
KubeCon
CloudNativeCon
OPEN SOURCE SUMMIT
China 2023



KubeCon | CloudNativeCon


OPEN SOURCE SUMMIT

China 2023

# 填补Kubernetes的空白：IO资源调度和隔离

*Theresa Shan，云软件工程师 @英特尔*
*Cathy Zhang，高级首席工程师/架构师 @英特尔*

# 议程



- 动机
- 用户故事
- IO资源调度和隔离堆栈
- IO资源规范
- IO资源调度和隔离工作流
- 高级平台特性
- 成果
- 总结

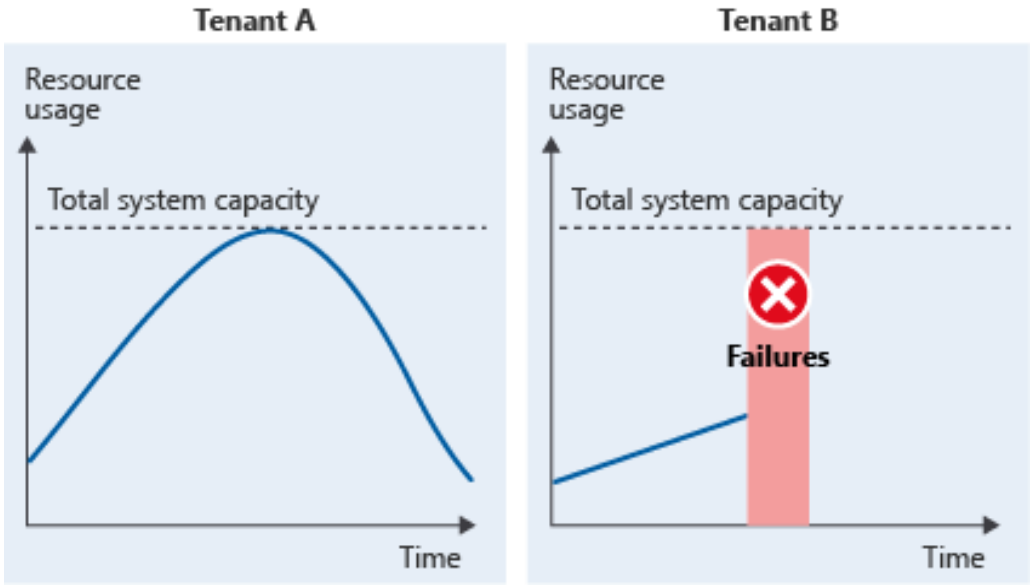
# 动机



提高资源利用率是云服务提供商 （CSP）的一个关键目标

- 根据 Vertiv 对 829 名数据中心专业人士的调查*，他们的主要目标之一是在 2025 年实现至少 60% 的 IT 资源利用率
- 当前大多数云数据中心未得到充分利用**

近邻干扰问题是CSP在提高资源利用率时的主要痛点，例如CPU，内存，存储，磁盘IO，网络IO***




*Data Center 2025 **Power Pollution and the Internet ***Azure Noisy Neighbor Antipattern

# 用户故事



工作负载类别:
- 类别 A（Guaranteed/GA）– 需要有 IO 资源保证的工作负载。（例如磁盘 IO、网络 IO）
- 类别 B（Best effort/BE）– 无需 IO 资源保证, 即可运行的工作负载

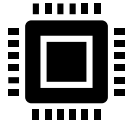
作为管理员，我想
- 保证 A 类工作负载的 IO 资源
- 在调度期间注意 IO 资源的可用性和工作负载的类别

作为独立软件供应商（ISV）提供商，我想
- 使用指定的 IO 带宽保证我的工作负载

# K8s资源调度和隔离现状



已支持的资源


CPU

内存

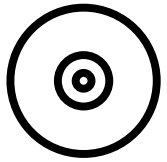

临时存储


内存大页


扩展设备

未支持的资源


网络 IO


磁盘 IO


缓存

# 填补K8s的空白



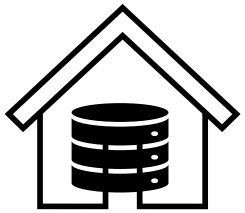
在k8s中添加了**磁盘 IO 感知调度和隔离**功能，以及**网络IO感知调度和隔离**功能

