HPC Scheduling with Kubernetes

Sprint 3

Siyuan Chen, Yilin Xu, Nidhi Shah, Soufiane Jounaid, Juhi Paliwal

Claudia Misale(IBM), Carlos Eduardo Arango Gutierrez (RedHat),
Daniel Milroy (Lawrence Livermore National Laboratory)

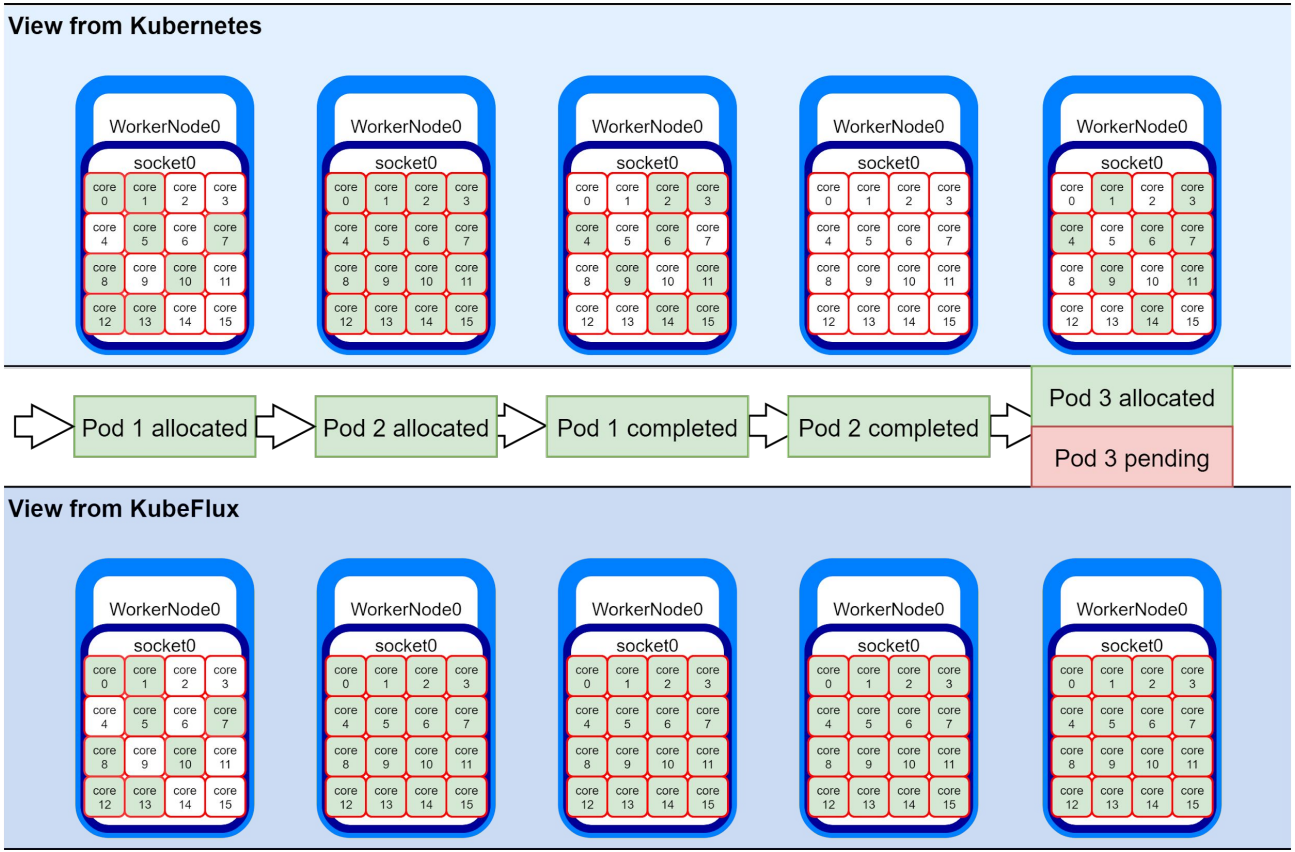
Agenda

- Recap of the project
- Introduction of Informer
- Demo 1 Pi test (Pod Succeeded)
- Demo 2 Pi test with segmentation fault (Pod Failed)
- Burndown chart
- Accomplished so far
- Plans for next sprint

About the Project so far

- State inconsistencies between the Kubernetes cluster and the Kube-Flux scheduler.
- Hinders utilization of the cluster, queued jobs wait forever.
- We want to design and implement an informer that carries updated cluster state information to the Kube-Flux resource graph.

