CERN openlab

Evaluating Kubernetes batch scheduling systems for containerized declarative data analyses

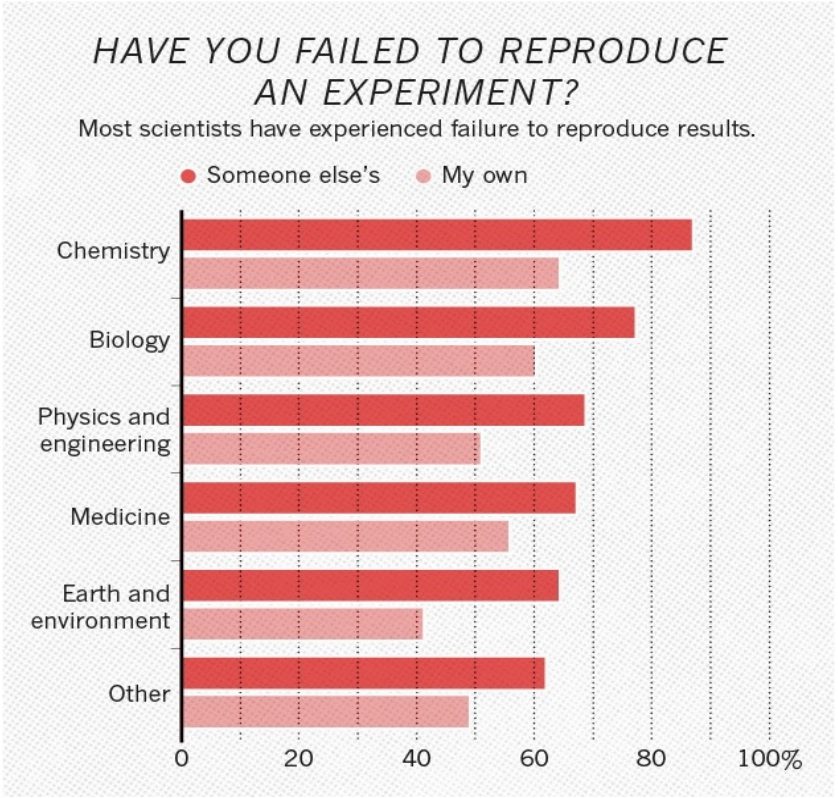
CERN openlab Summer Student Lightning Talks