避免了磁盘IO和网络IO的近邻干扰问题

# IO资源调度和隔离堆栈



Edge cloud



On-premise cloud

Public cloud

AI

Database

Messaging


Web

IO Aware Scheduling

Network

Disk

IO Boundary Isolation and Telemetry

Profiling

Policy Manager

Analyzing

Resource Isolation

Monitoring

Metrics Advisor

**Hardware**



- 启用IO感知调度，以支持部署各种类型的工作负载并提高资源利用率，这些工作负载包括AI、数据库、消息传递应用程序、Web服务等
- 隔离不同 QoS 类型的工作负载的IO 资源
- 监控和分析工作负载的实时 IO 使用情况，并根据预定义的策略对节点 IO 资源压力采取措施
- 通过云原生环境中的英特尔应用设备队列 （ADQ）、英特尔® 基础设施处理单元 （英特尔® IPU） 等高级平台功能优化系统性能

# IO资源规范



```
apiVersion: v1
kind: Pod
metadata:
 name: ga_pod
 annotations:
   blockio.kubernetes.io/
       container-xxxServer-io-request: |
{"rbps": "20M","wbps": "30M","blocksize": "4k"}
spec:
 containers:
  - name: xxxServer
    image: xxx
     volumeMounts:
      - name: xxx-storage
        mountPath: /data/xxx
 volumes:
 - name: xxx-storage
   emptyDir: {}
```

磁盘 IO

GA：指定吞吐量(rbps/wbps)
BE：未指定吞吐量

```
apiVersion: v1
kind: Pod
metadata:
 name: ga_pod
 annotations:
   networkio.kubernetes.io/container.xxx-
server.io-request: |
{“ingress": "20M",“egress": "30M"}
spec:
 containers:
  - name: xxx-server
    image: xxx
```