Problem Recap - A scheduler without feedback



- Cluster has 1 worker node with 1 cpu socket. It has 16 cores
- Pod 1, 2 ,3 requires 8 cores

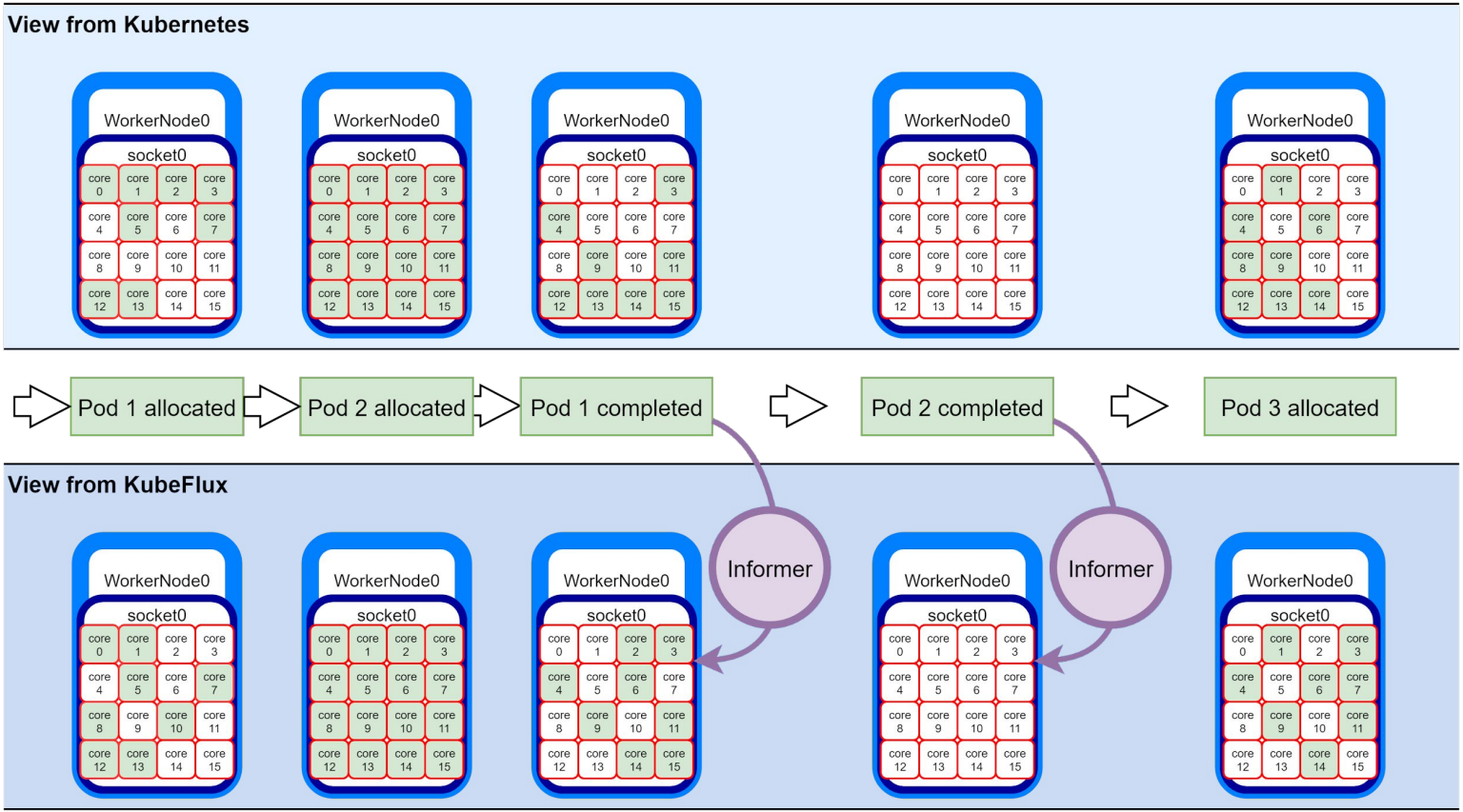
KubeFlux Log:

```
Pod cannot be scheduled by KubeFlux, noden
from a pool of 2 nodes (offset: 0, sample
Status after running PostFilter plugins fo
"Unable to schedule pod; no fit; waiting"
Could not construct reference to: '<nil>'
Updating pod condition for default/pi-job-
```

