Kubernetes Batch and Other News

Quick Update

Ricardo Rocha, CERN

CNCF Research User Group

https://community.cncf.io/research-end-user-group/ Agenda

1st and 3rd Wednesdays, 5pm CET / 8am PT

Discussion and advancement of Research Computing using *Cloud Native*

Topics on Batch, Baremetal Deployments, Notebooks, etc

https://www.youtube.com/results?search_query=cncf+research+user+group

Upcoming Events: GitOps Workshop @ CERN

https://indico.cern.ch/event/1145174/

Half day event, April 27th

Several use cases reporting on their choices and experiences

- ArgoCD, Flux and GitLab CI

Discussion to decide next steps and try some consolidation

Upcoming Events: Kubecon Europe 2022

Upcoming Events: Kubecon Europe 2022

Hybrid Event

- Expecting ~5000 people in person
- 10000+ additional attendees virtually

Academic / Non Profit registration for CERN people (CERN ID) … $USD 150

Unlimited free virtual passes, also valid for the co-located events

https://codimd.web.cern.ch/eGEghe2RRpezdNDuSdEYlA

Monday, May 16

08:45 CEST

- Data on Kubernetes Day 2022 Europe hosted by CNCF (Complimentary Registration Required)

09:00 CEST

- Cloud Native eBPF Day Europe Hosted By CNCF (Additional Registration + Fee Required)
- Cloud Native Telco Day Europe Hosted By CNCF (Additional Registration + Fee Required)
- Cloud Native Wasm Day Europe Hosted By CNCF (Additional Registration + Fee Required)
- Day 1: Cloud Native SecurityCon Europe Hosted by CNCF, Track 1 (Additional Registration + Fee Required)
- FluentCon Europe Hosted By CNCF (Additional Registration + Fee Required)
- Kubernetes AI Day Europe Hosted by CNCF (Additional Registration + Fee Required)

18:00 CEST

- KubeCon + CloudNativeCon Happy Hour hosted by Harness + Snyk (Complimentary Registration Required)

Tuesday, May 17

09:00 CEST

- Day 2: Cloud Native SecurityCon Europe Hosted by CNCF, Track 1 (Additional Registration + Fee Required, Includes both Tracks)
- Day 2: Cloud Native SecurityCon Europe Hosted by CNCF, Track 2 (Additional Registration + Fee Required, Includes both Tracks)
- GitOpsCon Europe hosted by CNCF, Track 1 (Additional Fee + Registration Required, Includes both Tracks)
- GitOpsCon Europe hosted by CNCF, Track 2 (Additional Fee + Registration Required, Includes both Tracks)
- KnativeCon Europe Hosted By CNCF (Additional Registration + Fee Required)
- Kubernetes Data Workshop hosted by Portworx by Pure Storage (Additional Registration + Fee Required)
- Kubernetes on Azure Day at KubeCon + CloudNativeCon Europe 2022 hosted by Microsoft Azure (Additional Registration + Fee Required)
- Kubernetes on Edge Day Europe Hosted by CNCF (Additional Registration + Fee Required)
- OpenShift Commons Gathering hosted by Red Hat (Complimentary Registration Required)
- PrometheusDay Europe Hosted By CNCF (Additional Registration + Fee Required)
- ServiceMeshCon Europe Hosted By CNCF (Additional Registration + Fee Required)

Monday, May 16

08:45 CEST

- Data on Kubernetes Day 2022 Europe hosted by CNCF (Complimentary Registration Required)

09:00 CEST

- Cloud Native eBPF Day Europe Hosted By CNCF (Additional Registration + Fee Required)
- Cloud Native Telco Day Europe Hosted By CNCF (Additional Registration + Fee Required)
- Cloud Native Wasm Day Europe Hosted By CNCF (Additional Registration + Fee Required)
- Day 1: Cloud Native Secur (Additional Registration + F
- FluentCon Europe Hosted Required)
- Kubernetes AI Day Europe Fee Required)

10:00 CEST

- Build with the Most Automated and Scalable Kubernetes hosted by Google Cloud (Complimentary Registration Required)
- CDEventsCon hosted by the CD Foundation (Complimentary Registration Required)

13:00 CEST

- Troubleshoot Kubernetes Clusters hosted by Replicated (Complimentary Registration Required)
- Kubernetes Batch + HPC Day Europe Hosted By CNCF (Additional Registration + Fee Required)