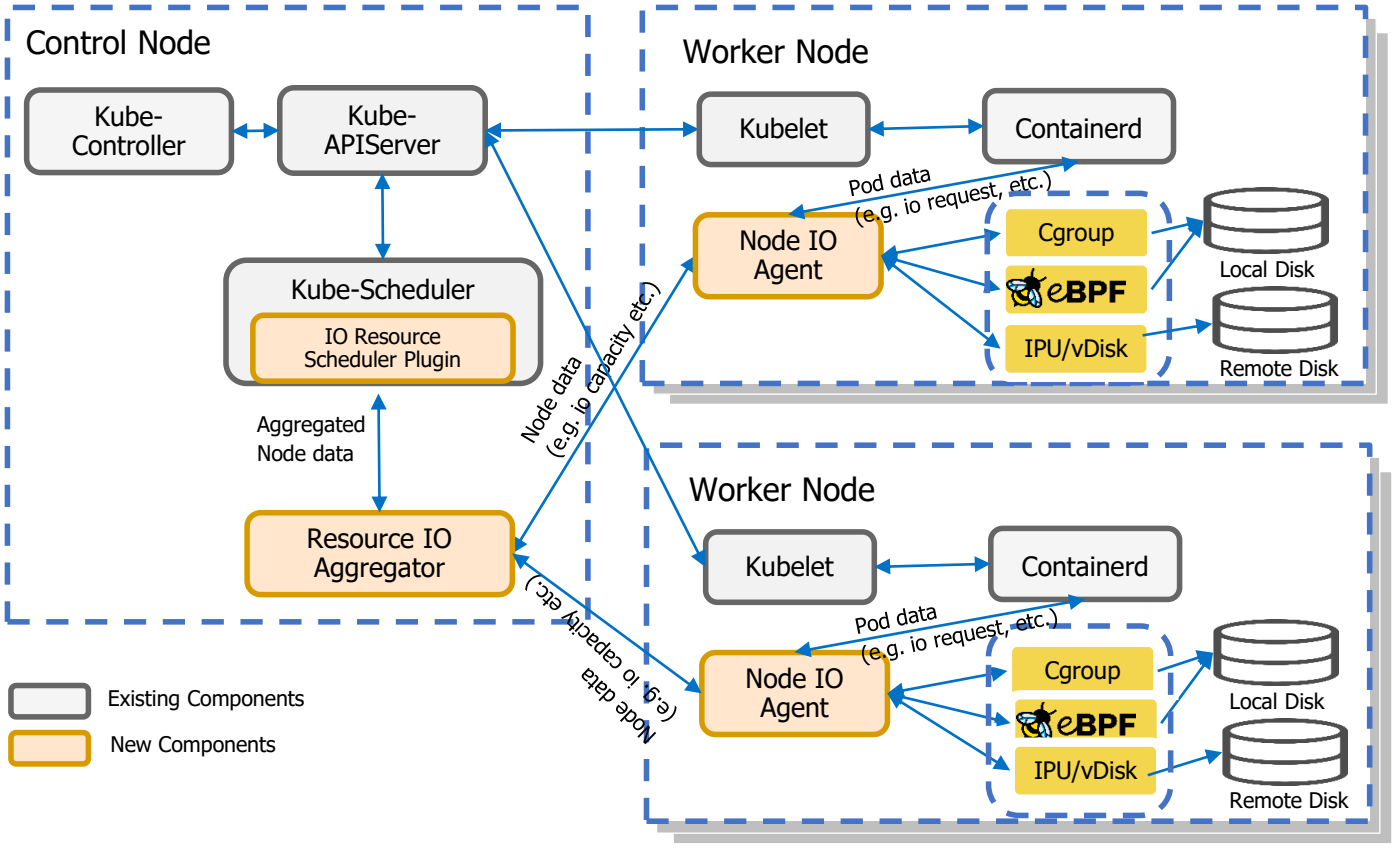
网络 IO

GA：指定ingress/egress
BE：未指定网络IO需求

# 磁盘IO资源调度和隔离



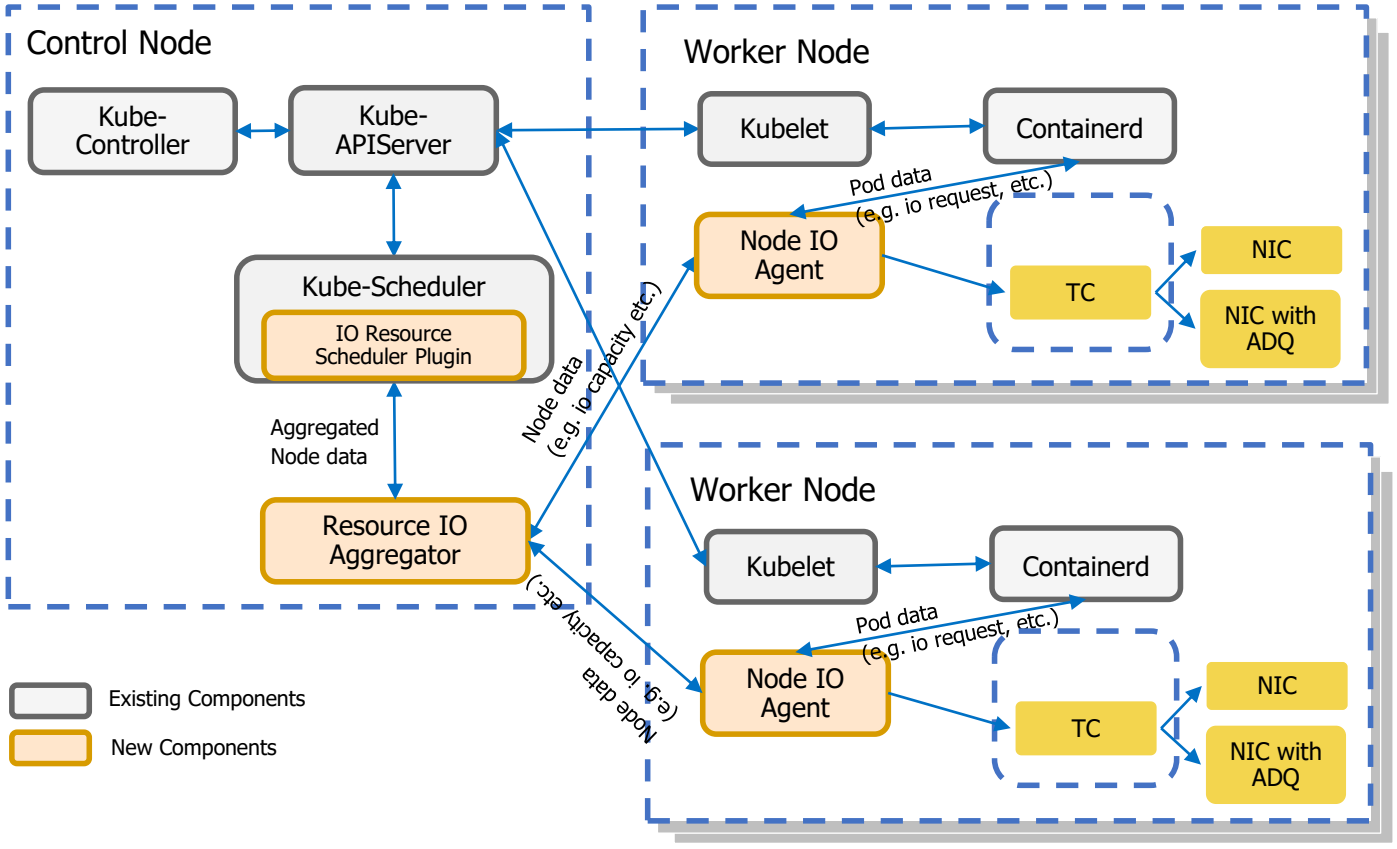
- 磁盘 IO 感知调度
  - 一个新的调度器插件，可为 Guaranteed，Burstable，和 Best Effort工作负载启用 IO 感知调度
- 磁盘 IO 资源隔离
  - Cgroup V2 用于磁盘隔离
- 磁盘 IO 资源的监控
  - 在每个节点上，监控每个工作负载的实时 IO 带宽，并将可用 IO 容量上报给 k8s 调度程序
  - 在每个节点上，动态调整 BE 工作负载的资源边界，以保证 GA 工作负载的 IO 性能
- 资源 IO 聚合
  - 聚合集群中每个节点的磁盘 IO 指标，并将其批量发送到 K8S 调度程序



# 网络IO资源调度和隔离



- 网络 IO 调度插件
  - 为 Guaranteed，Burstable，Best Effort 工作负载启用网络 IO 感知调度
- 网络 IO 资源隔离
  - 使用通用网卡或启用 ADQ 功能的网卡（例如英特尔® 800 系列网卡）的 TC 限制工作负载的 IO 资源
- 网络 IO 资源监控
  - 监控每个工作负载的实时 IO 带宽，并将其报告回节点 IO 代理



