
KubeCon
CloudNativeCon
North America 2023


KubeCon | CloudNativeCon

North America 2023

Kepler: Project Update and Deep Dive

Marcelo Amaral
Staff Research Scientist
IBM Research

Tatsuhiro Chiba
Senior Technical Staff Member
IBM Research

Agenda

- Introduction and Motivation
- Kepler deployment updates
- BM power model updates
- VM power model updates
- Performance analysis and optimization updates


KubeCon | CloudNativeCon

North America 2023

Introduction and Motivation

Sustainability is a rapidly growing area of focus for many organizations

51% of CEOs name sustainability as their greatest challenges for their organization over the next 2–3 years*

CIOs and CTOs identify sustainability as the **top area** where technology will have the greatest impact over the next 3 years**

Data centers around the world consume **200 to 250** Terrawatt-hours*** of electricity

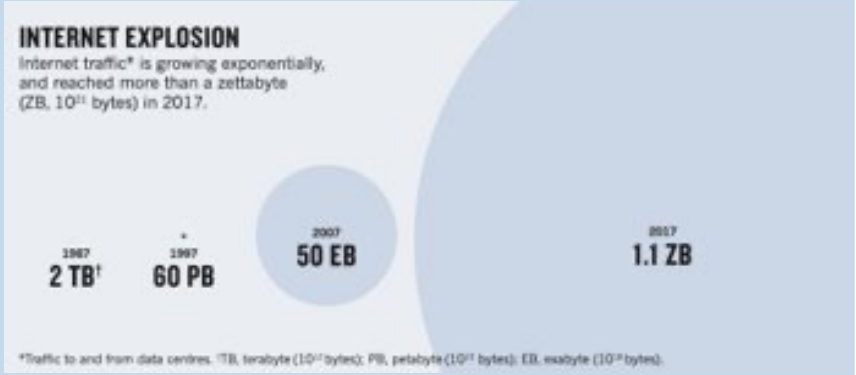
* The IBM Institute of business value: Own your impact
** The IBM Institute for Business Value: Sustainability as a transformation catalyst
*** The IBM Institute for Business Value: IT sustainability beyond the data center

The Computer Energy Problem

KubeCon CloudNativeCon North America 2023

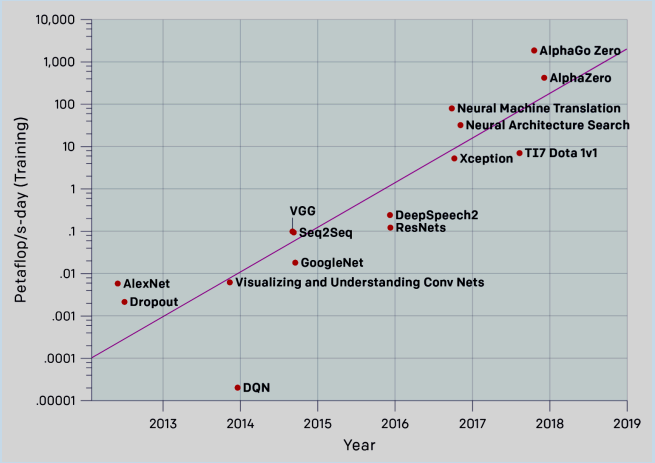
We are at an inflection point :

1. Demand is growing at exponential scale



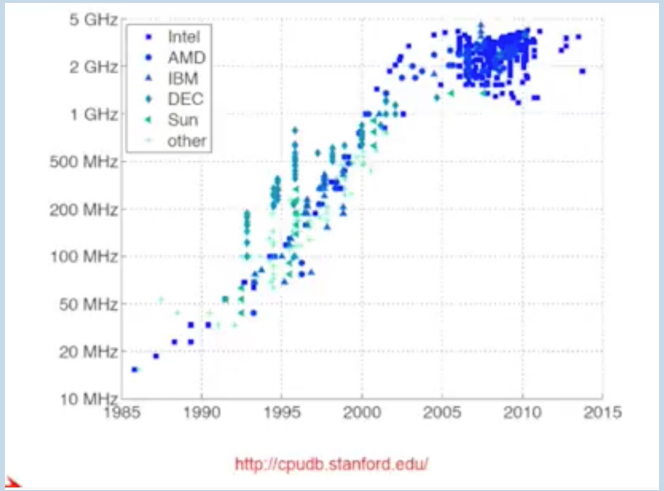
How to stop data centers from gobbling up the world's electricity
https://www.nature.com/articles/d41586-018-06610-y

2. The emergence of energy-demanding workloads(AI)



AI power consumption **doubles** every **3-4 months**
*** Green AI, R. Schwartz, J. Dodge, N. A. Smith, O. Etzioni 2019**

3. The end of Dennard Scaling means we can't keep up



➔ Some predict that electricity consumed by Data Centers will increase to 8% by 2030
➔ Golden Era for Chip Design

Towards Sustainability Computing on Clouds

Quantification	Assess	Optimization	Automation
Energy and **Carbon Footprint** per **workload**, **tenant**, **VM**, **container**, **Service**, etc.	Identify **hotspots** and **applicable strategies.** Calculate potential savings.	A set of frameworks and schedulers to optimize performance/Watt at scale.	A set of controllers to take actions dynamically to be an ideal state.

- How to collect workload energy for containers running on Clouds?
- Can we minimize the overhead of collector?
- What approach can we take if we could not see the power meter?
- Can we identify whether the power consumption comes from correctly?

Kepler: Kubernetes-based Efficient Power Level Exporter

https://github.com/sustainable-computing-io/kepler

- CNCF Sandbox
- GitHub Main Repository with 758 Stars and 116 Forks

Report

process, container, pod level energy consumption

GPU, CPU, DRAM, Node

K8S/OCP/RHEL on bare metal and VMs

x86_64, s390x, aarch64 in future

Advantage

eBPF to reduce overhead

Power Model and Server

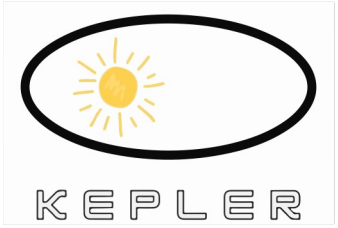
Standalone Mode

Power Model

Ratio Power Model

Trained Power Model

Project Roadmap



Inovation Timeline

Fev 2022

Birth of Kepler Red Hat working to measure container's Power Consumption on Kubernetes and OpenShift

IBM joined efforts with Red Hat

Jun 2022

Trained Power Models
VMs on public cloud does not expose power metrics, as a result Kepler needed to create power models

Model Server
a project to create power models from regression of data collected from bare-metal nodes

Idle Power
accordingly, to the GHG protocol for a fair power distribution among application the constant idle power should be divided based on the application size and dynamic power based on the load

Mar 2023

Kepler Standalone
to support Edge computing use cases, developers added the option to run Kepler outside Kubernetes and install via RPM

Process Power Consumption
developers added the support to measure the power per process not only containers

Jun 2023

CNCF Sandbox

Oct 2023

Disabled cGroup and Kubelet metric collection
the newer power model relies only on BPF metrics and disabling cGroup and Kubelet metrics further minimize the overhead

