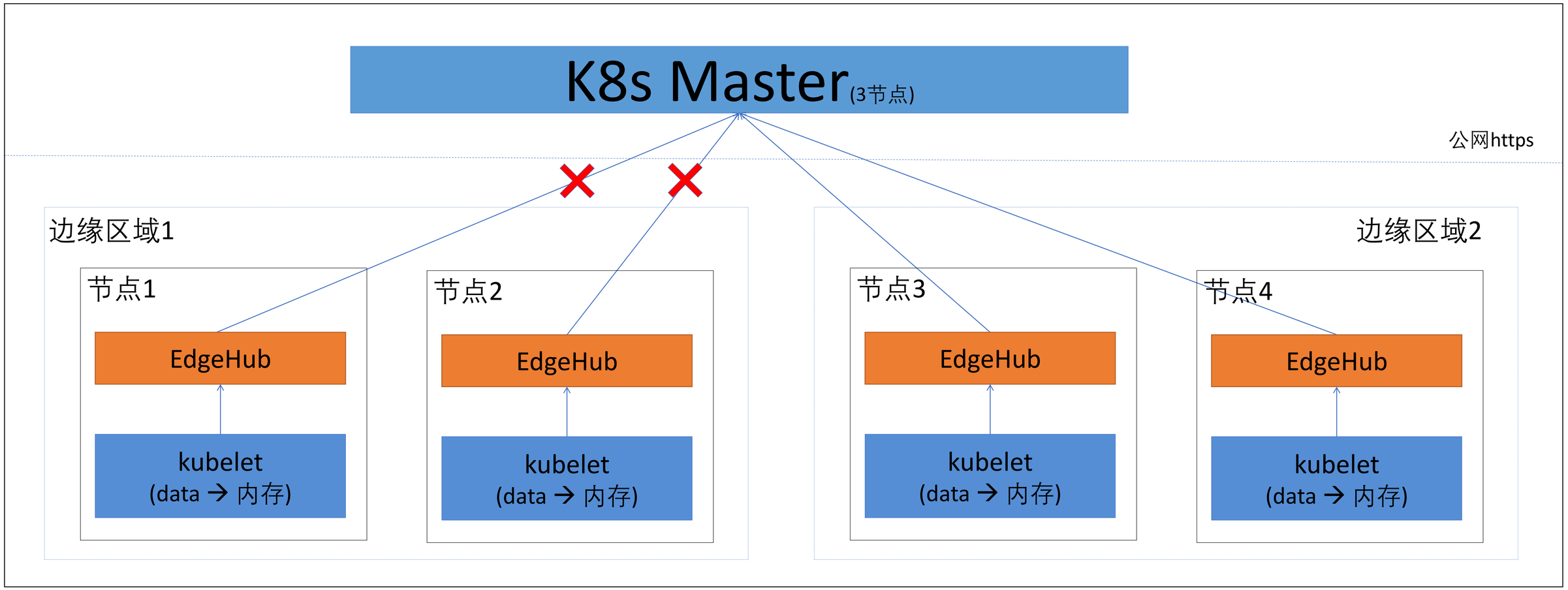
- 功能不足：仅缓存 POD，对 ConfigMap/Secret/PV/PVC等暂不支持。
- 可定制修改kubelet，但面临升级难题。