# 用于网络IO的ADQ



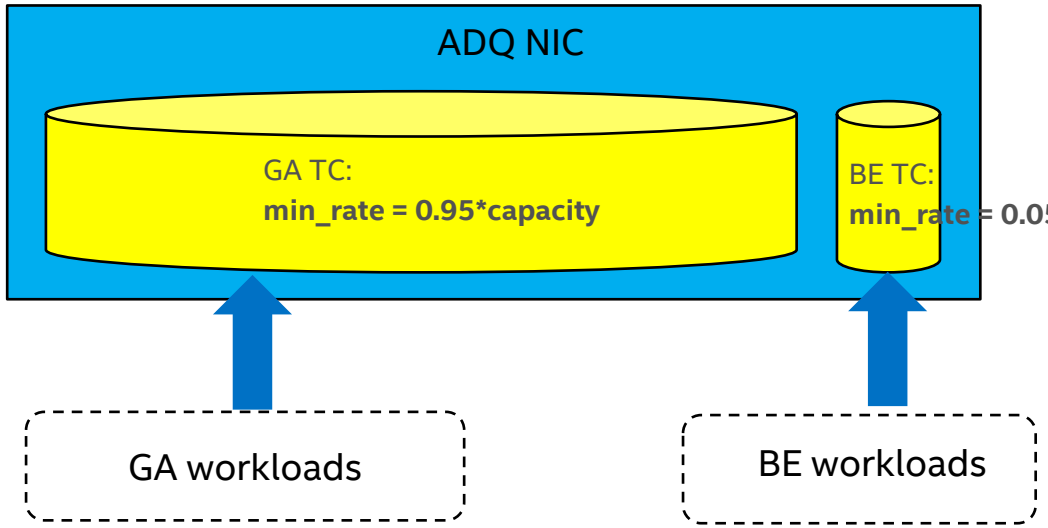
ADQ充当高速公路上的快速车道，可防止由交通拥堵而导致的数据中心关键应用程序的性能下降。它允许您在数据中心的网络硬件设备上保留专用通道/队列，以确保应用程序性能。