core i available

core j allocated

Solution - Informer



Cluster has 1 worker node with 1 cpu socket. It has 16 cores

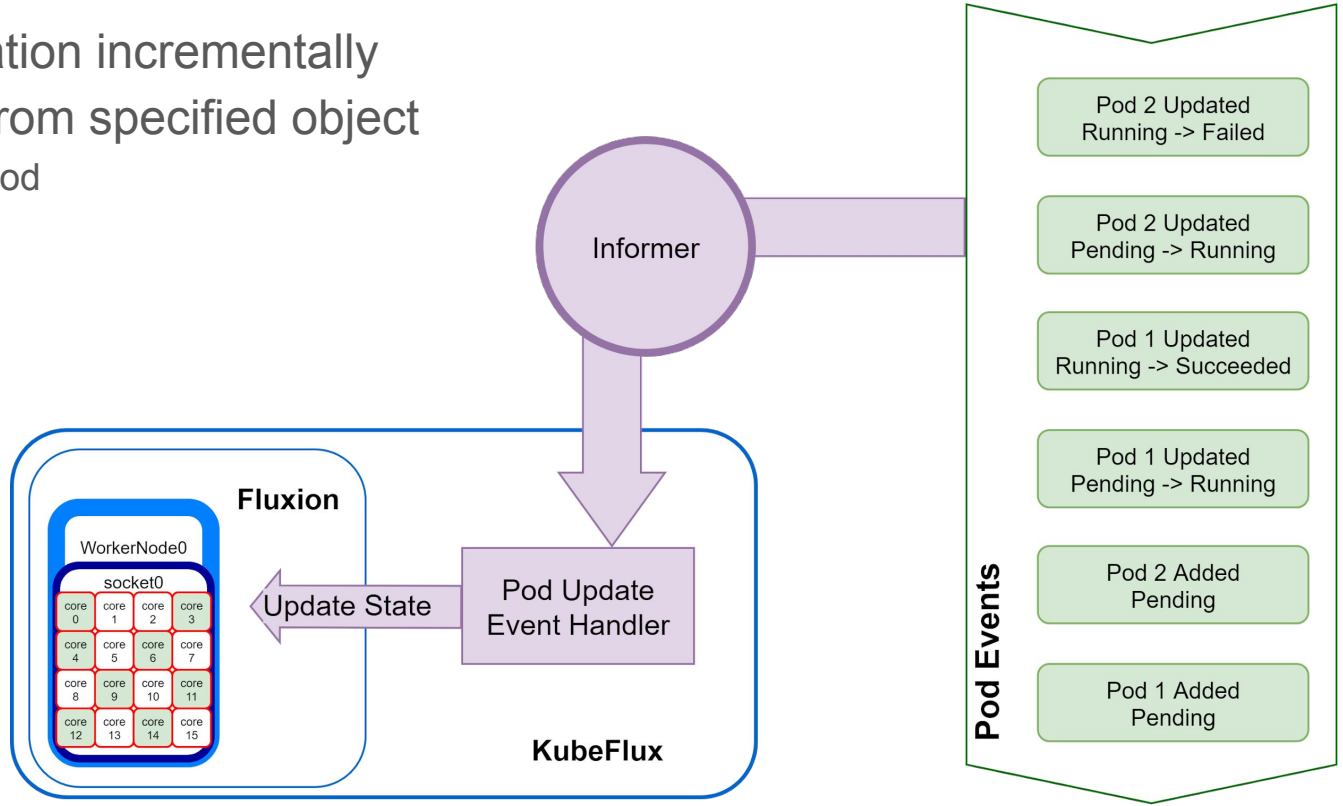
Pod 1, 2 ,3 requires 8 cores

core i — available

core j — allocated

Closer look: Kubernetes Informer

- Update information incrementally
- Watch events from specified object
  - In our case: pod