Author:
Xavier Tintin

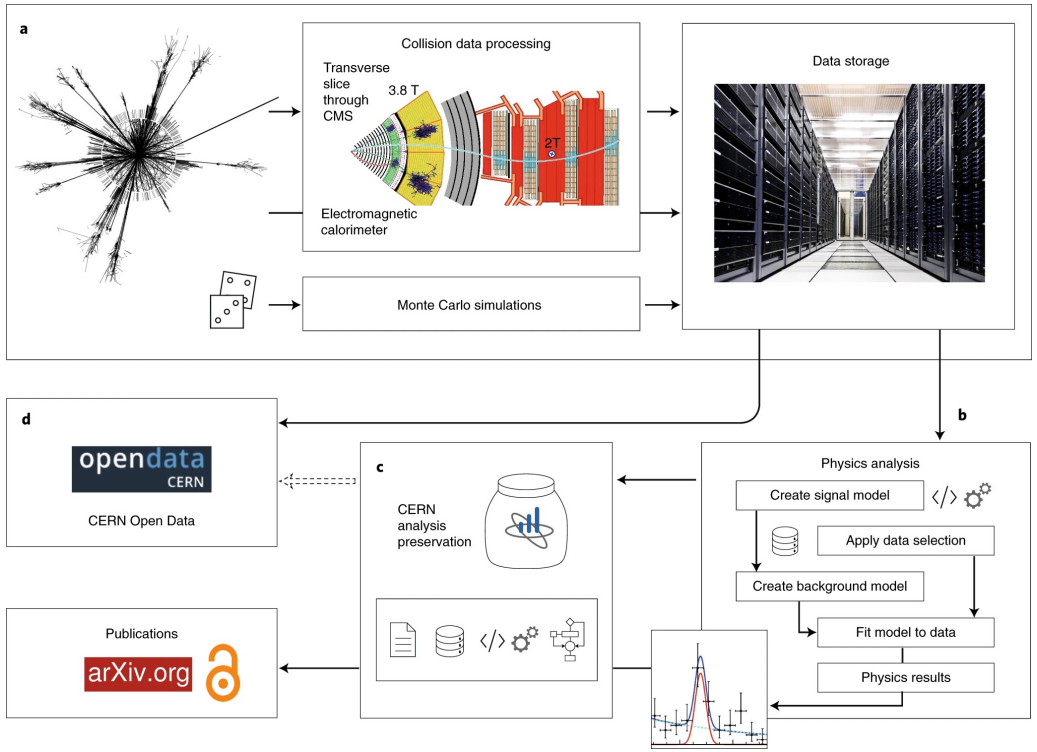
Supervisors:
Marco Donadoni
Tibor Šimko

Reproducible and reusable science

*When testing new theories, building on previous data analyses is necessary, but **very** challenging*



https://doi.org/10.1038/533452a

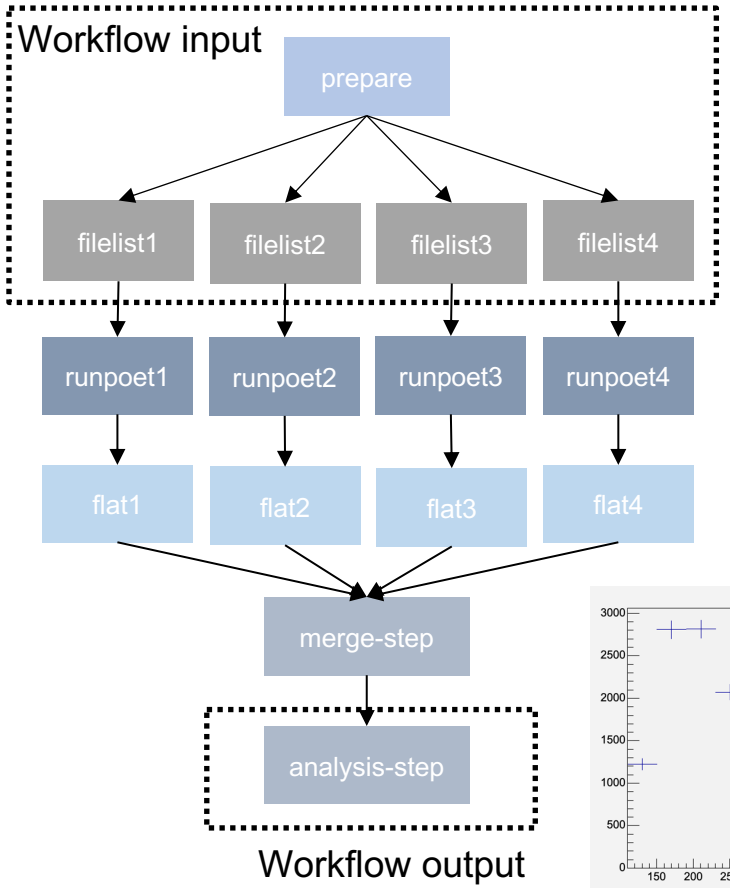


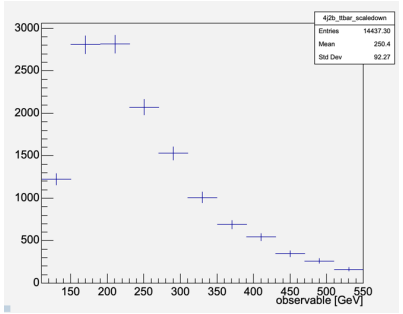
*Scientific workflows represent the **complex** flow of data that through various steps of collection, transformation, and analysis produces published results.*