Increase Prometheus scrape Interval
to minimize overhead

2024

eBPF Sampling Rate
to minimize overhead

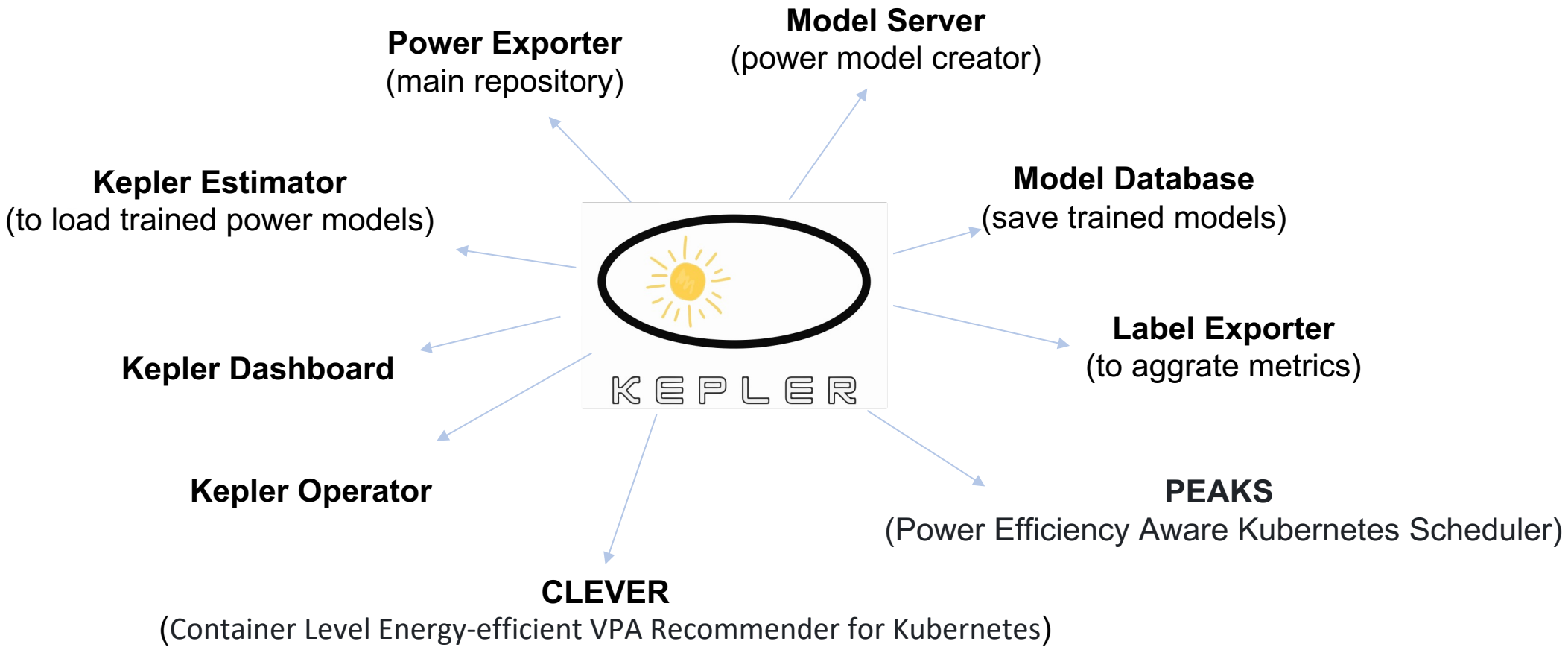
OpenTelemetry

VM power passthrough
expose VM power within the VM, this use case could be used by cloud provides in the future

New Idle Power Logic
needs regression to determine the true Idle power (e.g., with 0 resource utilization)

Kepler Project Ecosystem

Quantification	Assess	Optimization	Automation



https://github.com/sustainable-computing-io


KubeCon | CloudNativeCon

North America 2023

DEMO

Administrator
Home
Operators
OperatorHub
Installed Operators
Workloads
Networking
Storage
Builds
Observe
Alerting
Metrics
Dashboards
Targets
Compute

Project: openshift-operators

Installed Operators > Operator details

Kepler

0.8.0 provided by Kepler Operator Contributors

Actions

Details YAML Subscription Events Kepler

Provided APIs

Kepler

Kepler is the Schema for the keplers API

Create instance

Description

Introduction

Kepler (Kubernetes Efficient Power Level Exporter) uses eBPF to probe energy-related system stats and exports them as Prometheus metrics.

API Breakage

NOTE: `0.6.z` is **NOT** backward compatible with `0.5.0`. Before installing `0.6.z`, you must uninstall `0.5.0` or (earlier) completely (including CRDs) by following the steps in the uninstall operator section

Documentation

Documentation and installation guide can be found below:

Provider
Kepler Operator Contributors

Created at
Oct 5, 2023, 4:45 AM

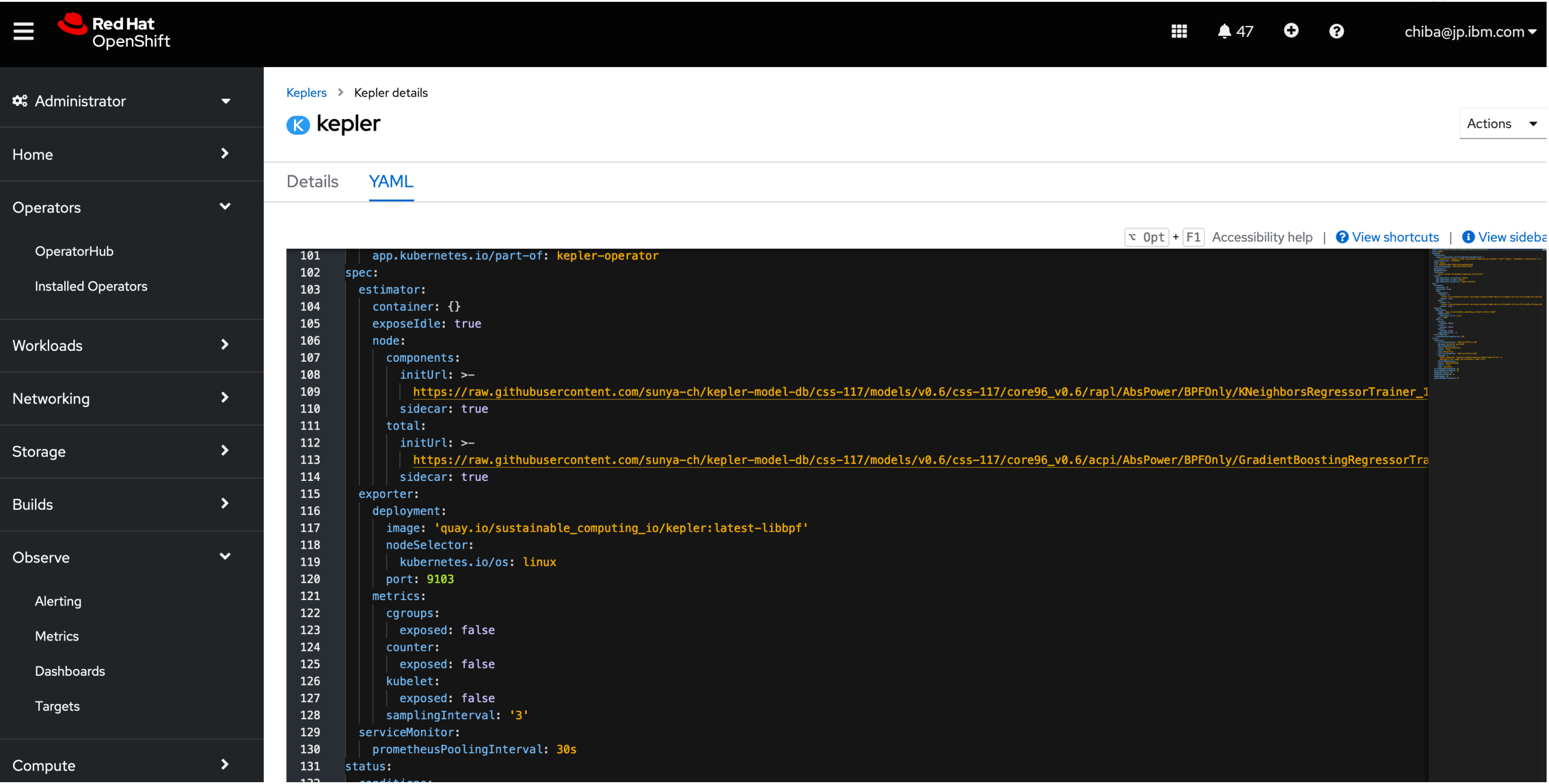
Links
Installation Guide
https://sustainable-computing.io/installation/community-operator/

