
KubeCon
CloudNativeCon
OPEN SOURCE SUMMIT
China 2023


KubeCon | CloudNativeCon

OPEN SOURCE SUMMIT

China 2023

Fill a Gap of Kubernetes: IO Resource Scheduling and Isolation

Theresa Shan, Cloud Software Engineer @Intel
Cathy Zhang, Senior Principal Engineer/Architect @Intel

Agenda

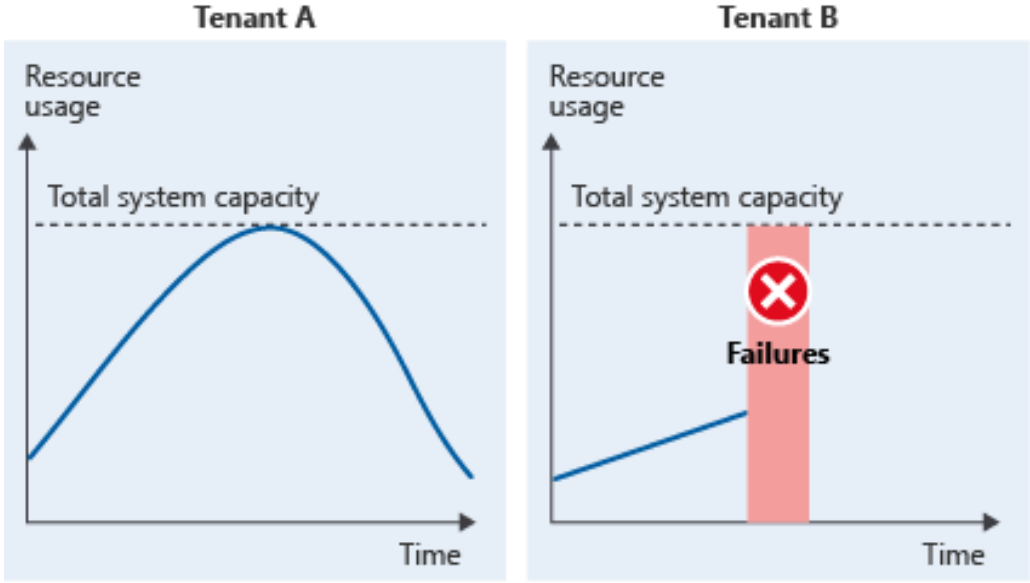
- Motivation
- User Story
- IO Scheduling and Isolation Stack
- IO Resource Specification
- IO Scheduling and Isolation Workflow
- Advanced Platform Features
- Result
- Summary

Motivation

Improving Resource Utilization is one key goal for Cloud Service Providers (CSPs)

- According to Vertiv survey* of 829 data center professionals, one of their key goals is for IT resource utilization rates to be at least 60% in 2025
- Most current cloud data centers are under-utilized**

Noisy neighbor problem is a main pain point for CSPs when improving resource utilization, e.g. CPU, memory, storage, disk IO, network IO ***



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

User Story

Workload Categories:

- Category A (guaranteed/GA) – workloads that require guaranteed IO resource.(e.g. disk IO, network IO)
- Category B (best effort/BE) – workloads that can function without guaranteed IO Resource