Demo Outline

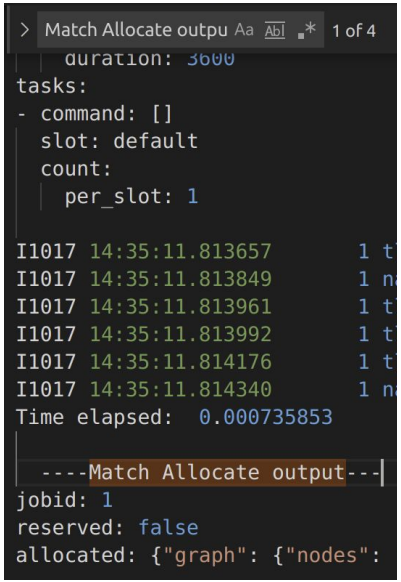
1. Demo description
2. Run Pi test without segmentation fault
3. Run Pi test with segmentation fault
4. Expected results of the experiment

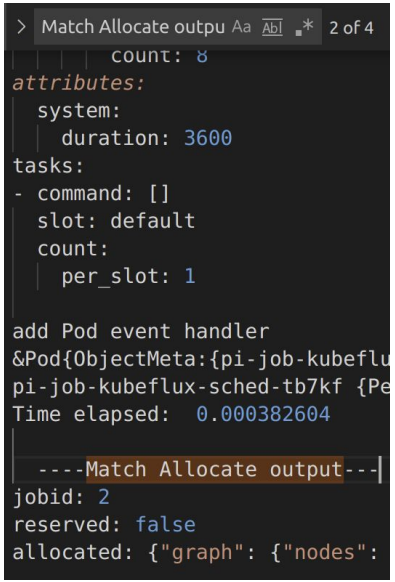
Demo Setup

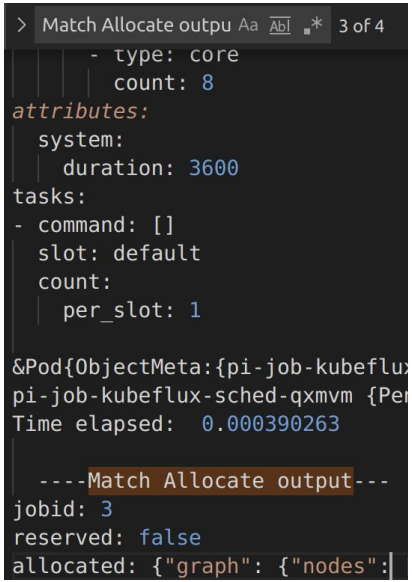
- Environment: Cluster running on local machine
  - 2 compute nodes
    - 1 Master node: Kubernetes control plane
    - 1 worker node
      - 16 virtual cores in 1 socket and 1 memory node
- Tools:
  - Kind
  - Docker
- Task:
  - Deploy 4 pods each on 8 compute cores.
- Goal
  - Demonstrate job cancellation feature in succeeded pods
  - Demonstrate informer in failed pods

