Slurm Bridge

Alan Mutschelknaus
Skyler Malinowski
Marlow Warnicke

slurm | SCHEDMD

Slinky Components - 0.3.0

- Slurm Operator
  - Helm charts (slurm, operator)
- Slurm Client
  - Go client library
- Slurm Exporter
  - Slurm metrics and grafana dashboard
- **Slurm Bridge** (new!)

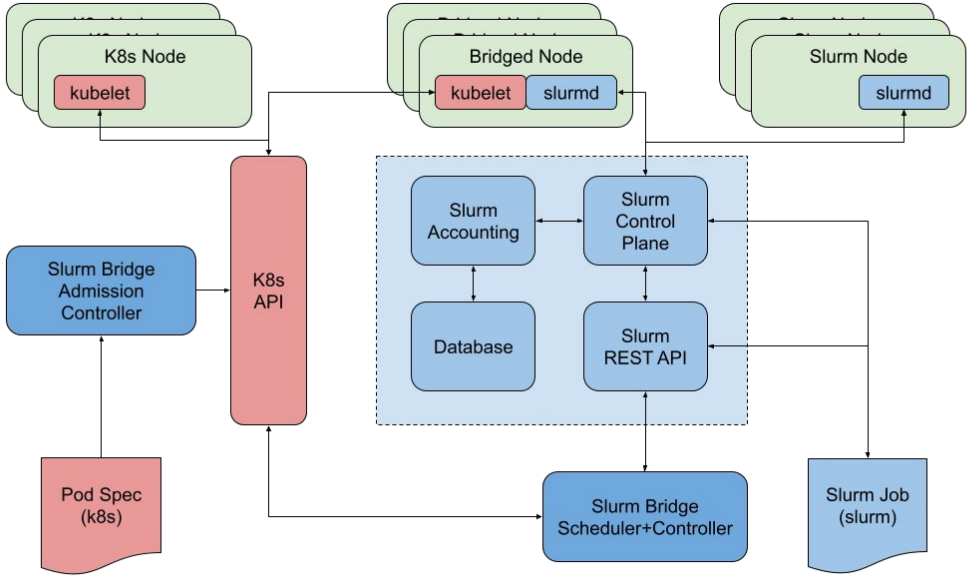
https://github.com/slinkyproject

Slurm Bridge - Requirements

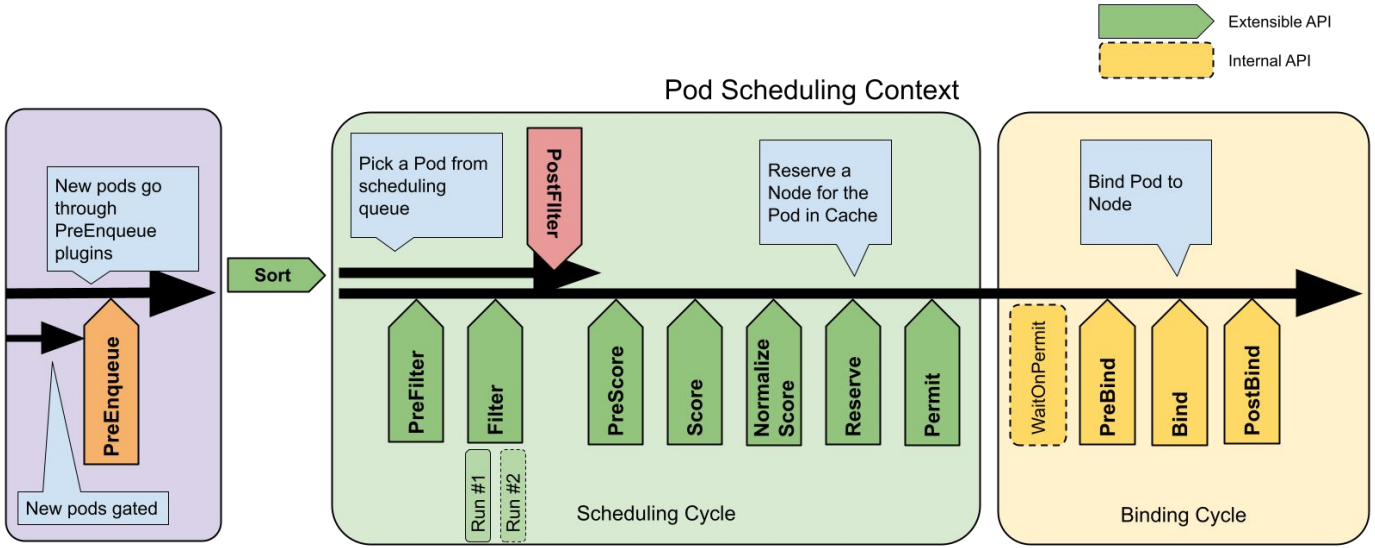
- Schedule K8s resources using Slurm's scheduler
- Scheduled by Slurm, run by K8s
- Can run Slurm and Kubernetes workloads on pools of nodes
- Translate resource requirements for Kubernetes workloads and make sure appropriate resources are available and schedule within the cluster
- Filter nodes intended for work scheduled by Slurm
- Schedule pods via namespace and/or "SchedulerName"
- Handle Device Plugins, such as GPUs