18:00 CEST

- KubeCon + CloudNativeCo (Complimentary Registration Required)

Tuesday, May 17

09:00 CEST

- Day 2: Cloud Native SecurityCon Europe Hosted by CNCF, Track 1 (Additional Registration + Fee Required, Includes both Tracks)
- Day 2: Cloud Native SecurityCon Europe Hosted by CNCF, Track 2 (Additional Registration + Fee Required, Includes both Tracks)
- GitOpsCon Europe hosted by CNCF, Track 1 (Additional Fee + Registration Required, Includes both Tracks)
- GitOpsCon Europe hosted by CNCF, Track 2 (Additional Fee + Registration Required, Includes both Tracks)
- KnativeCon Europe Hosted By CNCF (Additional Registration + Fee Required)
- Kubernetes Data Workshop hosted by Portworx by Pure Storage + Fee Required)
- ay at KubeCon + CloudNativeCon Europe 2022 re (Additional Registration + Fee Required)
- ay Europe Hosted by CNCF (Additional uired)
- athering hosted by Red Hat (Complimentary
- e Hosted By CNCF (Additional Registration + Fee
- e Hosted By CNCF (Additional Registration + Fee Required)

https://events.linuxfoundation.org/kubernetes-batch-hpc-day-europe/

13:00 CEST	Opening + Welcome - Abdullah Gharaibeh & Ricardo Rocha, Kubernetes Batch + HPC Day Program Committee Members
13:10 CEST	Keynote: High Performance Computing on Google Kubernetes Engine- Maciek Różacki, Google Cloud
13:15 CEST	Kueue: A Kubernetes-native Job Queueing - Abdullah Gharaibeh, Google
13:45 CEST	Resource Orchestration of HPC on Kubernetes: Where We Are Now and the Journey Ahead! - Swati Sehgal & Francesco Romani, Red Hat
14:15 CEST	Volcano – Cloud Native Batch System for AI, BigData and HPC - William(LeiBo) Wang, Huawei Cloud Computing Co., Ltd
14:45 CEST	How to Handle Fair Scheduling in a Private Academic K8s infrastructure - Lukas Hejtmanek, Masaryk University & Dalibor Klusacek, CESNET
14:55 CEST	Coffee Break + Networking

15:10 CEST	Best Practices Considerations When Running MPI-Operator at Scale - Carlos Eduardo Arango Gutierrez, Red Hat
15:25 CEST	Get More Computing Power by Helping the OS Scheduler - Antti Kervinen, Intel & Alexander Kanevskiy, Intel
15:35 CEST	Fast Data on-Ramp with Apache Pulsar on K8 - Timothy Spann, StreamNative
15:50 CEST	Apache YuniKorn A Kubernetes Scheduler Plugin for Batch Workloads - Wilfred Spiegelenburg, Cloudera & Craig Condit, Cloudera
16:20 CEST	Efficient Deep Learning Training with Ludwig AutoML, Ray, and Nodeless Kubernetes - Anne Marie Holler, Elotl & Travis Addair, Predibase
16:45 CEST	Closing - Aldo Culquicondor, Kubernetes Batch + HPC Day Program Committee Member
17:00 CEST	CNCF-hosted Co-located Events Happy Hour

Batch and HPC

Motivation

Enhance the support for Batch (eg. HPC, AI/ML) workloads in Kubernetes

Unify the way users deploy batch workloads, improve portability

Enhancements

- **Extend the batch API** group (Job, CronJob)
- Add Job level **queueing**, potentially multi-cluster
- Improve runtime and scheduling support for **accelerators**

Out of Scope: workflows, pipelines, …

Initiatives

Batch Working Group in Kubernetes

- Most active: organized by Apple, Google, VMWare, RedHat, Intel
- Meetings on Thursdays 7am and 3pm PT (alternating)
- Focus on support in upstream Kubernetes, working closely with SIGs
- https://github.com/kubernetes/community/tree/master/wg-batch

CNCF Batch System Initiative

- Slow start, promoted by projects like Volcano, Armada, …
- Batch system specification to be incorporated into Kubernetes, Volcano, Armada, etc
- https://github.com/cncf/tag-runtime/issues/38