Declarative HEP analysis workflows

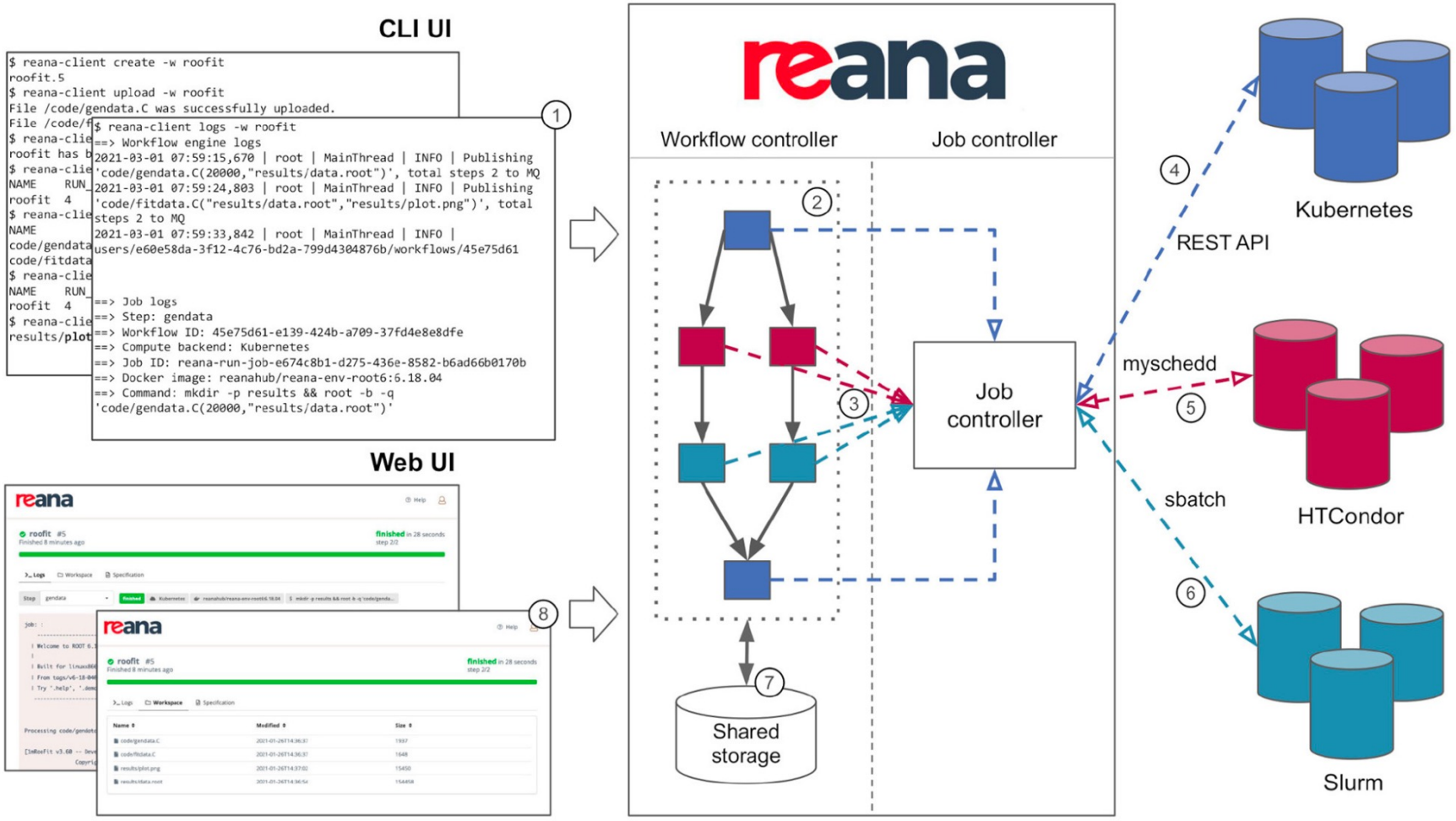
Workflows

- A workflow is a recipe explaining how to compute results from the input data
- What is needed?
  - Input data
  - Code
  - Computing environment
  - Workflow steps
- Each step can use different container images
- Steps can be described by means of declarative workflow languages