Big Picture

- Slurm-Bridge represents k8s pod(s) as a Slurm job, for scheduling purposes
- Kubernetes handles pods launch, after scheduling
- Slurm handles job scheduling
- Both Slurm and Kubernetes can still schedule other workload on non-Bridged Nodes

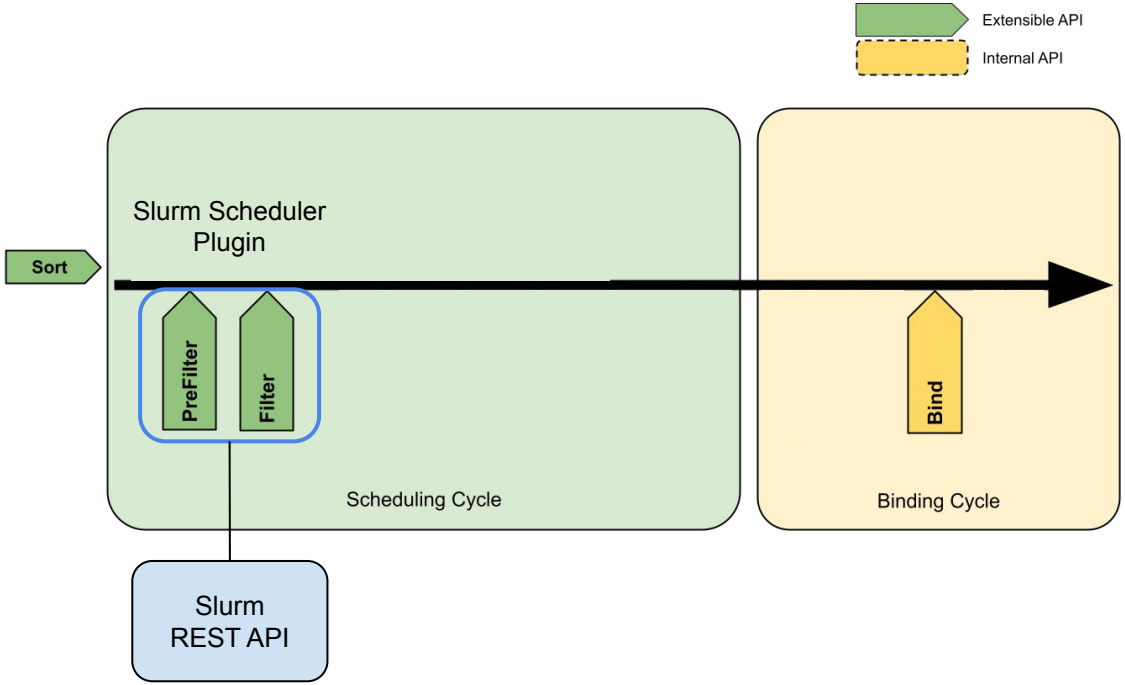


Kubernetes Scheduler Framework



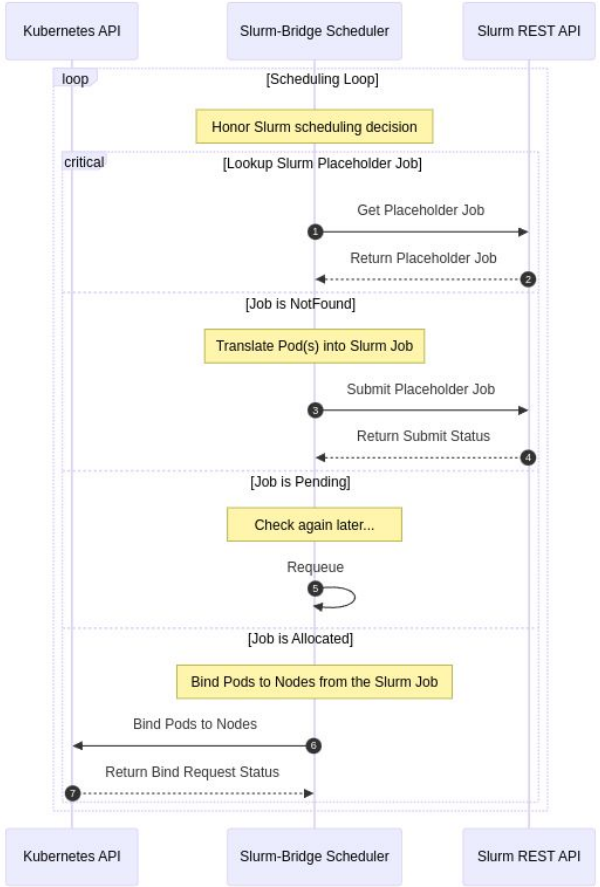
Slurm Scheduler Plugin

- Implement PreFilter and Filter Plugins
- Use default Sort and Bind Plugins
- Translate Pod spec into Slurm job spec, and submit as a placeholder Slurm job in PreFilter
- Bind pod to Kubernetes Node based on Slurm node allocation, from Slurm job
- Let kubelet handle the pod initialization and resources