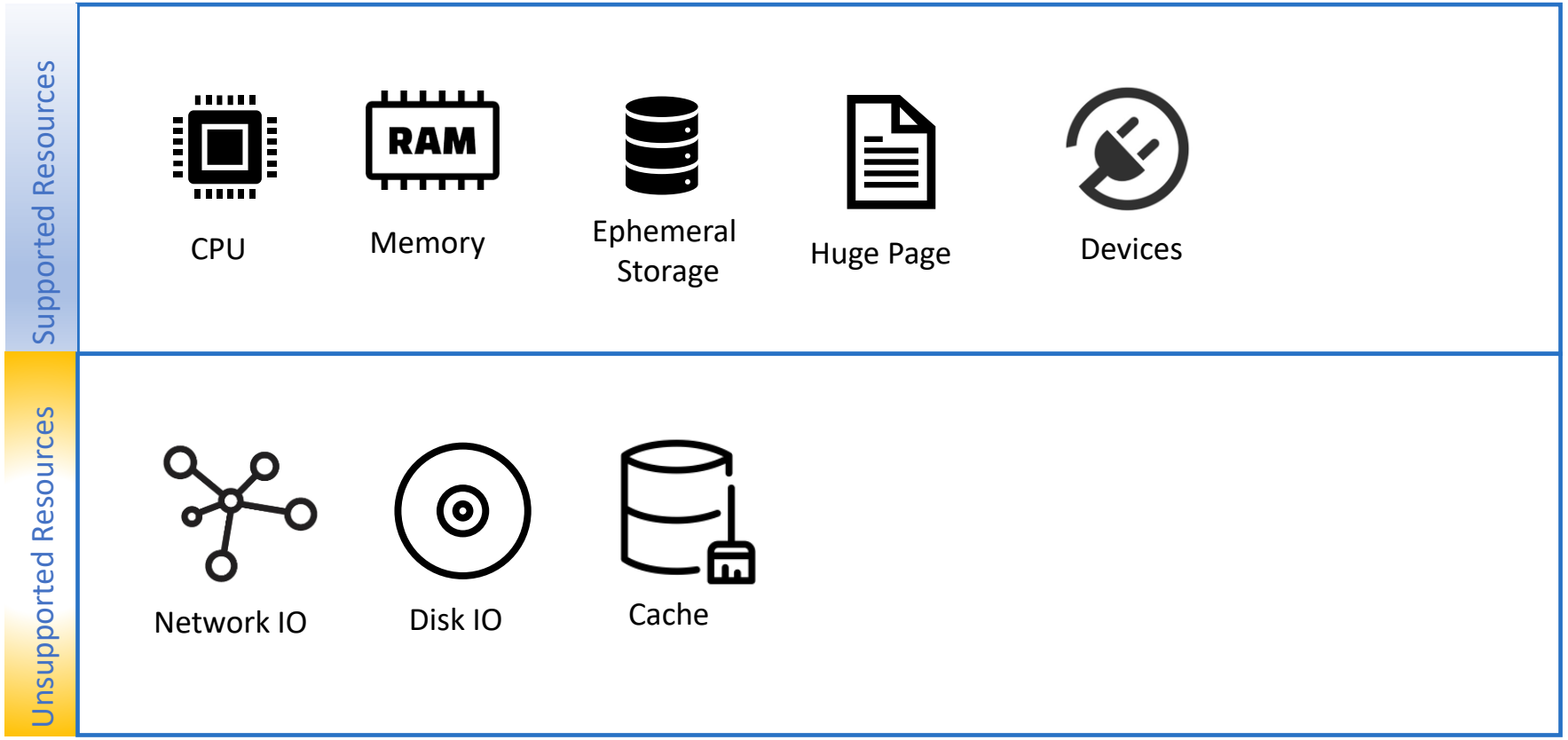
As admin, I want to

- Guarantee IO resource for the category-A workloads
- Be aware of IO availability and workloads' category during scheduling

As Independent Software Vendor (ISV) provider, I want to

- Guarantee my workloads with a specified IO bandwidth

K8s Resource Scheduling and Isolation Landscape



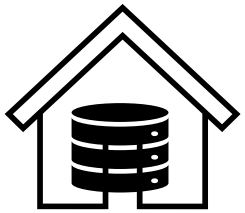
Fill The Gap

Add **disk IO aware scheduling and isolation as well as network IO aware scheduling and isolation** functionality to K8S. An end-to-end SW stack that enables cloud providers to pack as more workloads to increases resource utilization rate and at the same time meet critical workloads' IO performance goals

It avoids Disk IO and Network IO noisy neighbor problem

IO Scheduling and Isolation Stack

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

- Enable IO Aware scheduling to support deployment of different types of workloads such as AI, Database, Messaging App, Web Service etc. with improved resource utilization
- Isolate workloads' IO resource of different QoS types
- Monitor and analyze workloads' real-time IO usage and take actions upon node IO resource pressure based on the pre-defined policies
- Optimized system performance through advanced platform features, such as Application Device Queues (ADQ), Intel® Infrastructure Processing Unit (Intel® IPU) etc. in cloud native environment

IO Resource Specification


```
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: {}
```