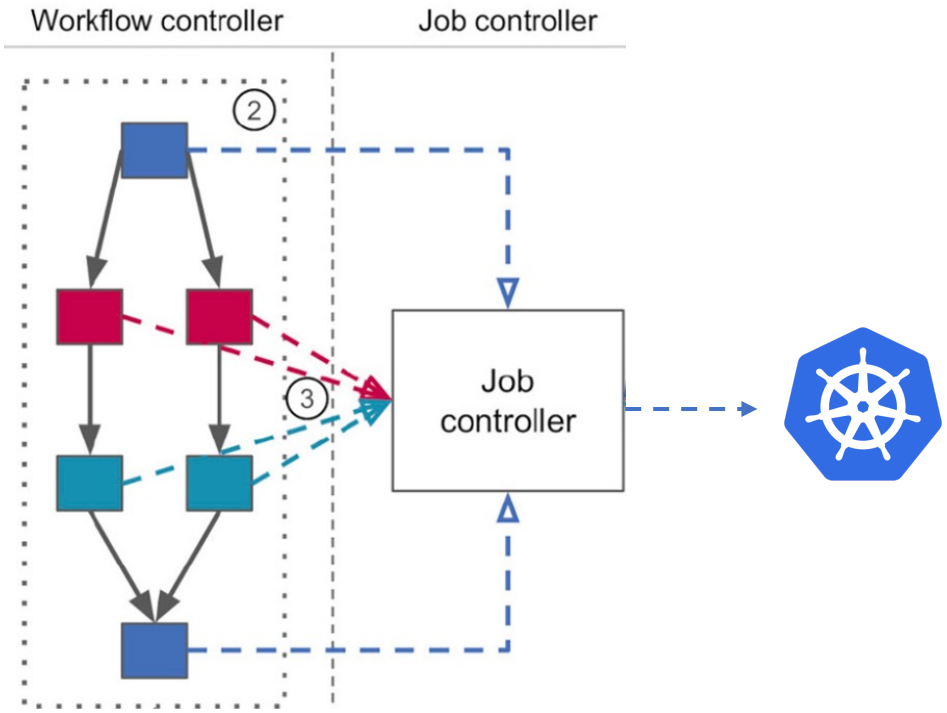


REANA: reusable analyses



Current solution: Kubernetes API

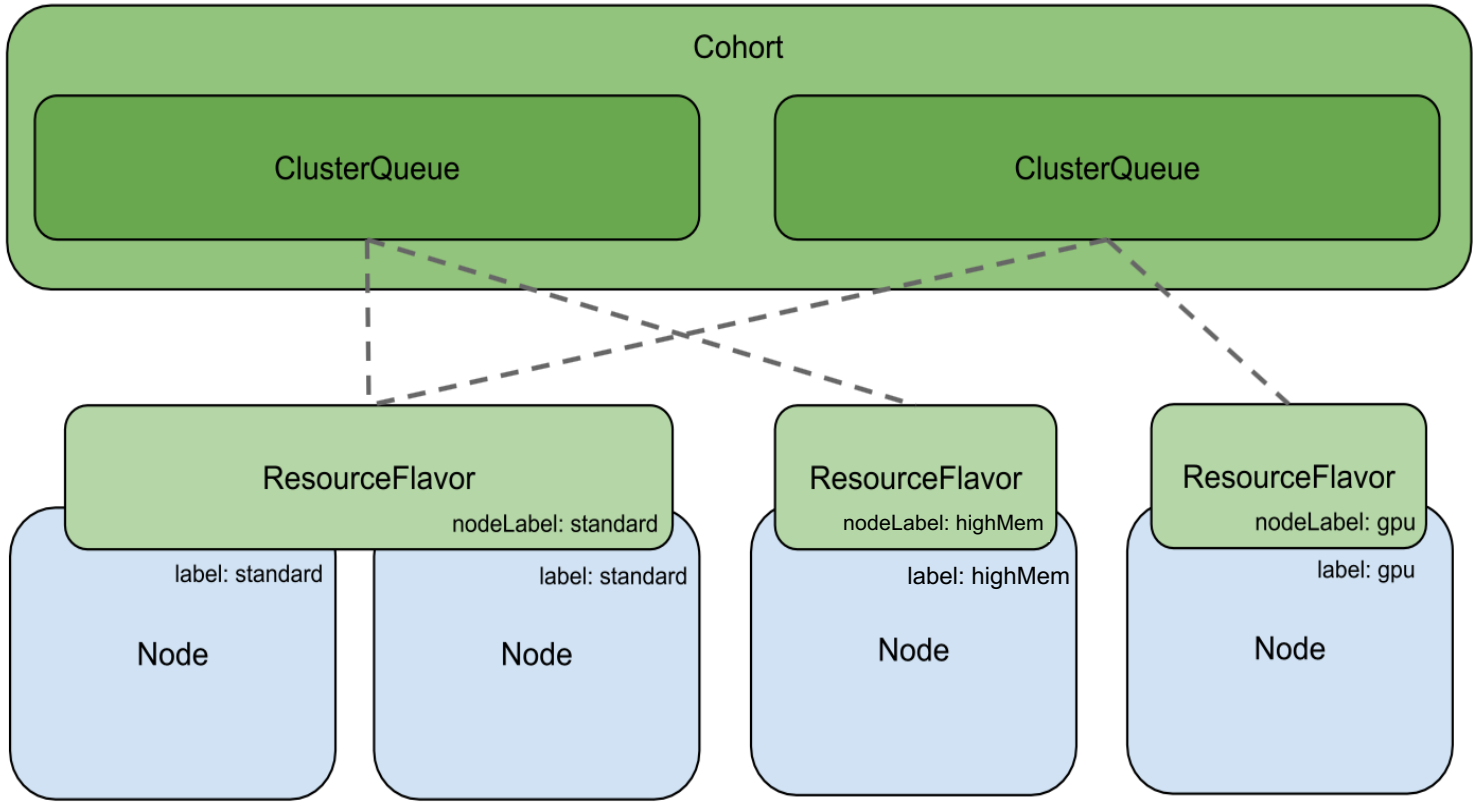
```python
class KubernetesJobManager(JobManager):
    @JobManager.execution_hook
    def execute(self):
        """Execute a job in Kubernetes."""
        backend_job_id = build_unique_component_name("run-job")
        self.job = {
            "kind": "Job",
            "apiVersion": "batch/v1",
            "metadata": {
                "name": backend_job_id,
                "namespace": REANA_RUNTIME_KUBERNETES_NAMESPACE,
            },
            "spec": {
                "backoffLimit": KubernetesJobManager.MAX_NUM_JOB_RESTARTS,
                "autoSelector": True,
                "template": {
                    "metadata": {
                        "name": backend_job_id,
                        "labels": {"reana-run-job-workflow-uuid": self.workflow_uuid},
                    },
```