Kepler Project
https://sustainable-computing.io/

Maintainers
Parul Singh
parsingh@redhat.com

Huamin Chen
hchen@redhat.com

Sunil Thaha
sthaha@redhat.com


Red Hat OpenShift
47
chiba@jp.ibm.com
Administrator
Home
Operators
OperatorHub
Installed Operators
Workloads
Networking
Storage
Builds
Observe
Alerting
Metrics
Dashboards
Targets
Compute
Keplers
Kepler details
kepler
Actions
Details
YAML
Opt + F1 Accessibility help | View shortcuts | View sideba
app.kubernetes.io/part-of: kepler-operator
spec:
estimator:
container: {}
exposeIdle: true
node:
components:
initUrl: >-
https://raw.githubusercontent.com/sunya-ch/kepler-model-db/css-117/models/v0.6/css-117/core96_v0.6/rapl/AbsPower/BPFOnly/KNeighborsRegressorTrainer_1
sidecar: true
total:
initUrl: >-
https://raw.githubusercontent.com/sunya-ch/kepler-model-db/css-117/models/v0.6/css-117/core96_v0.6/acpi/AbsPower/BPFOnly/GradientBoostingRegressorTra
sidecar: true
exporter:
deployment:
image: 'quay.io/sustainable_computing_io/kepler:latest-libbpf'
nodeSelector:
kubernetes.io/os: linux
port: 9103
metrics:
cgroups:
exposed: false
counter:
exposed: false
kubelet:
exposed: false
samplingInterval: '3'
serviceMonitor:
prometheusPoolingInterval: 30s
status:

Workloads
Networking
Storage
Builds
Observe
Alerting
Metrics
Dashboards
Targets
Compute
Nodes
Machines
MachineSets

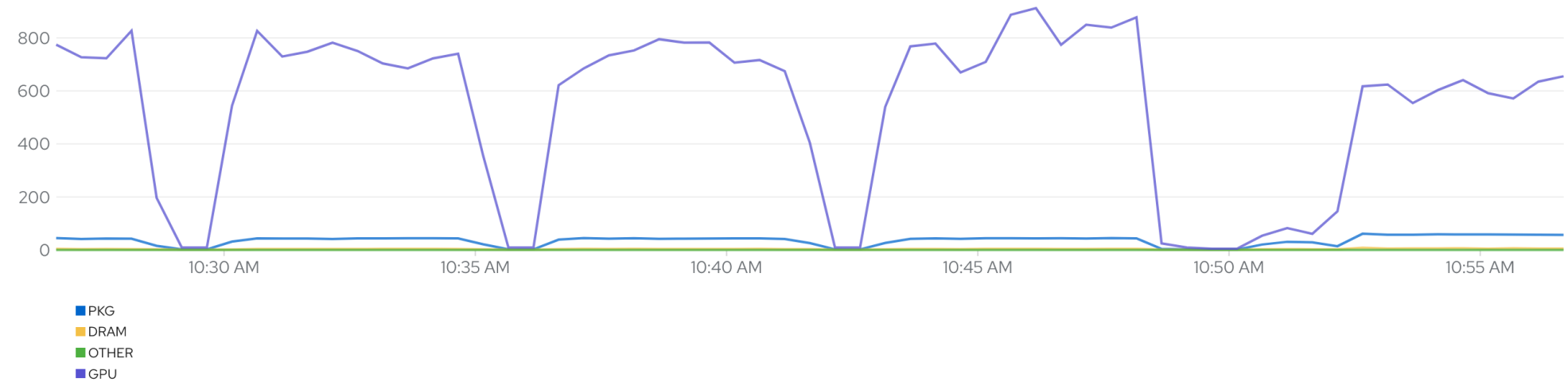
Dashboards

Time Range: Last 30 minutes
Refresh Interval: 30 seconds

Dashboard: Power Monitoring / Namespace
Namespace: rina
Pod: All

Total Power Consumption (W) in Namespace

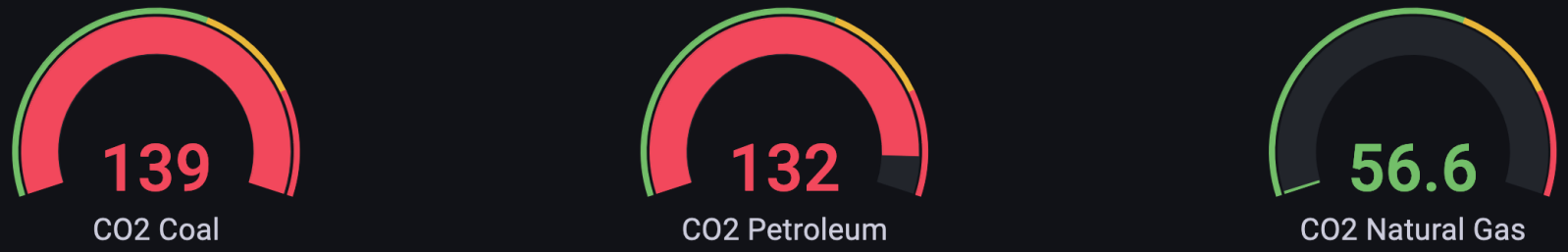
Inspect



datasource: Prometheus | Namespace: All | Pod: All | Coal Coeff: 2.23 | Natural Gas Coeff: 0.91 | Petroleum Coeff: 2.13

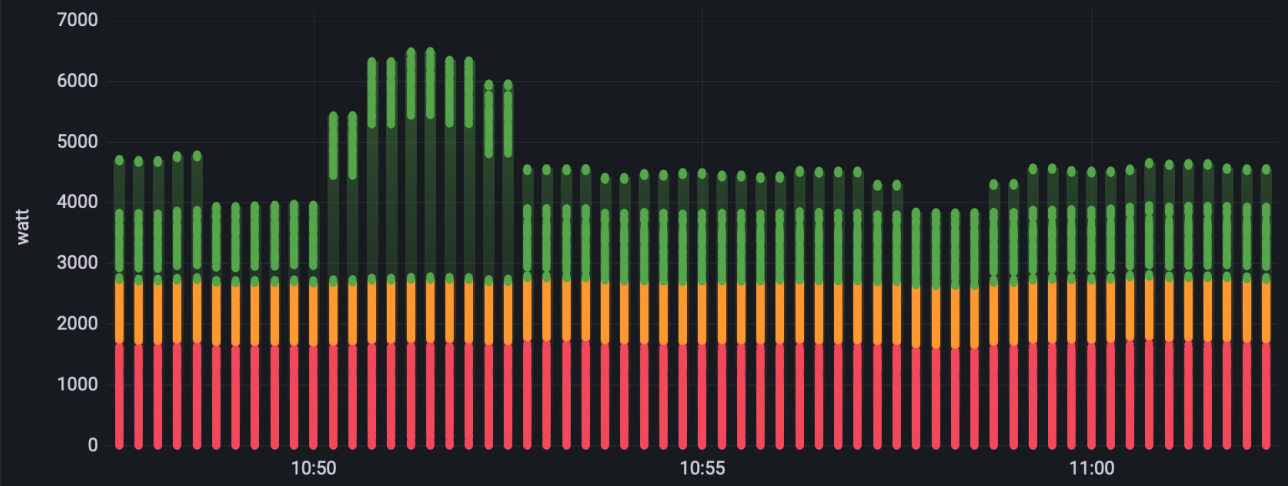
Carbon Emissions

Carbon Footprint (pounds per kWh per day) in Namespace: All (PKG+DRAM)