Disk IO

GA: throughput(rbps/wbps) is specified
BE: throughtput is not specified

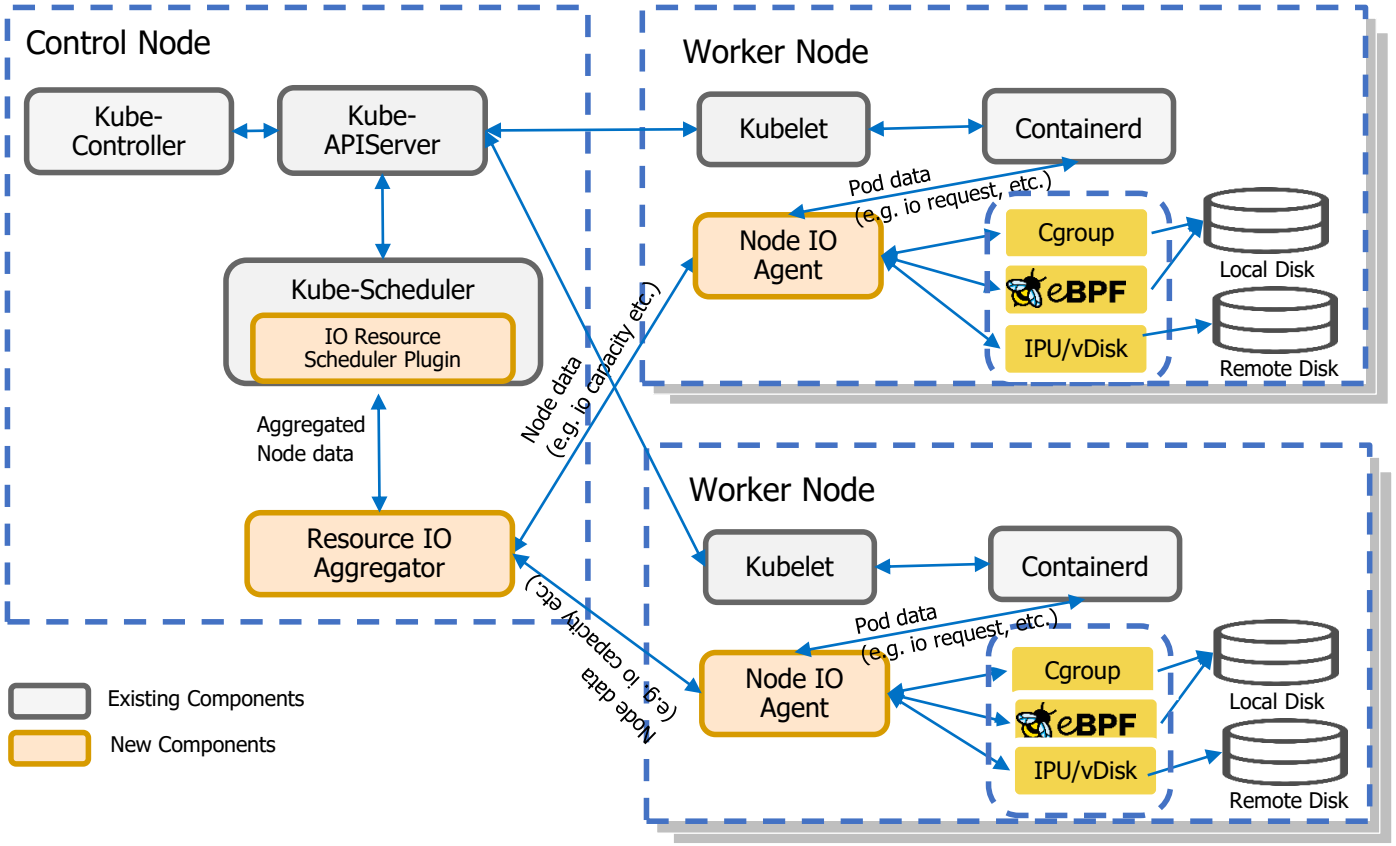
```
apiVersion: v1
kind: Pod
metadata:
 name: ga_pod
 annotations:
   networkio.kubernetes.io/container.xxx-
server.io-request: |
{“ingress": "20M",“egress": "30M"}
spec:
 containers:
  - name: xxx-server
    image: xxx
```