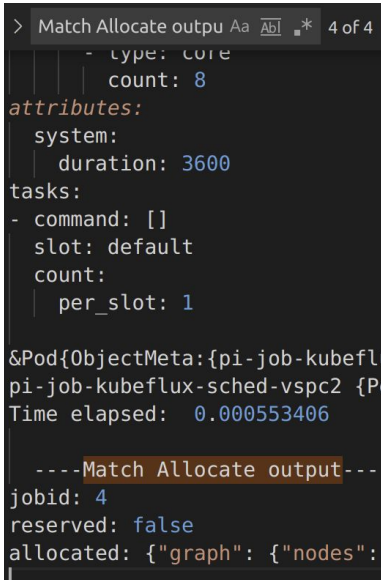
Experiment Logs

- 4 match allocate outputs found in the log

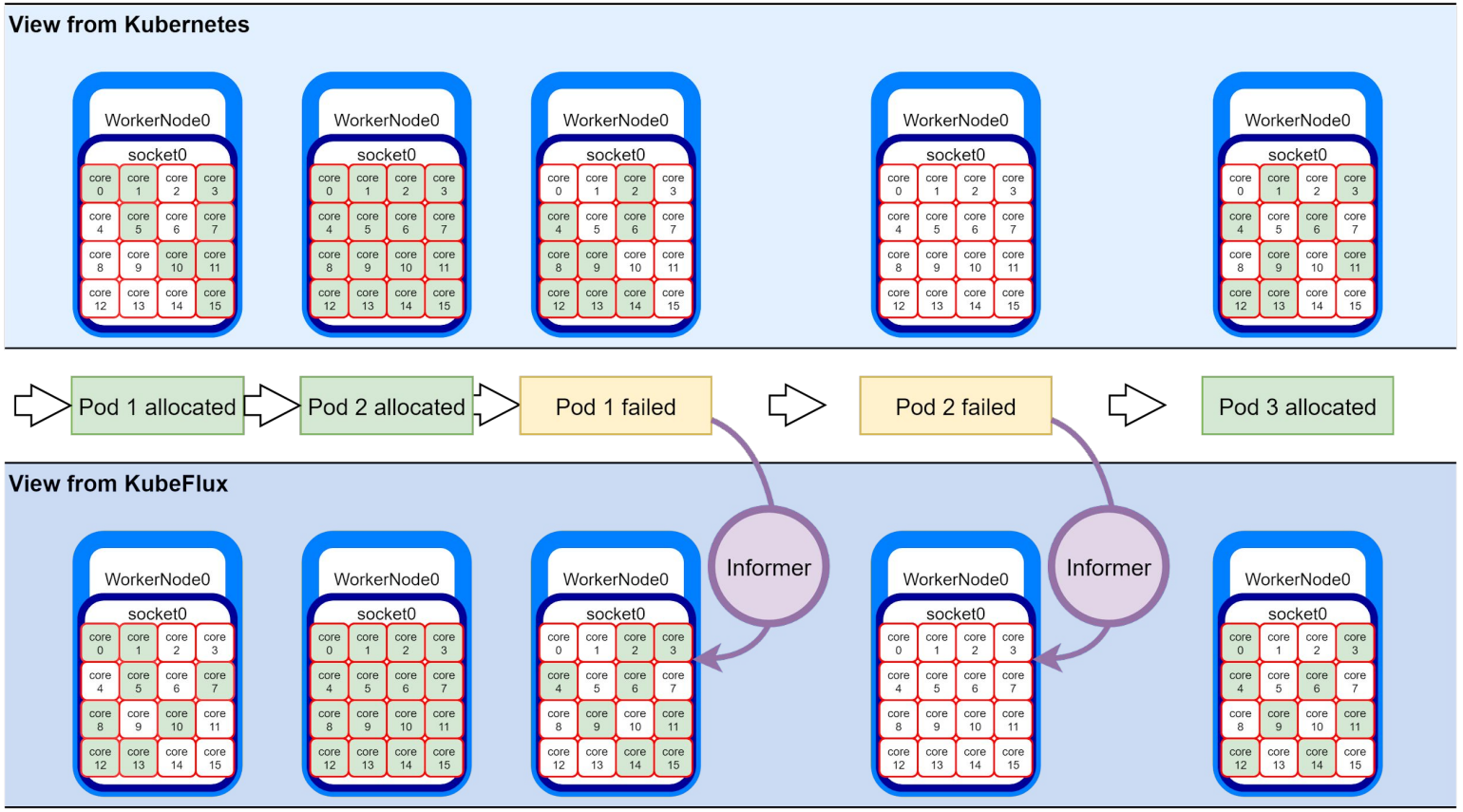








Corner Case: Failed Pod



Cluster has 1 worker node with 1 cpu socket. It has 16 cores

Pod 1, 2 ,3 requires 8 cores

core i available

core j allocated

Experiment Logs

- Failed pods' resources have been freed

```
pi-job-kubeflux-sched-segfault-mdtht {Failed [{Initialized Tru
Pod pi-job-kubeflux-sched-segfault-mdtht failed, kubeflux need
Cancel flux job: 1 for pod pi-job-kubeflux-sched-segfault-mdth
Time elapsed (Cancel Job) : 6.8945e-05
Job cancellation for pod pi-job-kubeflux-sched-segfault-mdtht
```

```
pi-job-kubeflux-sched-segfault-4cwjn {Failed [{Initialized Tru
Pod pi-job-kubeflux-sched-segfault-4cwjn failed, kubeflux need
Cancel flux job: 2 for pod pi-job-kubeflux-sched-segfault-4cwj
Time elapsed (Cancel Job) : 7.012e-05
Job cancellation for pod pi-job-kubeflux-sched-segfault-4cwjn
```

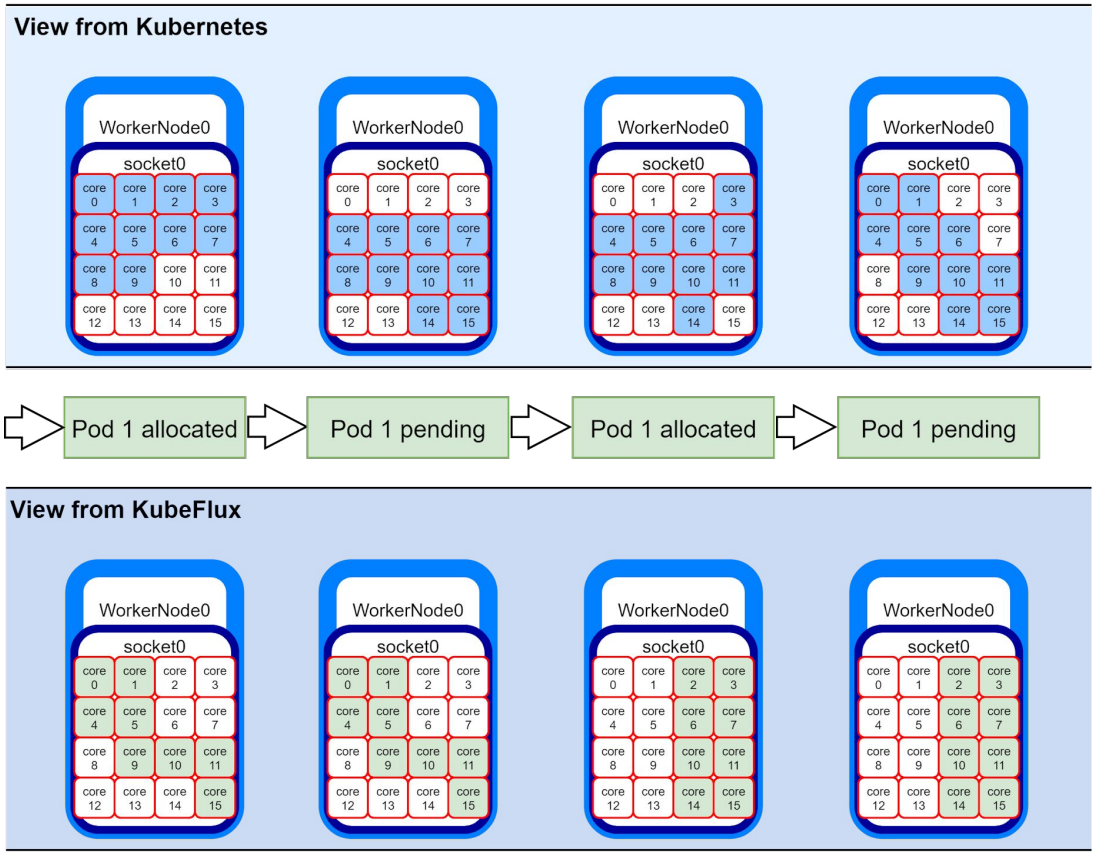
```
pi-job-kubeflux-sched-segfault-4fw72 {Failed [{Initialized Tr
Pod pi-job-kubeflux-sched-segfault-4fw72 failed, kubeflux nee