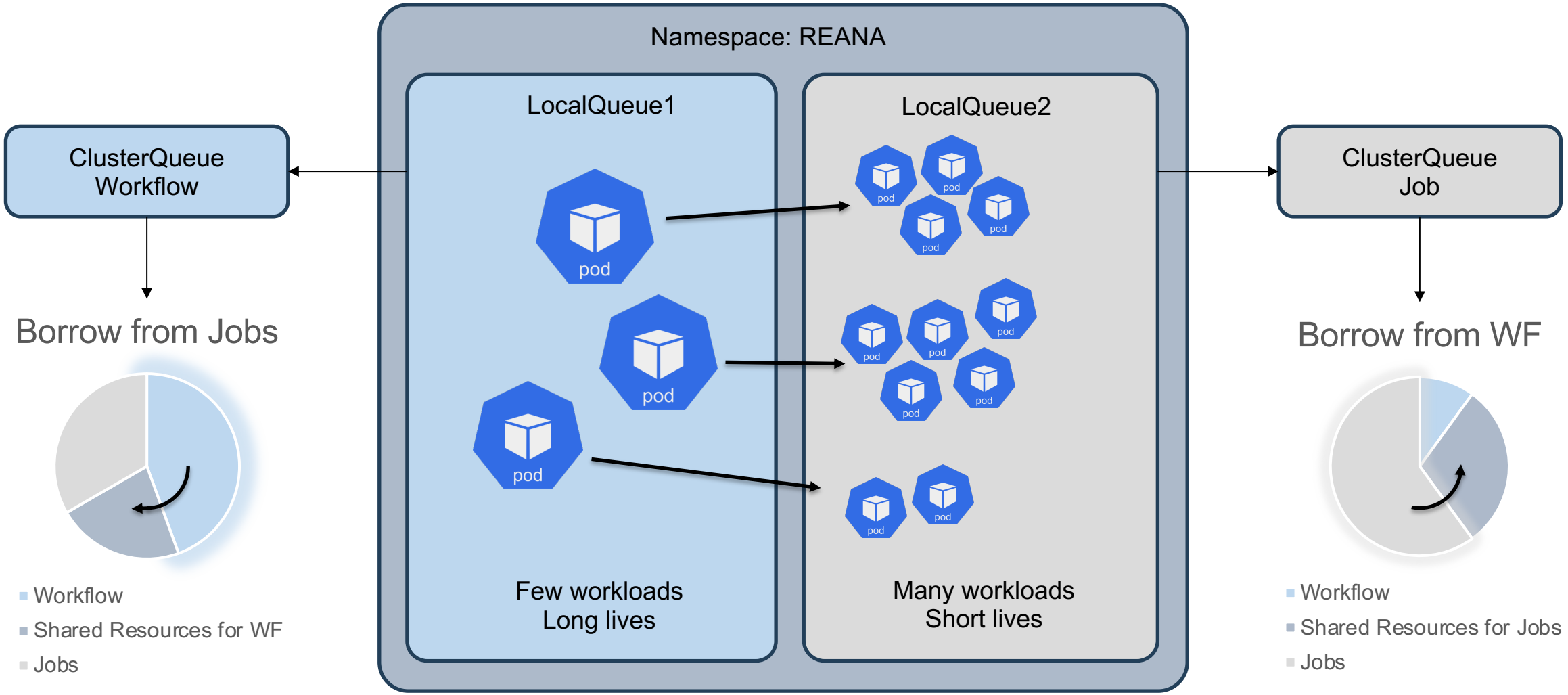
REANA job controller uses native Kubernetes Job API and necessitated custom developments for efficient and fair workflow scheduling.