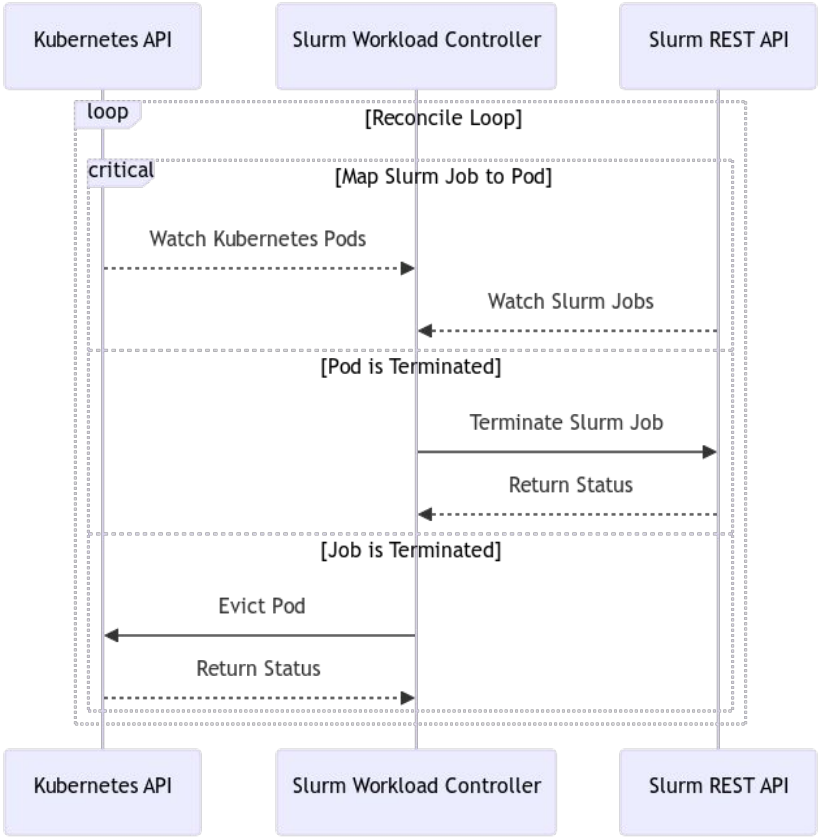
Slurm Scheduler Plugin - Sequence

- Translate a pod spec to Slurm job spec
- Submit a placeholder job in Slurm for the pod
- Wait for placeholder job to be running
- Bind the pod to a node, given where the Slurm job was allocated in Slurm



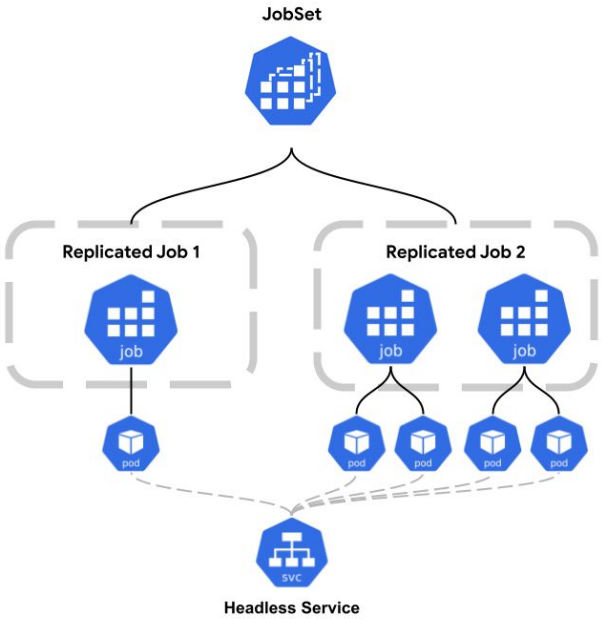
Slurm Workload Controller - Sequence

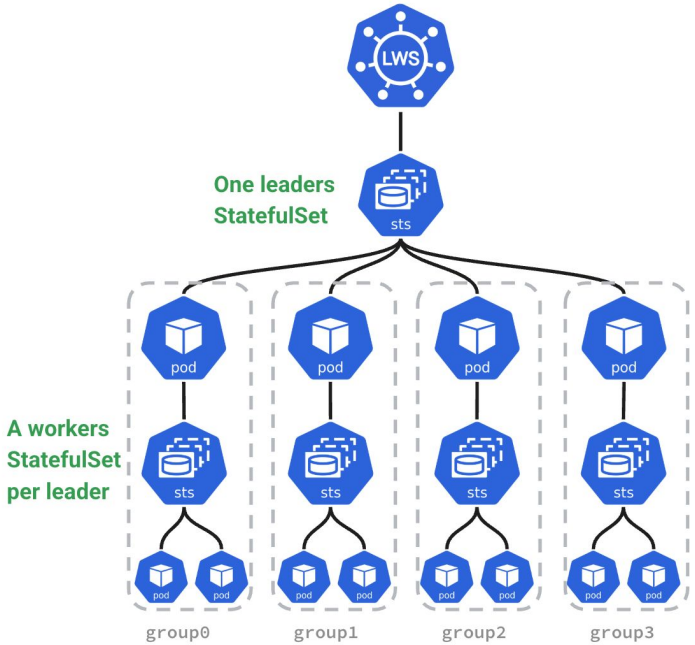
- Responsible for cleaning up:
  - Slurm jobs after pods complete/terminate
  - Pods after Slurm job complete/terminate



Demo

JobSet and LeaderWorkerSet





Future Work

Future Work

- Work with the Kubernetes community to be able to handle fine-grained control and understanding of native resources
- Be able to handle Dynamic Resource Allocation (DRA)
- Allow Slurm to schedule Kubernetes workloads without slurmd needing to run alongside kubelet
- Support LeaderWorkerSet