Cancel flux job: 7 for pod pi-job-kubeflux-sched-segfault-4fw
Time elapsed (Cancel Job) : 7.7859e-05
Job cancellation for pod pi-job-kubeflux-sched-segfault-4fw72
```

```
pi-job-kubeflux-sched-segfault-5psmc {Failed [{Initialized Tru
Pod pi-job-kubeflux-sched-segfault-5psmc failed, kubeflux need
Cancel flux job: 8 for pod pi-job-kubeflux-sched-segfault-5psm
Time elapsed (Cancel Job) : 6.7277e-05
Job cancellation for pod pi-job-kubeflux-sched-segfault-5psmc
```

```
→ kubeflux git:(dev-kubeflux) ✗ kubectl get pods
NAME                                    READY   STATUS
pi-job-kubeflux-sched-7kzsc             0/1     Completed
pi-job-kubeflux-sched-84pz8             0/1     Completed
pi-job-kubeflux-sched-csgsb             0/1     Completed
pi-job-kubeflux-sched-segfault-4cwjn    0/1     Error
pi-job-kubeflux-sched-segfault-4fw72    0/1     Error
pi-job-kubeflux-sched-segfault-5psmc    0/1     Error
pi-job-kubeflux-sched-segfault-mdtht    0/1     Error
pi-job-kubeflux-sched-segfault-qndq8    0/1     Error
pi-job-kubeflux-sched-wt6b4             0/1     Completed
```