Network IO

GA: ingress/egress is specified
BE: no network IO request

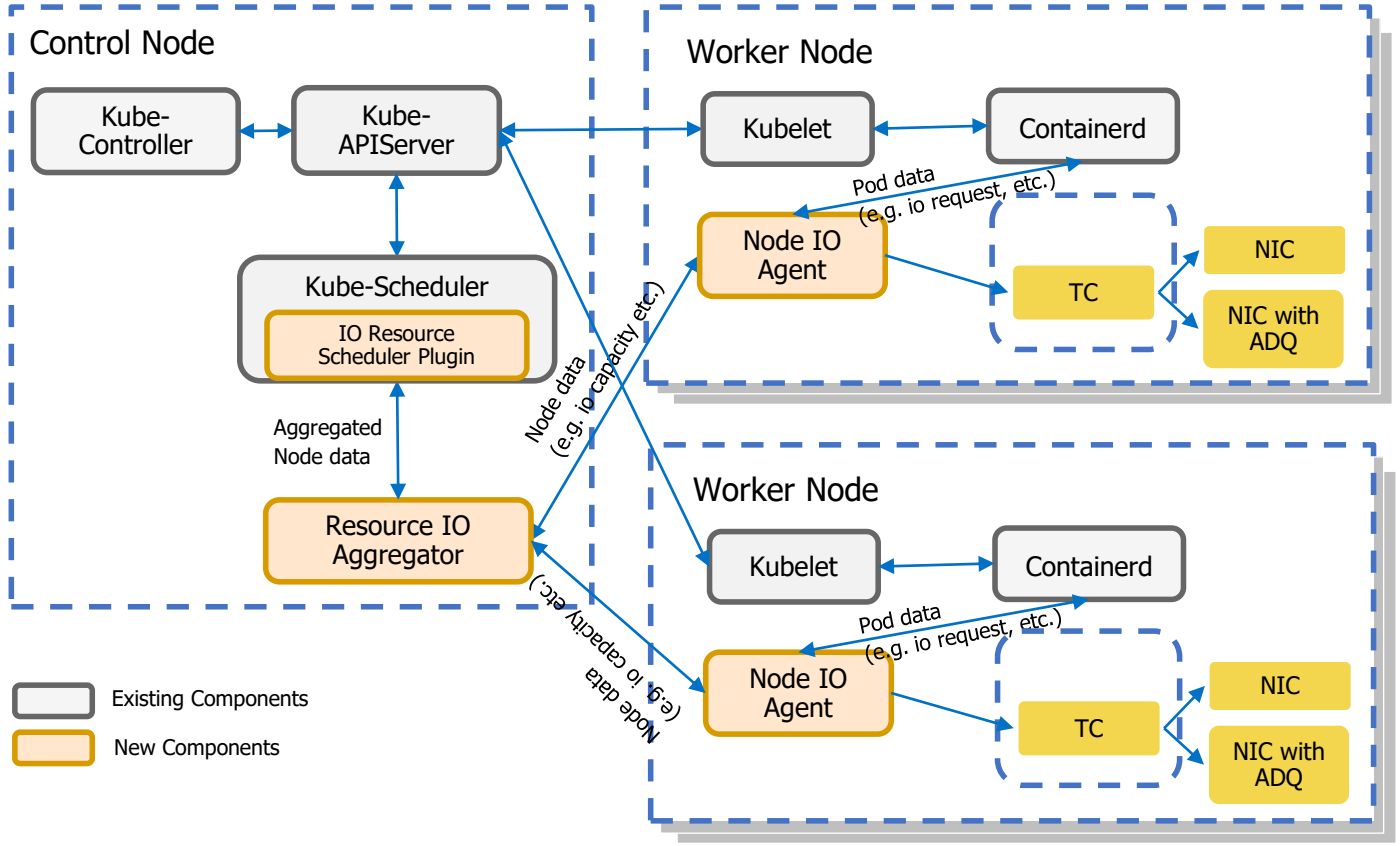
Disk IO Scheduling and Isolation

- Disk IO Aware Scheduling
  - A new scheduler plugin to enable io aware scheduling for Guaranteed, Burstable, and Best Effort workloads
- Disk IO Resource Isolation
  - Cgroup V2 for disk isolation
- Disk IO Resource Monitoring
  - On each node, monitor the real-time IO BW of each workload and report the available IO capacity to k8s Scheduler
  - On each node, dynamically adjust BE workloads' resource boundary to ensure GA workloads' IO performance
- Resource IO Aggregation
  - Aggregate Disk IO metrics from each node in the cluster and batch send them to the K8S Scheduler