优势

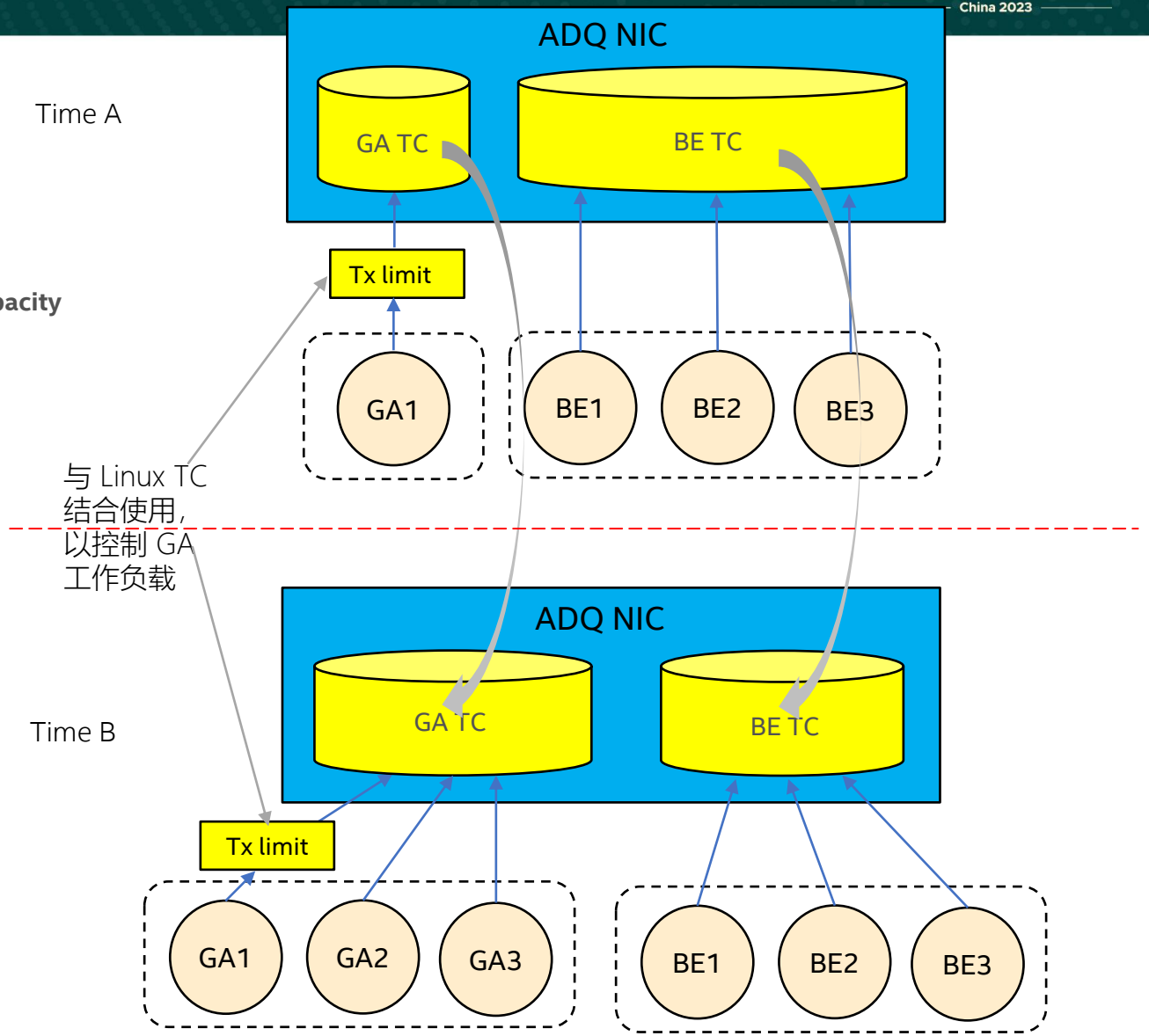
- 专用资源 = 提高可预测性
- 减少上下文切换 = 低延迟
- 高效的数据包处理 = 高吞吐量和可扩展性
- 可定制的流量整形 = 应用程序级QoS控制