- 8 pods has been scheduled, 4 failed and 4 succeeded. Note, a failed pod has been rescheduled in the above screen shot.

Next Challenge: state synchronization



Example: In a cluster with 1 worker node, a kubeflux scheduled pod could be assigned to a node of insufficient cpu.

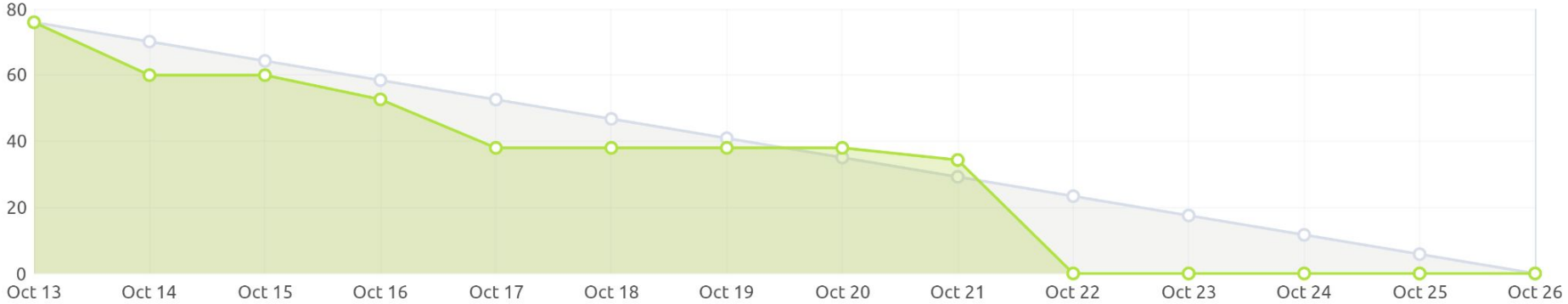
- It causes unwanted rescheduling
- Solution: implement state synchronization

Sprint 3 Tasks

- ❖ Complete the infrastructure setup
- ❖ Codewalk by our mentors
- ❖ Analysis the existing code
- ❖ Develop the informer

Challenge:

- State Synchronization



Plan for Next Sprint

- Performance comparison between both solutions experiment proposal
  - Goal: show the difference in performance between updating or rebuilding the graph.
- Approach to Task 2: state sync between Kube-flux and other schedulers.
  - Problem
  - Discussion
  - Potential solution

Thank You!