Batch WG Roadmap

http://bit.ly/wg-batch-roadmap

Job API

- Multi-pod templates, resizable jobs, completion policies
- Support features required for MPI, TensorFlow, Spark, …
- Integrate features required for workflows (Tekton, Argo, Kubeflow, …)

Batch WG Roadmap

http://bit.ly/wg-batch-roadmap

Job Management

- Queueing, Co-Scheduling, Fair Share
- Job level preemption, bulk provisioning in the cluster auto scaler
- Multi cluster support, likely later in the effort
- Look at existing schedulers
  - Kubernetes: Volcano, Yunikorn, Kueue, …
  - Traditional: SLURM, HTCondor, …

Batch WG Roadmap

http://bit.ly/wg-batch-roadmap

Specialized Hardware, Accelerators

- NUMA awareness in the upstream scheduler
- Improved GPU and generic resource scheduling

Kueue

Kubernetes Native Job Queueing

CNCF Research User Group Presentation
https://www.youtube.com/watch?v=ft5kBOFcXqg

Main Features

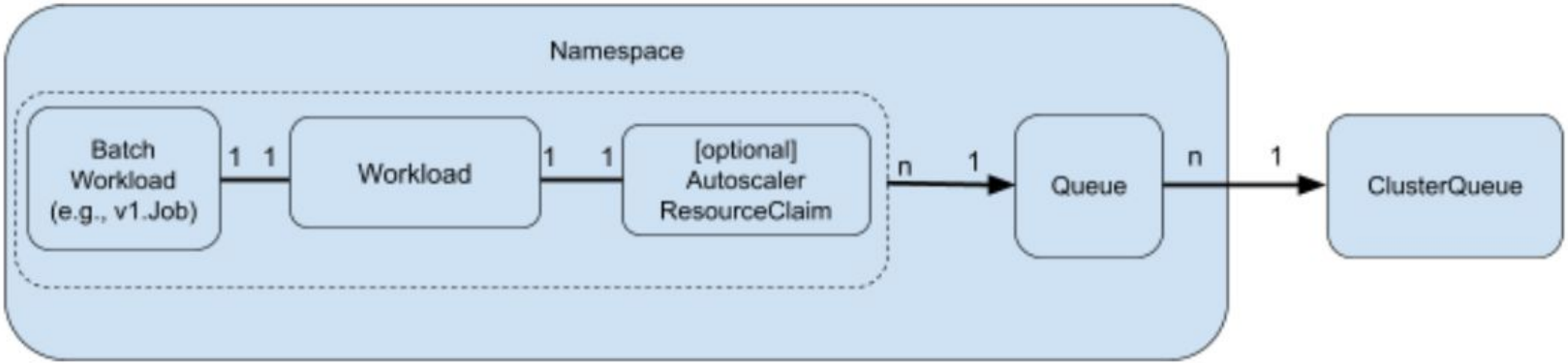
Allow sharing unused capacity, *max-min* ***fairness***

Queueing features, priorities and policies

> “*Use on demand up to committed use discounts, spot otherwise”*

Execution order, co-scheduling of pods for a Job, array jobs

Budgets to manage tenant resource usage over time


Namespace
Batch Workload (e.g., v1.Job)
1
1
Workload
1
1
[optional] Autoscaler ResourceClaim
n
1
Queue
n
1
ClusterQueue

Queues		

metadata:
  name: queue
  namespace: tenantA1
spec:
  clusterQueue: tenantA-cluster-queue
	metadata:
	  name: queue
	  namespace: tenantA2
	spec:
	  clusterQueue: tenantA-cluster-queue
		metadata:
		  name: queue
		  namespace: tenantB
		spec:
		  clusterQueue: tenantB-cluster-queue

Queues

```yaml
metadata:
 name: queue
 namespace: tenantA1
spec:
 clusterQueue: tenantA-cluster-queue
```

```yaml
metadata:
 name: queue
 namespace: tenantA2
spec:
 clusterQueue: tenantA-cluster-queue
```

```yaml
metadata:
 name: queue
 namespace: tenantB
spec:
 clusterQueue: tenantB-cluster-queue
```