Network IO Scheduling and Isolation

- Network IO Scheduler Plugin:
  - Enable network io aware scheduling for Guaranteed, Burstable, and Best Effort workloads
- Network IO Resource Isolation
  - Throttle workload's IO Resource using TC with generic NIC or with ADQ enabled NIC (e.g. Intel® 800 Series Ethernet Driver)
- Network IO Resource Monitoring
  - Monitor the real-time IO BW of each workload and report it back to Node IO Agent



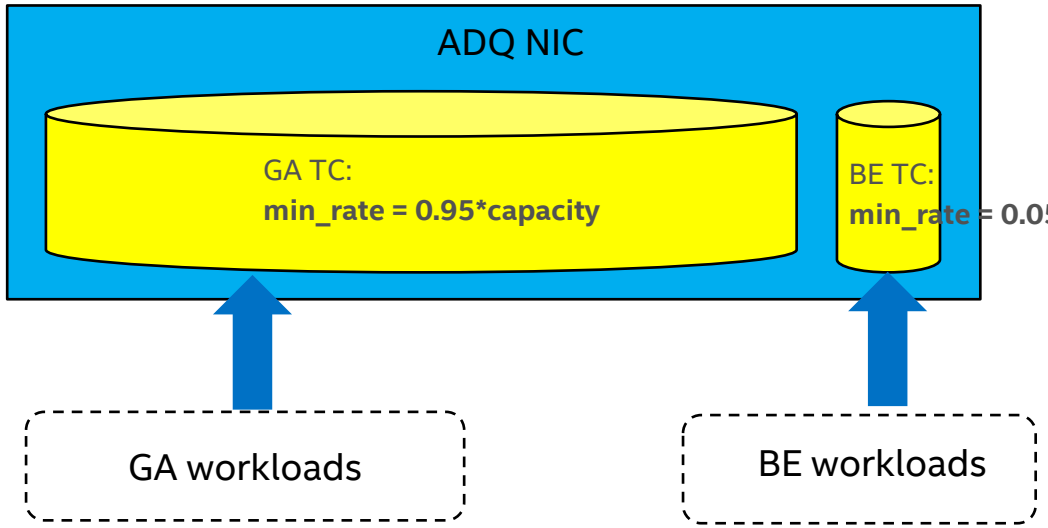
ADQ for Network IO

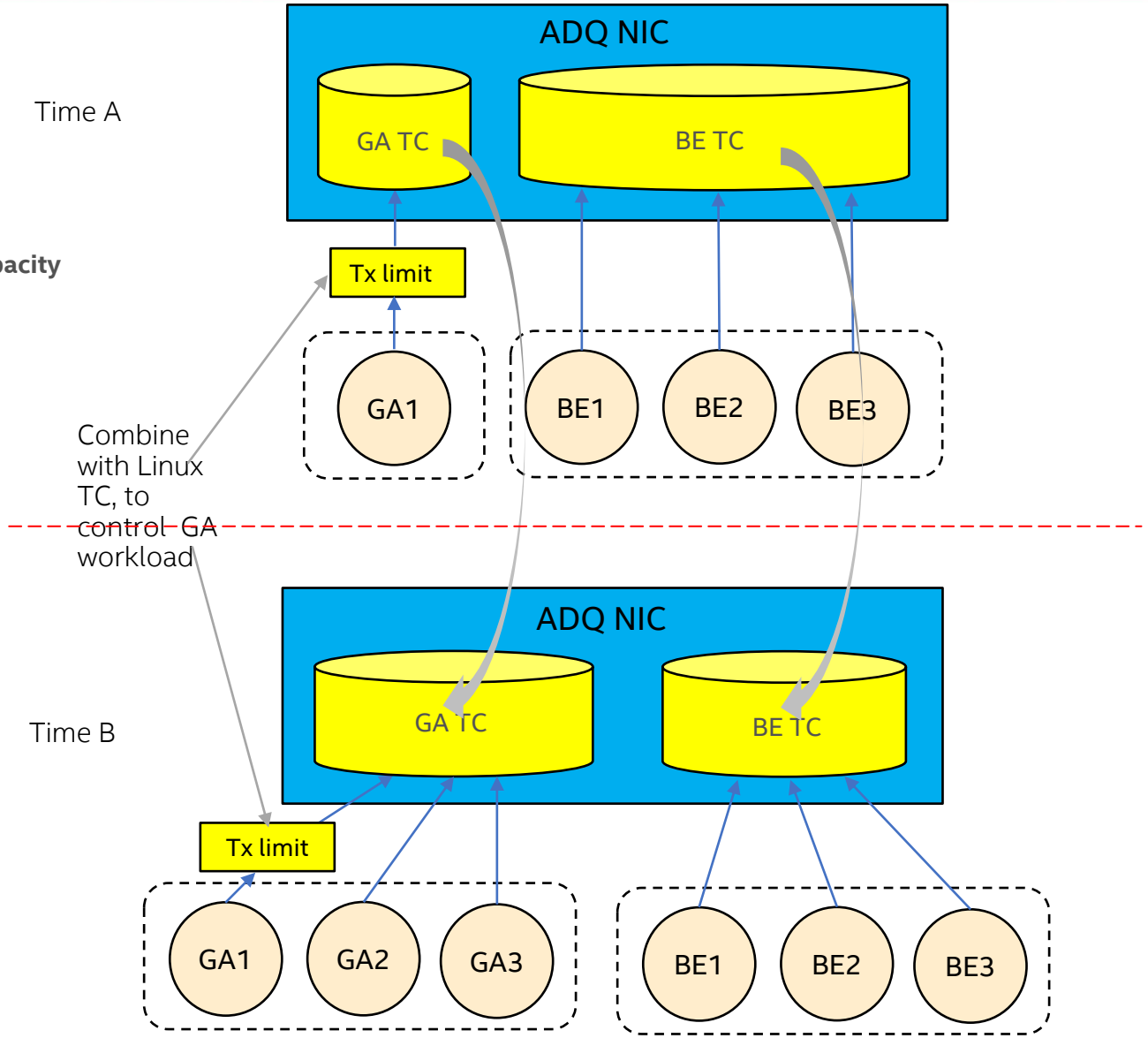
ADQ acts as an express lane on the freeway that prevents traffic jams which leads to degrading performance of critical applications in your data center. It allows you to reserve dedicated lanes/queues on the network hardware devices in your data center to ensure the application performance.

Strength

- Dedicated resources = Increased predictability
- Less context switching = Low latency
- Efficient packet processing = High throughput and scalability
- Customizable traffic shaping = Application level QoS control