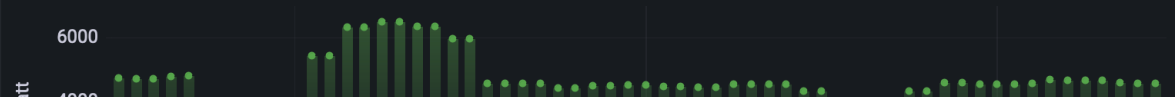
Power Consumption

Pod/Process Power Consumption (W) in Namespace: All

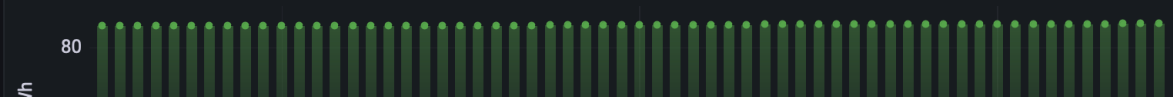


	Mean
fmperf-static-llama-13b-4gpu-inference-server-78c4d97649-7x48j / rina / GPU	513
granite-13b-cos-v2-master-0 / default / GPU	457
fmperf-static-llama-13b-8gpus-inference-server-74cbb6b998-7zc9k / rina / GPU	454
system_processes / system / PKG	91.2
multus-cpcnj / openshift-multus / GPU	46.9
fmperf-static-llama-13b-8gpus-inference-server-74cbb6b998-7zc9k / rina / PKG	45.1
multus-additional-cni-plugins-dt9vc / openshift-multus / GPU	31.3
multus-additional-cni-plugins-bplnw / openshift-multus / GPU	30.2
ibm-vpc-block-csi-controller-fcc4466c9-dl862 / openshift-cluster-csi-drivers / GPU	26.7
system_processes / system / DRAM	25.7
fmperf-static-llama-13b-4gpu-inference-server-78c4d97649-7x48j / rina / PKG	25.7
ovnkube-node-jlzt8 / openshift-ovn-kubernetes / GPU	22.4
nvidia-driver-daemonset-412.86.202306132230-0-dclg2 / nvidia-gpu-operator / GPU	22.4

Total Power Consumption (W) in Namespace: All



Total Power Consumption (kWh per day) in Namespace: All

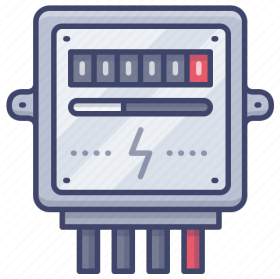



KubeCon | CloudNativeCon

North America 2023

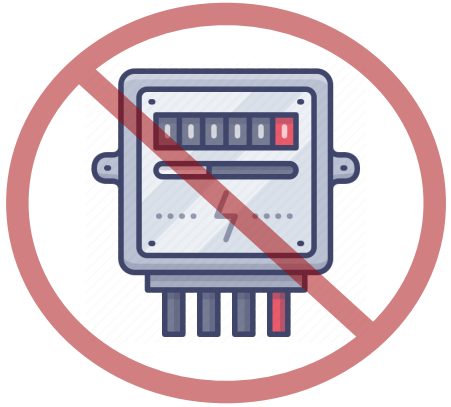
Updates related to Kepler deployment

Kepler different deployment motivation



Hardware Power Meter

- On Bare-Metal nodes
- Kepler can collect power metrics from hardware sensors and distribute this power to the running processes
- On Bare-metal Kepler can collect hardware counters
- With real-time power metrics the process power accuracy is very high

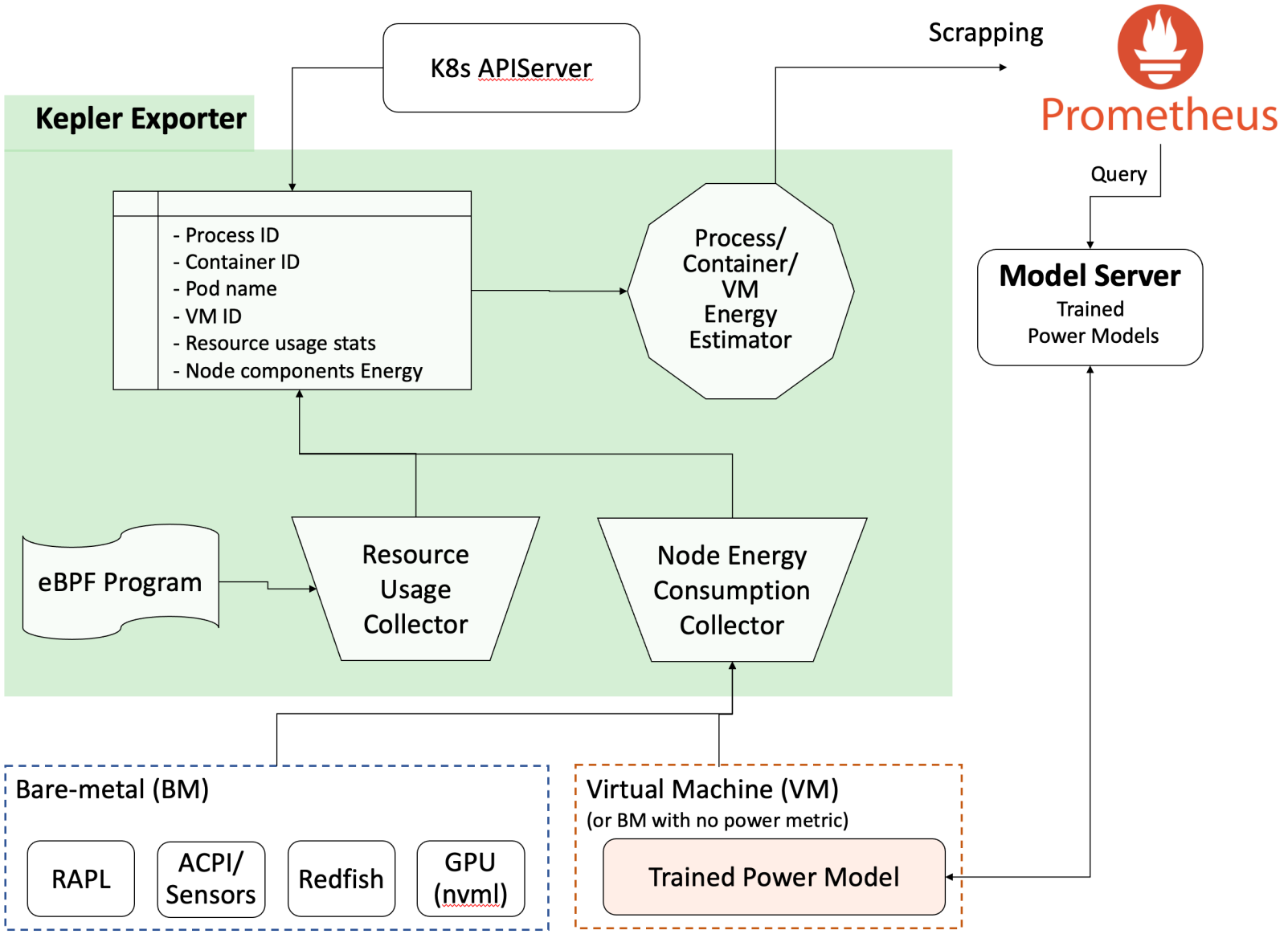


Without Power Meter

- On Virtual-Machine nodes
- Kepler cannot collect power metrics
- Trained Power models with data from BM are needed
- But power models are specific to architectures
- Idle power cannot be reported because we cannot split it without the knowledge of running VMs in the Host