New kids on the block?

Kueue: *Kubernetes-native batch scheduling system*

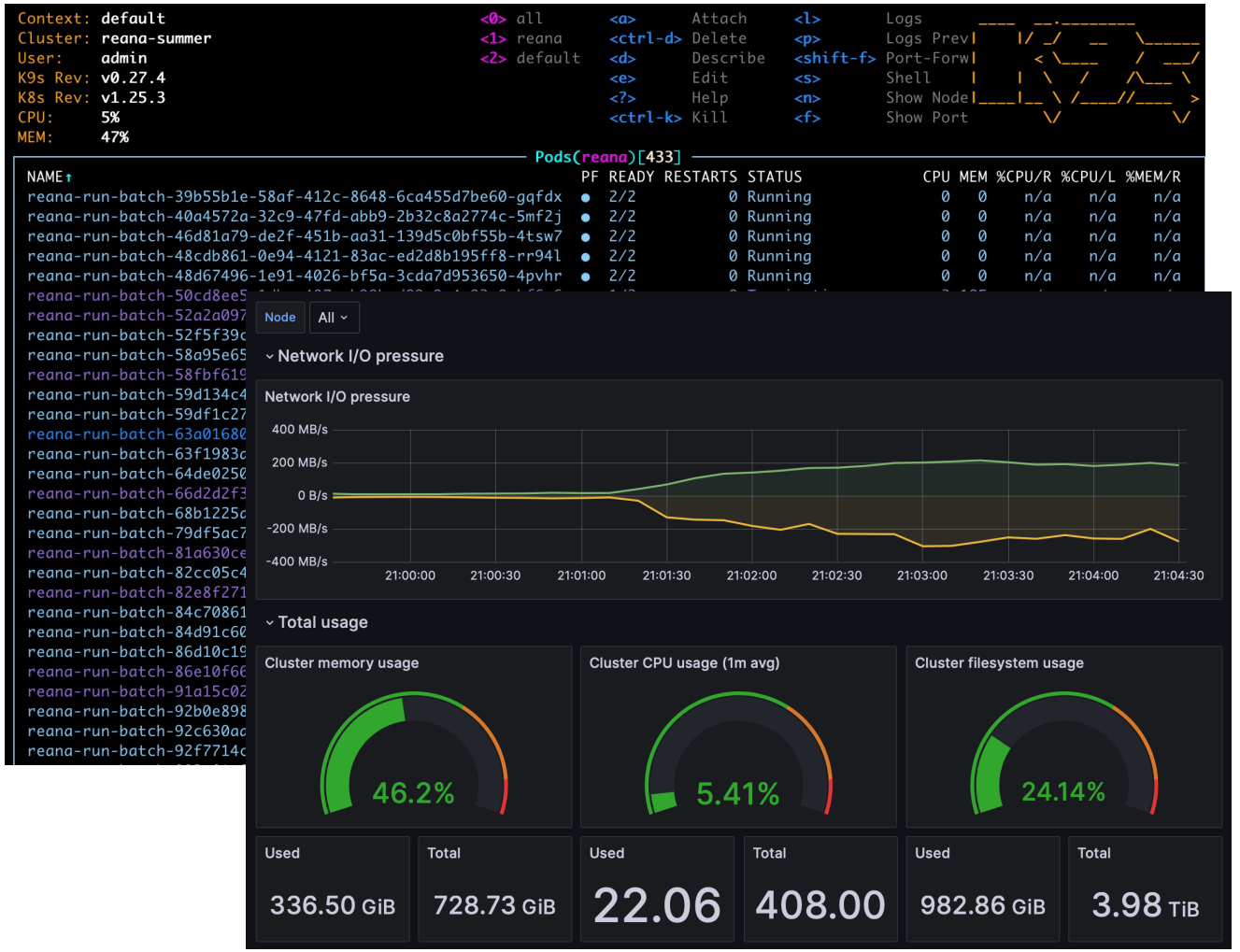


Kueue evaluation setup

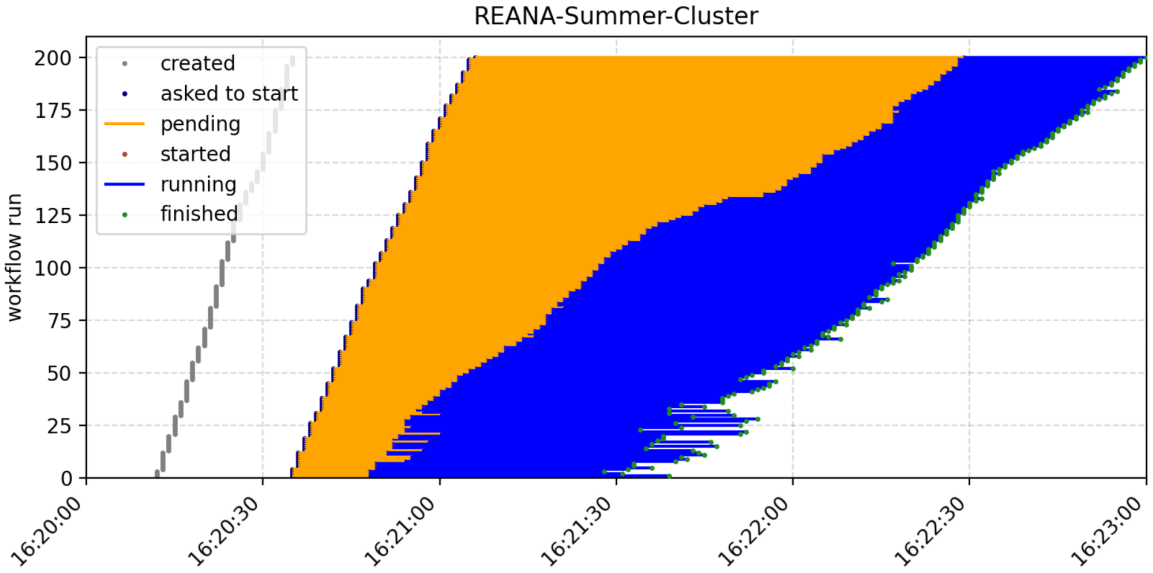


Typical evaluation experiment

```yaml
apiVersion: batch/v1
kind: Job
metadata:
  namespace: reana # Job under reana namespace
  generateName: sample-job-
  annotations:
    kueue.x-k8s.io/queue-name: lq-jobs # Point to the LocalQueue
spec:
  parallelism: 1 # This Job will have 1 replica running at the same time
  completions: 1 # This Job requires 1 completion
  suspend: true # Set to true to allow Kueue to control the Job when it starts
  template:
    spec:
      containers:
      - name: dummy-job
        image: gcr.io/k8s-staging-perf-tests/sleep:latest
        args: ["5s"] # Sleep for 5 seconds
        resources:
          requests:
            cpu: "4m"
            memory: "150M"
          limits:
            cpu: "4m"
            memory: "150M"
      restartPolicy: Never
```



First results



200 workflows running in parallel on 46 nodes

Kueue succeeded on several demo examples for thousands of submissions

Encountered a few corner cases when submitting a high number of jobs

- Kueue UUID job allocation issues
- Unresponsive when overloaded with 20k jobs

```
Error from server (InternalError): error when creating "Workflows/wf-
999.yaml": Internal error occurred: failed calling webhook
"mjob.kb.io": failed to call webhook: Post "https://kueue-webhook-
service.kueue-system.svc:443/mutate-batch-v1-job?timeout=10s": dial tcp
10.254.242.140:443: connect: connection refused
Error from server (InternalError): error when creating "Workflows/wf-
9990.yaml": Internal error occurred: failed calling webhook
"mjob.kb.io": failed to call webhook: Post "https://kueue-webhook-
service.kueue-system.svc:443/mutate-batch-v1-job?timeout=10s": dial tcp
10.254.242.140:443: connect: connection refused
Error from server (InternalError): error when creating "Workflows/wf-
```

Kueue is promising… but still maturing

Thank you
Questions?

xavier.tintin@cern.ch

https://www.linkedin.com/in/xavier-tintin/

@reanahub