# 用于网络IO的ADQ



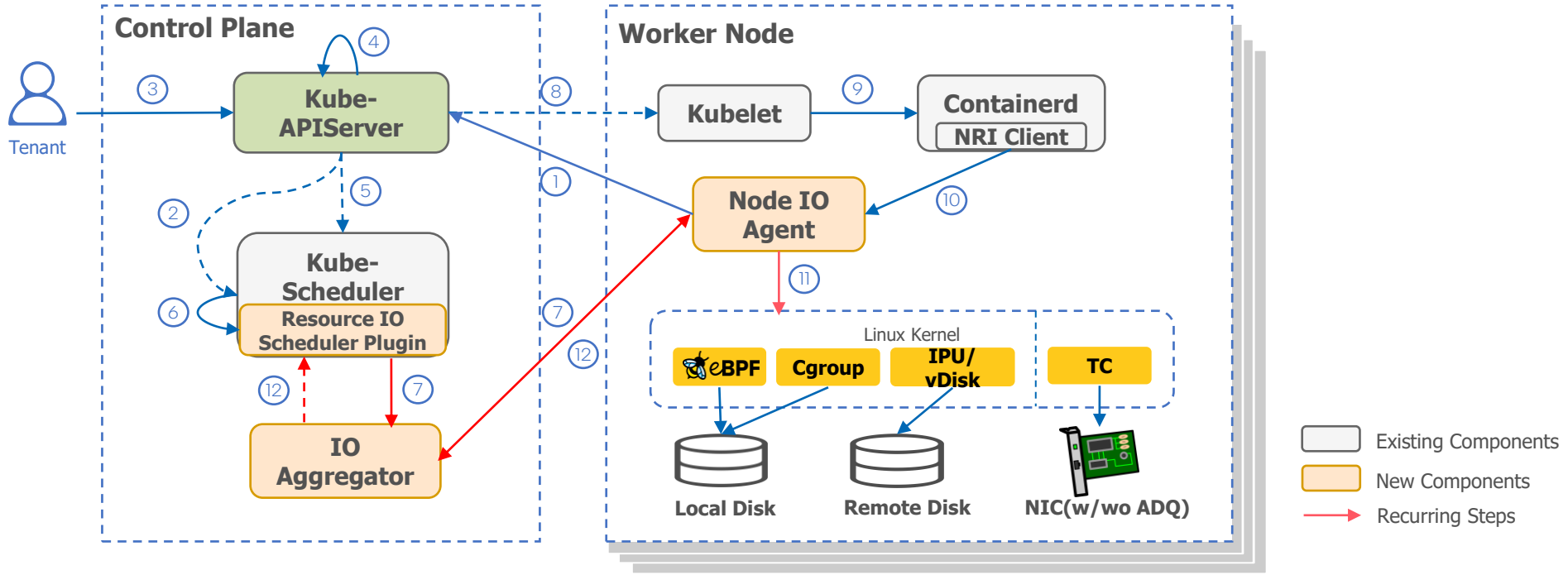



- 我们为用户定义了新的 POD 规范原语，以直接指定其网络 IO 带宽要求
- 我们通过利用 ADQ 的速率限制功能实现网络 IO 带宽隔离
- 支持动态网络IO资源边界调整，为每个工作负载启用弹性资源边界



# 资源IO感知的调度流程






1. 更新节点 IO 设备的静态信息（设备 ID、IO 转换系数、池大小、容量）
2. 将静态信息同步到调度程序缓存
3. 创建 Pod
4. 验证资源列表中的 IO 请求
5. 监控 Pod
6. 根据可分配的 IO 带宽做出调度决策
7. 将上下文（Pod 列表、可用 IO 带宽）同步到 IO 聚合器和节点 IO 代理
8. 监控 Pod
9. 创建 Pod
10. 通知 Pod 的创建
11. 轮询每个 Pod 的实时 IO 带宽
12. 将该节点的可分配 IO 带宽更新到调度程序缓存

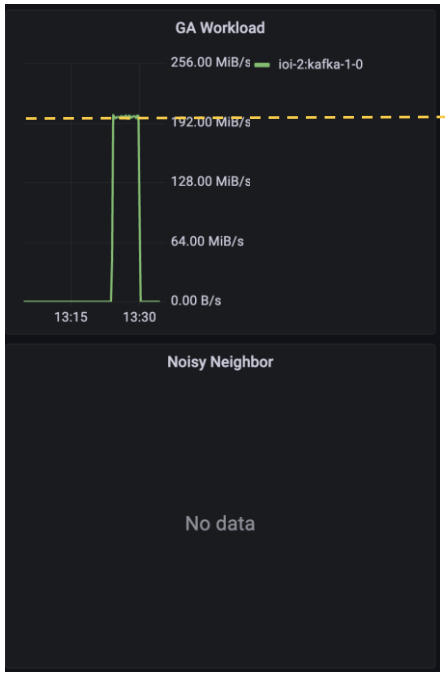
# 成果



GA:
Kafka-1(200M/4k)

NN:
Fio(200M/4k) * 4

1 Kafka




1 Kafka + 4 NN (测试 10 次)
不启用 IO 资源调度和隔离


1 Kafka + 4 NN (测试 10 次)
启用IO资源调度和隔离


Fio Pod 作为BE工作负载, 已经被压缩

该项目正在英特尔站台展示

# 总结



- 该项目填补了 K8s 的空白，在GA和BE的工作负载共存的情况下，保证GA工作负载 IO 资源
- 高级平台功能（例如ADQ NIC 和英特尔® IPU）可以帮助 IO 资源隔离和控流

# 问答



联系方式


- Cathy Zhang, cathy.h.zhang@intel.com
- Theresa Shan, theresa.shan@intel.com


基于磁盘 IO 资源感知调度的KEP： https://github.com/kubernetes-sigs/scheduler-plugins/pull/628

问题或意见?


KubeCon
CloudNativeCon
OPEN SOURCE SUMMIT
China 2023
谢谢

# 通知和免责声明



英特尔技术可能需要启用硬件、软件或服务激活。

没有任何产品或组件是绝对安全的。

您的费用和结果可能会有所不同。