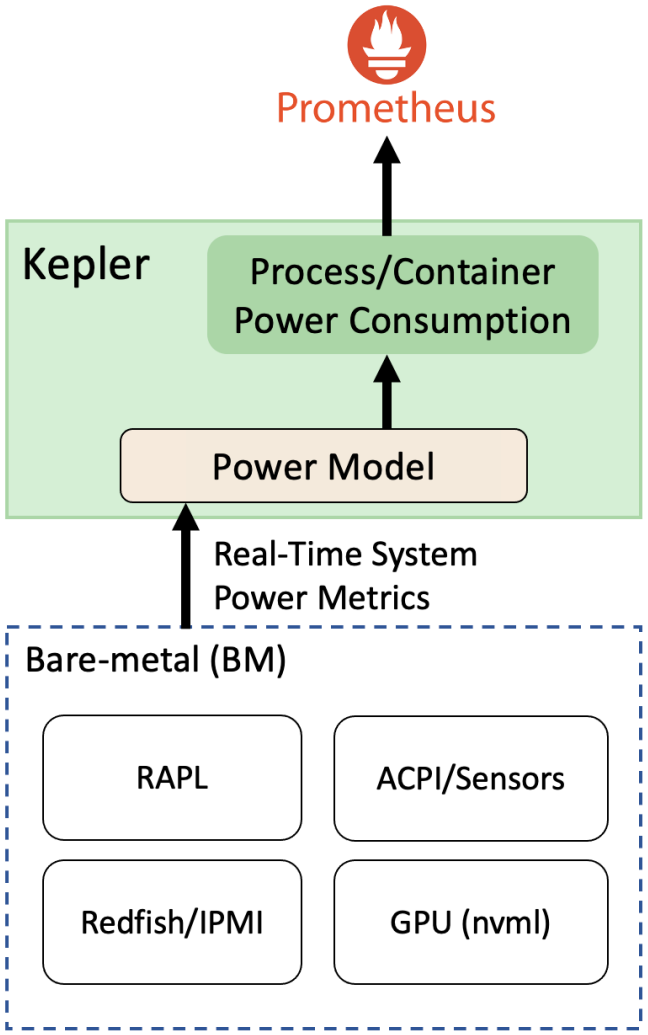
Kepler Architecture

- **Kepler** estimates the process power consumption based on the system power consumption using the resource utilization
- For example, the **Ratio Power Model** has the premise of if a process has 10% of CPU utilization, 10% of the power utilization is related to this process
- **Process resource utilization** is determined by eBPF counters
- If a system does not have power metrics, such as VMs, a trained Power Model via regression will be used