ClusterQueues

```yaml
# Defines a quota for on-demand C2 machine type and k80 GPUs.
metadata:
  name: tenantA-cluster-queue
spec:
  namespaceSelector:
    matchExpressions:
    - key: tenant
      operator: In
      values:
      - tenantA
  requestableResources:
  - name: cpu
    flavors:
    - name: c2-on-demand
      min: 1000
      labels:
        cloud.provider.com/vm-family: c2
  - name: nvidia.com/gpu
    flavors:
    - name: a100
      min: 20
      labels:
        cloud.provider.com/accelerator: nvidia-tesla-a100
```

```yaml
# Defines a smaller quota for on-demand C2 and k80 GPUs.
metadata:
  name: tenantB-cluster-queue
spec:
  namespaceSelector:
    matchExpressions:
    - key: tenant
      operator: In
      values:
      - tenantB
  requestableResources:
  - name: cpu
    flavors:
    - name: c2-on-demand
      min: 100
      labels:
        cloud.provider.com/vm-family: c2
  - name: nvidia.com/gpu
    flavors:
    - name: k80
      min: 20
      labels:
        cloud.provider.com/accelerator: nvidia-tesla-k80
```

Queues

```yaml
metadata:
 name: queue
 namespace: tenantA1
spec:
 clusterQueue: tenantA-cluster-queue
```

```yaml
metadata:
 name: queue
 namespace: tenantA2
spec:
 clusterQueue: tenantA-cluster-queue
```

```yaml
metadata:
 name: queue
 namespace: tenantB
spec:
 clusterQueue: tenantB-cluster-queue
```

ClusterQueues

```yaml
# Defines a borrowing-cohort. TenantA can borrow up to 100 more
# C2 cores. A workload could start by using a100 GPUs from this
# cluster-queue and borrowed C2 cores from
# tenantB-cluster-queue. TenantA can't borrow k80 because the
# type is not defined in the ClusterQueue.
metadata:
  name: tenantA-cluster-queue
spec:
  cohort: borrowing-cohort
  namespaceSelector:
    matchExpressions:
    - key: tenant
      operator: In
      values:
      - tenantA
  requestableResources:
  - name: cpu
    flavors:
    - name: c2-on-demand
      min: 1000
      labels:
      - cloud.provider.com/vm-family: c2
  - name: nvidia.com/gpu
    flavors:
    - name: a100
      min: 20
      labels:
      - cloud.provider.com/accelerator: nvidia-tesla-a100
```

```yaml
# tenantB-cluster-queue is part of the "borrowing-cohort";
# however, by setting the borrowingWeight to 0, it can't borrow
# from tenantA-cluster-queue, but tenantA-cluster-queue can.


metadata:
  name: tenantB-cluster-queue
spec:
  cohort: borrowing-cohort
  borrowingWeight: 0
  namespaceSelector:
    matchExpressions:
    - key: tenant
      operator: In
      values:
      - tenantB
  requestableResources:
  - name: cpu
    flavors:
    - name: c2-on-demand
      min: 100
      labels:
      - cloud.provider.com/vm-family: c2
  - name: nvidia.com/gpu
    flavors:
    - name: k80
      min: 20
      labels:
      - cloud.provider.com/accelerator: nvidia-tesla-k80
```

Queues

```yaml
metadata:
 name: queue
 namespace: tenan
spec:
 clusterQueue: te
```

```yaml
metadata:
```

```yaml
metadata:
```

```
uster-queue
```

```
type WorkloadSpec struct {
  // The name of the Queue the workload is sent to.
  QueueName string

  // A workload may include one or more sets of pods of different specs.
  // This is needed to communicate to Kueue the resources needed by the
  // workload. This is supposed to be a copy from the workload object itself. In
  // theory, Kueue could use the WorkloadReference to query this information from
  // the workload object directly, but this is not practical because Kueue aims to
  // support custom workloads without having a dependency on the workload api
  // itself.
  PodSets []PodSet

  // priority determines the workload's order in the ClusterQueue.
  // Higher is more important.
  Priority int64

  // A reference to the actual workload resource. Kueue will use that to
  // start/stop the actual workload.
  WorkloadReference WorkloadReference
}
```

```yaml
      min: 20
      labels:
      - cloud.provider.com/accelerator: nvidia-tesla-a100
```

```yaml
    - name: k80
      min: 20
      labels:
      - cloud.provider.com/accelerator: nvidia-tesla-k80
```

Questions?