社区方案二



- 高成本：所有边缘区域均有 Master。
- 难运维：架构较复杂，集群规模数倍增加，监控、日志等复杂度。

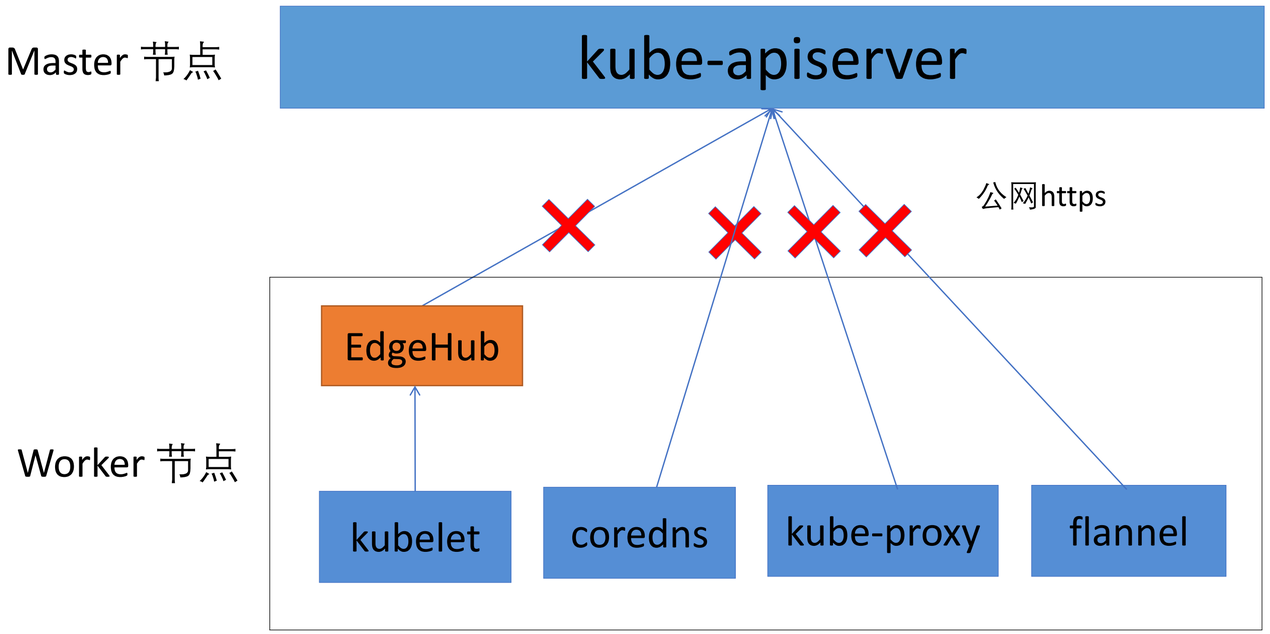
节点自治 – EdgeHub



边缘代理组件 EdgeHub：

1. 作为 kubelet 和 apiserver 之间的缓存和代理。
2. 在网络正常情况下，EdgeHub直接代理转发 Kubelet 请求到 apiserver，并缓存结果到本地。
3. 在节点断网的情况下，EdgeHub

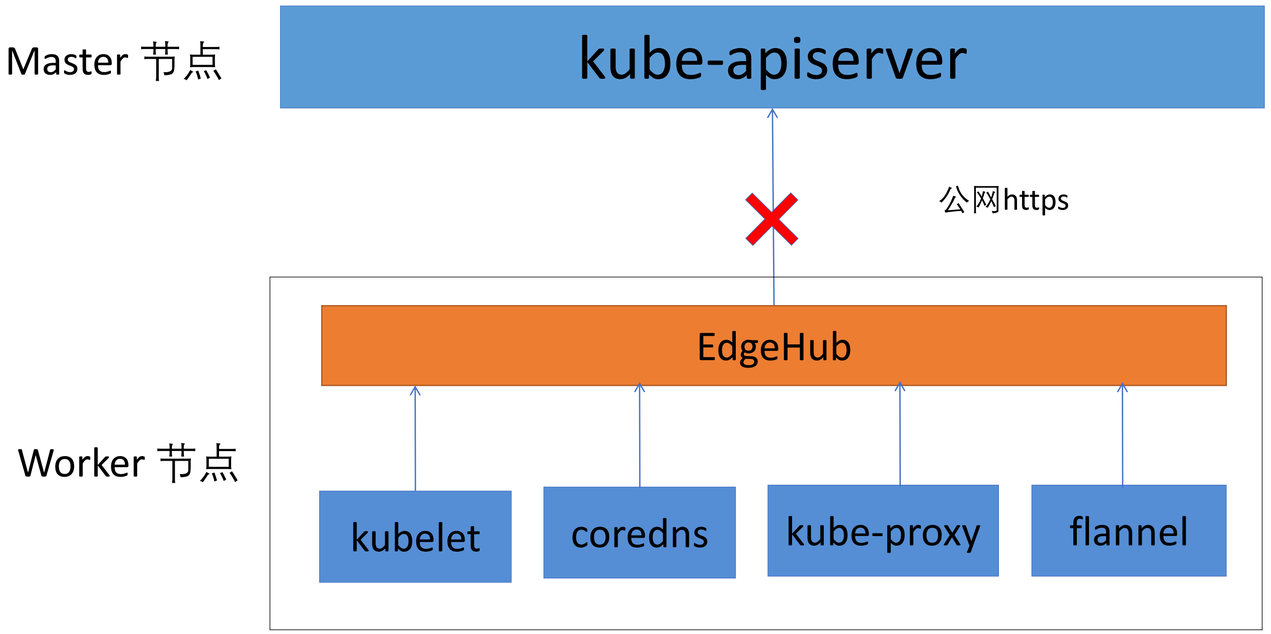
网络自治



断网后，需要解决的问题：

1. 网络组件重启，因无法连接 apiserver 导致无法正常启动。
2. 业务容器重启时，IP 被重新分配。
3. 节点被重启时，POD 网络不通（原因：重启后 flannel vtep MAC 改变，其他节点无法感知，另因为vetp的IP由node.spec.podCIDR决定,故保持不变)。

网络自治-解决方案



方案：

1. coredns/kube-proxy/flannel 均接入 EdgeHub 缓存代理组件。
2. 修改 IPAM ip 分配算法，缓存并复用 POD <-> IP。
3. 修改 flannel 源码，在 flannel 重启后可以复用之前为 flannel.1 vtep 分配的 MAC 地址 。

边缘容器(Edge Kubernetes) | ACK
2019.06 公测

阿里云 | 为了无法计算的价值