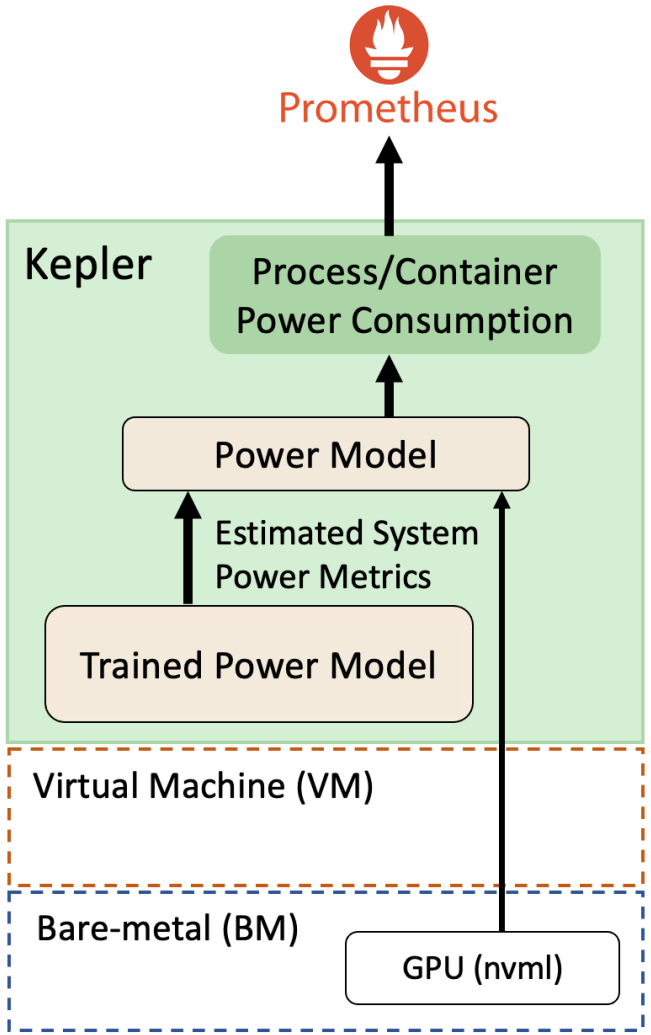


Kepler Deployment Approaches

1) Direct Real-Time System Power Metrics

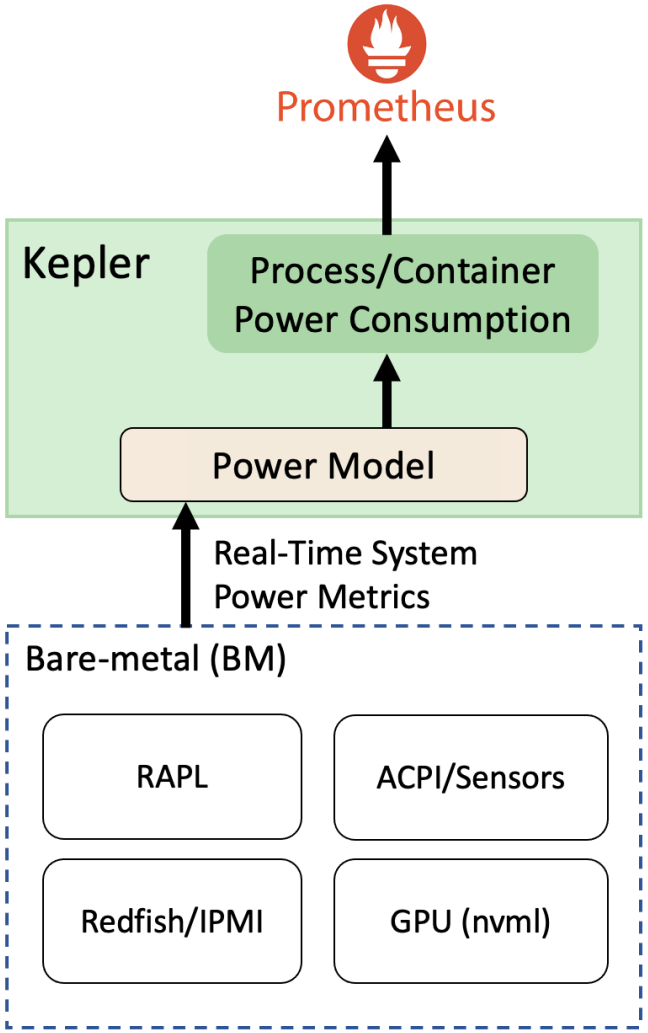


2) Estimated System Power Metrics for a VM

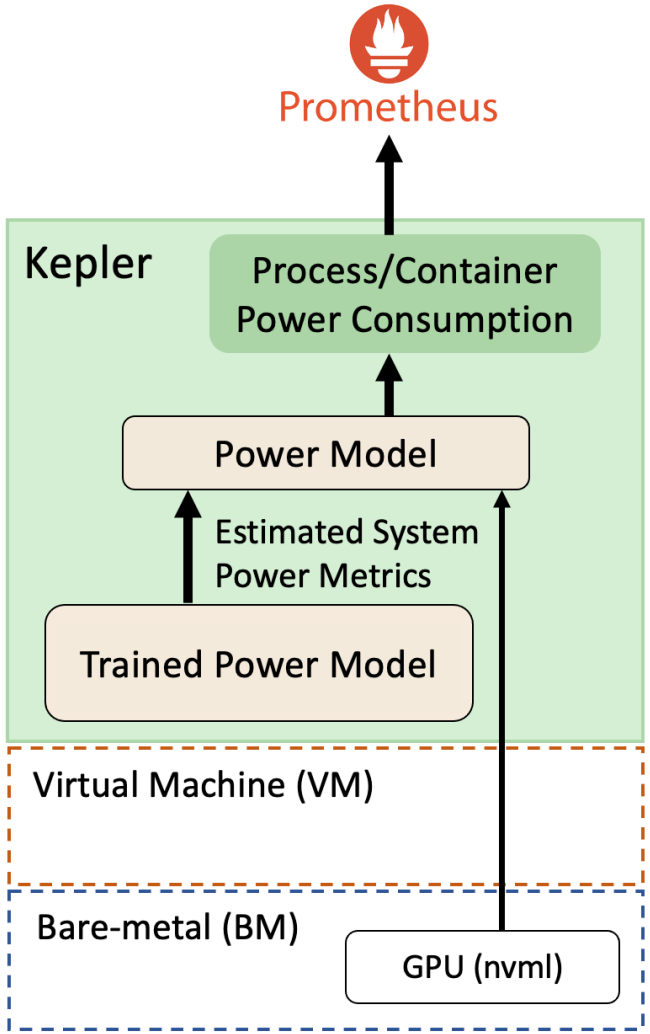


Kepler Deployment Approaches

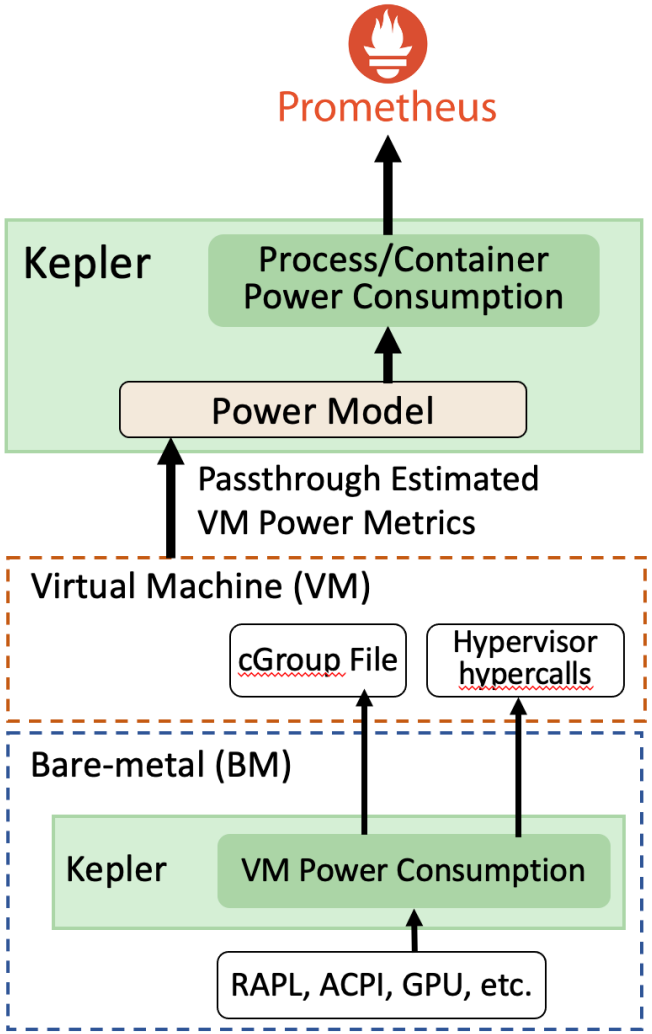
1) Direct Real-Time System Power Metrics



2) Estimated System Power Metrics for a VM



3) Passthrough Estimated VM Power Metrics




KubeCon | CloudNativeCon

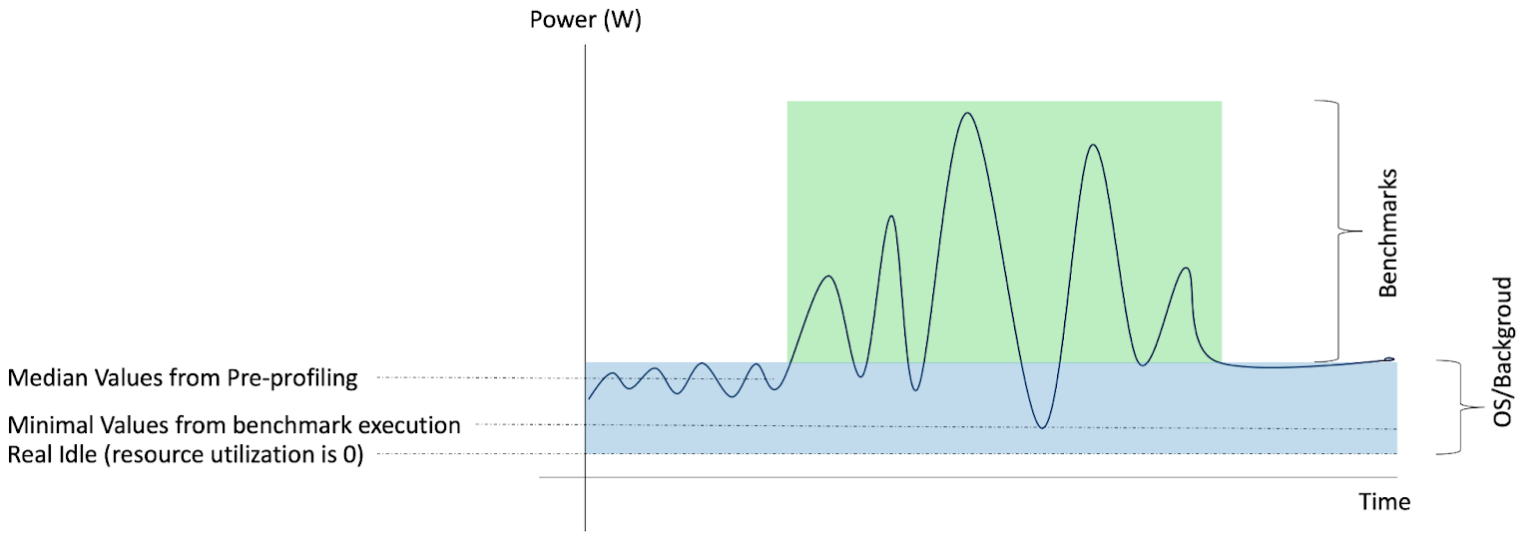
North America 2023

Updates related to BM deployment

Bare-Metal Power Model

- Idle Power
- Kepler standalone
- Process power and aggregate to Container, Pod and VMs

Bare-Metal Idle Power



- **Idle Power** refers to the power consumption **not related to resource utilization**
- Measuring idle power consumption is important for getting an accurate picture of overall energy usage
- Therefore, **it's not fair to assign all idle energy to a process only considering the resource utilization** since the idle power does not vary with the utilization
- The GHG protocol define that constant power consumption (e.g. idle power) should be divided based on the application size
- However, as process and containers typically do not have resource limits, **we currently evenly divide the idle power among all user processes + OS/Background processes**

Updates related to VM deployment

Virtual-Machine Power Model

- Estimate memory operation without hardware counter is challenge
- Memory utilization (e.g., read and write) depends on several factors from CPU cache to virtual memory and disk cache
- Memory allocation does not represent memory utilization
- CPU utilization cannot be used to estimate memory power consumption, otherwise any process using CPU will have memory power even if it is not using memory
- From the Software Counter perspective, we can estimate the memory operation related to virtual memory
- Therefore, we enabled the metric **Virtual Memory Page Cache Hit**
- Page Cache Hit can account memory operations when the application is using virtual memory, but memory operation directly from the hardware (CPU cache -> memory) will not be accounted
- The new DRAM power model will include this metric