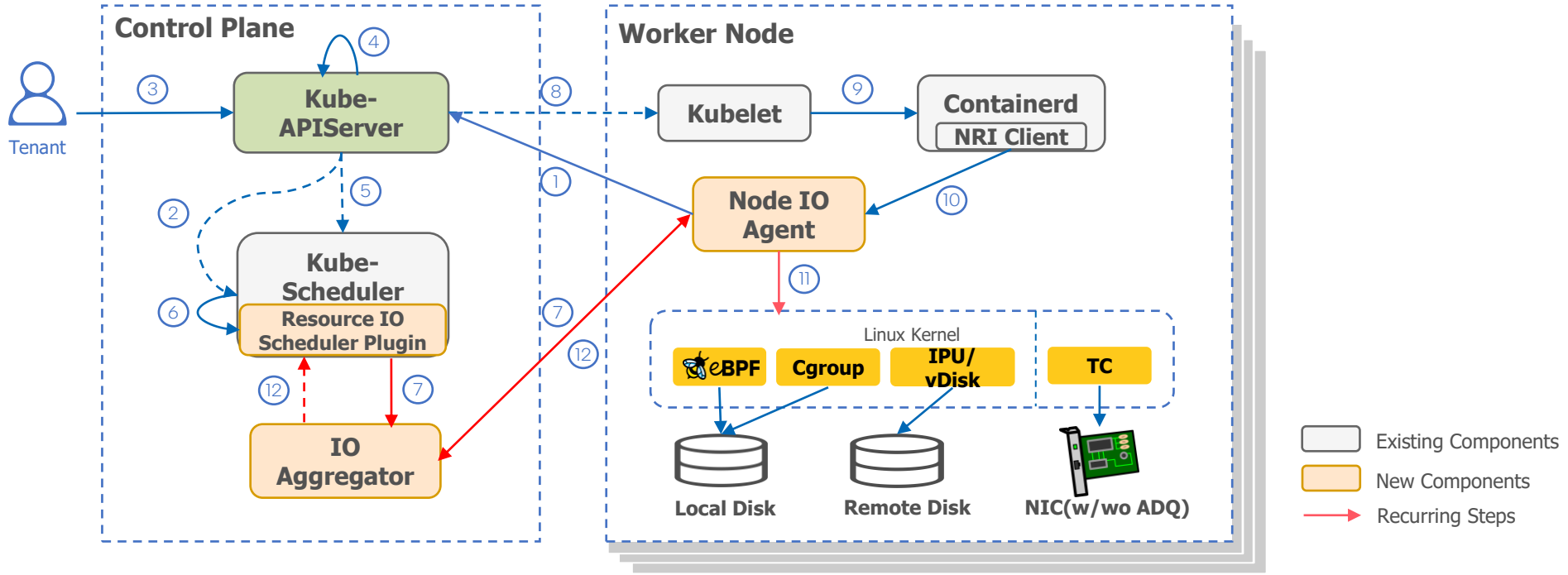
ADQ for Network IO





- We define new POD spec primitives for the users to directly specify their network IO BW requirement
- We enable network IO BW isolation by leveraging ADQ's rate limiting feature
- We support dynamic network IO resource boundary adjustment, which enables elastic resource boundary for each workload

Resource IO Aware Scheduling Flow



1. Update node IO devices' static info (device ID, IO conversion coefficients, pool size, empty capacity)
2. Sync the static info to scheduler cache
3. Create Pod
4. Validate IO request in resource list
5. Watch Pod
6. Make schedule decision based on allocable IO BW
7. Sync the context (pod list, available IO BW) with IO Aggregator and Node IO Agent
8. Watch Pod
9. Create Pod
10. Notify Pod Creation
11. Poll real-time IO BW per Pod
12. Update the node's allocable IO BW to scheduler cache

Result

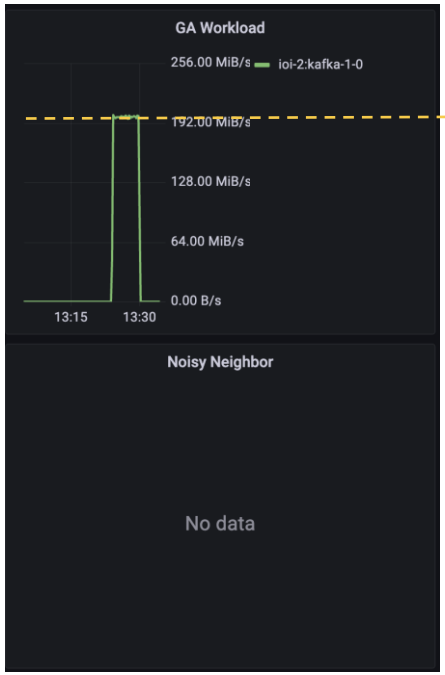
GA:
Kafka-1(200M/4k)

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

1 Kafka

1 Kafka + 4 NN (run 10 times)
Without IO Scheduler & Isolation

1 Kafka + 4 NN (run 10 times)
With IO Scheduler & Isolation



Fio Pod work as BE level, been compressed

The project is being demonstrating @ Intel booth

Summary

- The project fills the gap in K8s to co-locate guaranteed and best effort workloads and provide guaranteed IO resource to guaranteed workloads
- Advanced platform features (e.g. ADQ NIC and Intel® IPU) can help IO resource Isolation and rate limiting

Q&A

Contact Info:

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

KEP for disk IO aware scheduling: https://github.com/kubernetes-sigs/scheduler-plugins/pull/628

Any questions or comments?


KubeCon
CloudNativeCon
OPEN SOURCE SUMMIT
China 2023
Thank You

Notices & Disclaimers

Intel technologies may require enabled hardware, software or service activation.

No product or component can be absolutely secure.

Your costs and results may vary.