KubeCon | CloudNativeCon

North America 2023

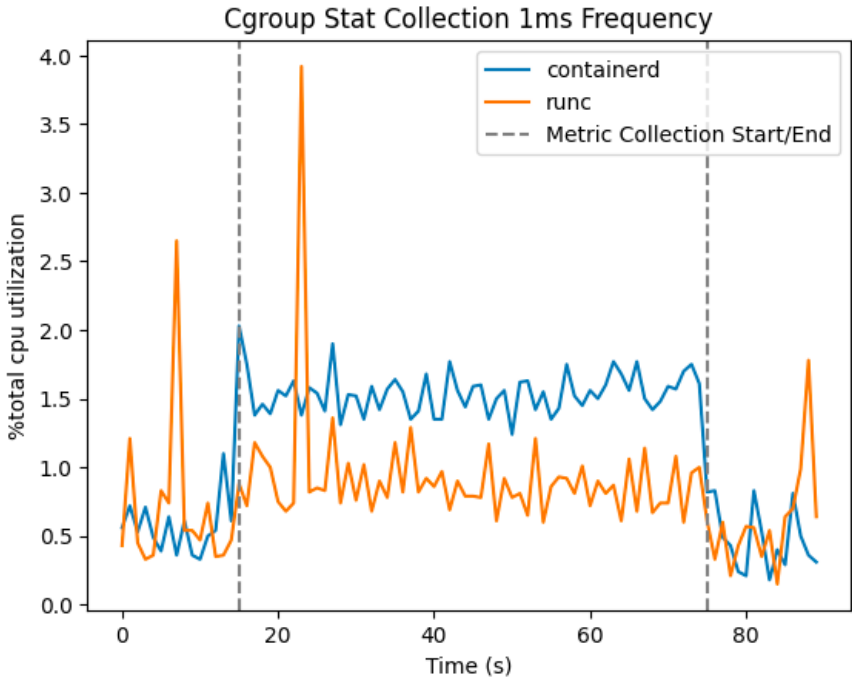
Updates related to Performance Analysis and Optimizations

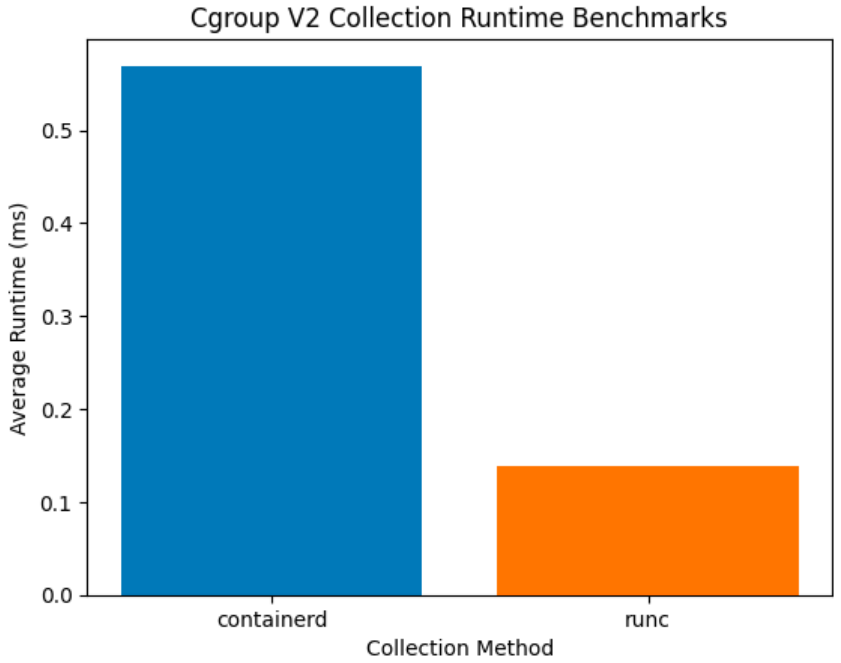
Performance Analysis and Optimization Update

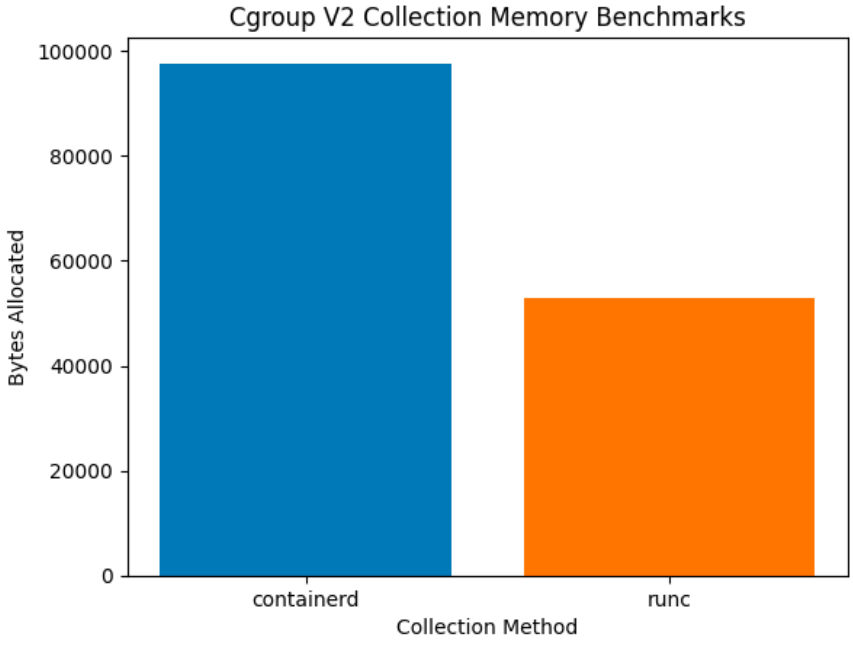
- Updated library to collect cGroup v2 metrics
- Disabled cGroup and Kubelet metrics
- Increased the Prometheus scrape interval

Performance Analysis and Optimization Update

- We replaced the use of the containerd package for collecting cgroup stats with the analogous package from runc when using v2 cgroups
- This reduces overhead as runc's package has lower overhead for collecting v2 cgroups
- Up to 25% less overhead



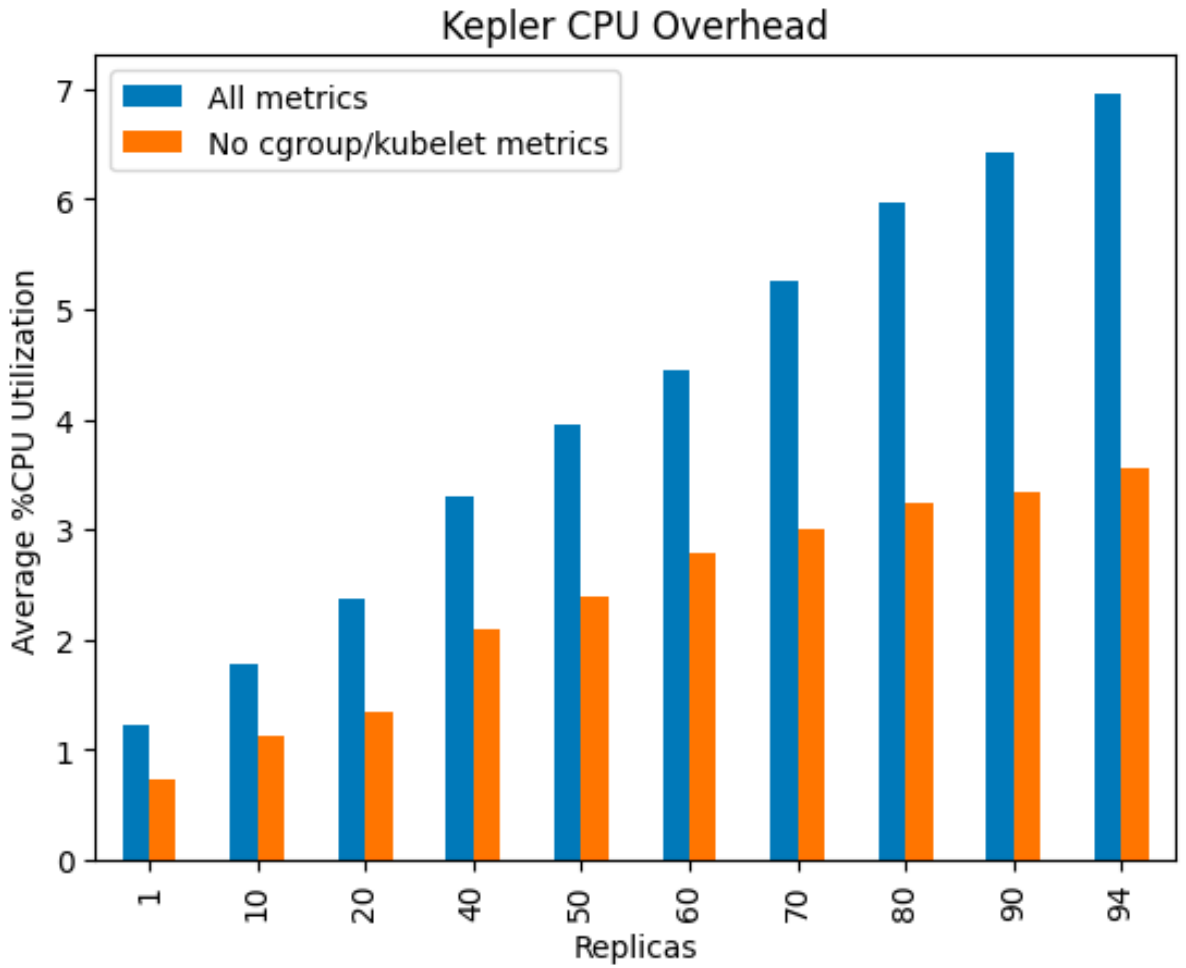




https://github.com/sustainable-computing-io/kepler/pull/840

Performance Analysis and Optimization Update

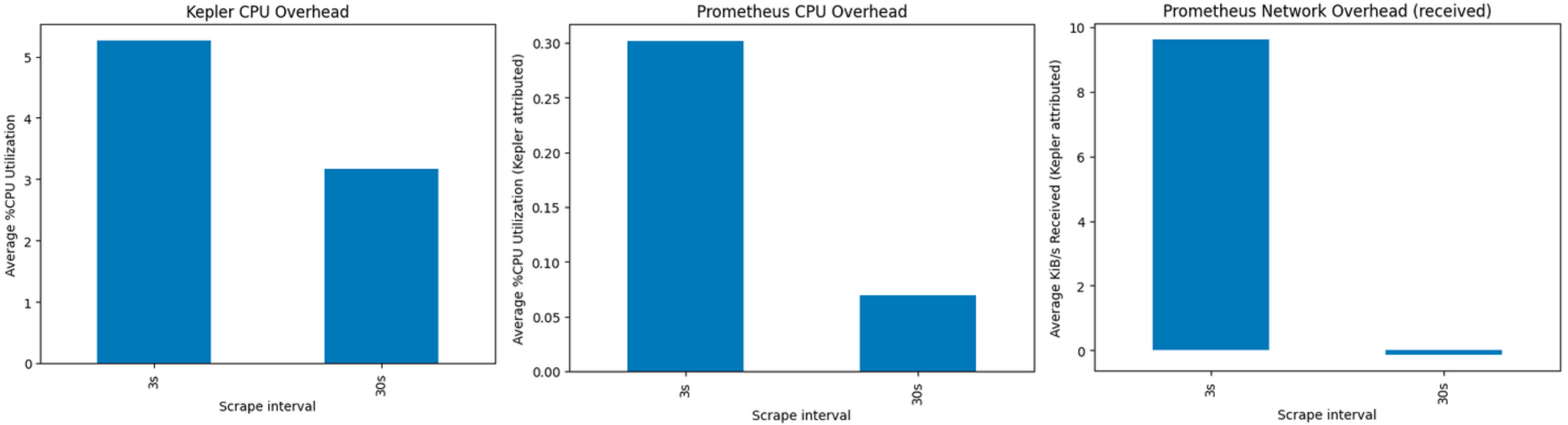
- **The newer default power models is based on BPF metrics**
- Therefore, we can disable the **cgroup and kubelet metrics** by default
- Disabling the cgroup and kubelet metrics **reduces 42% of the Kepler CPU overhead**



https://github.com/sustainable-computing-io/kepler/discussions/915
https://github.com/sustainable-computing-io/kepler/pull/983

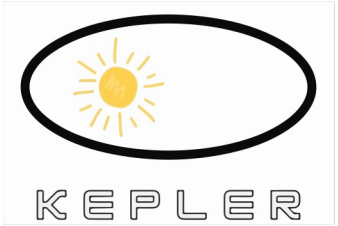
Performance Analysis and Optimization Update

- Originally, we set the Prometheus scraping interval to 3 seconds because Kepler collects metrics internally at a 3-second interval
- However, Prometheus typically scrapes metrics at a default interval of 30 seconds to reduce overhead
- As a result, **we have adjusted the Kepler service monitor to have the Prometheus scraping interval of 30 seconds**



https://github.com/sustainable-computing-io/kepler/issues/963
https://github.com/sustainable-computing-io/kepler/pull/970

Future Works



Inovation Timeline

Fev 2022

Birth of Kepler Red Hat working to measure container's Power Consumption on Kubernetes and OpenShift

IBM joined efforts with Red Hat

Jun 2022

Trained Power Models
VMs on public cloud does not expose power metrics, as a result Kepler needed to create power models

Model Server
a project to create power models from regression of data collected from bare-metal nodes

Idle Power
accordingly, to the GHG protocol for a fair power distribution among application the constant idle power should be divided based on the application size and dynamic power based on the load

Mar 2023

Kepler Standalone
to support Edge computing use cases, developers added the option to run Kepler outside Kubernetes and install via RPM

Process Power Consumption
developers added the support to measure the power per process not only containers

Jun 2023

CNCF Sandbox

Oct 2023

Disabled cGroup and Kubelet metric collection
the newer power model relies only on BPF metrics and disabling cGroup and Kubelet metrics further minimize the overhead

Increase Prometheus scrape Interval
to minimize overhead

2024

OpenTelemetry

eBPF Sampling Rate
to minimize overhead

VM power passthrough
expose VM power within the VM, this use case could be used by cloud provides in the future

New Idle Power Logic
needs regression to determine the true Idle power (e.g., with 0 resource utilization)

Thanks to Kepler Contributors

IBM

Marcelo Amaral
Sunyanan Choochotkaew
Rina Nakazawa
Chen Wang
Sam Yuan
Qi Feng Huo
Max Calman
Peng Hui Jiang
Peng Li
Xiao-Peng Zhang
Guang Han Sui
John Rofrano
Scott Trent

Huamin Chen
Ji Chen
Parul Singh
Kaiyi Liu
Sally O'Malley
William Caban
Sunil Thaha
Sally O'Malley
Vibhu Prashar
Chris Laprun
Chris Procter
Anthony Aharivel
Adrian Hammond
Andreas Spanner
Vimal Kumar
Ryan Cook
Holly Cummins
Ramachandran Ravi
Gabriel Bernal
Marc Methot
David Szegedi

intel

Jie Ren
Filip Skirtun
Ruomeng Hao
Lu Ken
Lei Zhou

NTT
NIPPON TELEGRAPH AND TELEPHONE CORPORATION

Takuya Iwatsuka
Yasumasa Suenaga

Niki Manoledaki

DB Systel GmbH

Bastien Grasnick

Apex®

Yanbo Xu

encora

Alyson Deives

ryax technologies

Michael Mercier

EnableIT

Sanskar
Bhushan

Brad McCoy

Sorry if I missed someone…

Questions?

Please scan the QR Code above
to leave feedback on this session