The Best of Both Worlds: Slurm on Kubernetes
SLUG '23
CoreWeave is a specialized cloud, built for GPU accelerated workloads on top of the industry’s fastest and most flexible infrastructure.
## One Orchestration Layer To Rule Them All

### Bare Metal Containers
- Distributed Training Workloads
- Inference Workloads
- General MicroServices
- KNative Serverless Compute
- KEDA Auto Scaling
- MPI Operator
- Argo Workflows
- Helm Application Management
- + Thousands of community supported projects

### Virtual Servers
- KVM / QEMU Hypervisor Workloads
- Supports workloads not containerizable
- **Great for Virtual Desktops and Developer Workstations**
- Available on all CPU and GPU platforms
- Linux and Windows support
- Does not support GPUDirect RDMA

### Storage
- **Scalable, Highly Available Network Attached Storage**
- All-NVMe and HDD Tiers Available
- Block Device and Shared Filesystem Volume Types (Supports Multi-Attach)
- Interoperable with Containers and Virtual Servers
- Up to 1PB Volume Sizes
- 10MM+ IOPS per Volume

### Networking
- Infiniband GPUDirect RDMA (Bare Metal)
- Private & Public IP Ranges
- Load Balancer Services (internal & external)
- Firewalls via Network Policies
- **Bare Metal Containers and Virtual Servers in same network domain**
- Direct Connects and Cloud On-Ramps Available

---

All resources are adjacent to one another and governed by the same Kubernetes concepts, making orchestration much more manageable.
Datacenter Footprint

225 MW of Power across 15 Datacenters
HGX H100 Supercomputer Instances

- NVIDIA HGX H100 Platform built on Intel Sapphire Rapids platform
- 1:1 Non-Blocking GPUDirect Fabric built rail optimized using NVIDIA InfiniBand NDR HCAs and Quantum-2 Switches
- Standard 8 rail configuration
- Topology supports NVIDIA SHARP in-network collections

For more information: CoreWeave H100 Documentation
Partners & Customers

- Investor and strategic partner
- Executive level collaboration and planning
- Elite North American partner for cloud services

- Top AI companies and labs
- Scale scientific and drug discovery
- Operations and platform providers
- Expanding compliance in 2024 for government related contracts
Why Slurm On Kubernetes?

Seamless Experience

Support for both **burst** and **batch** workloads on the same central platform.

- CoreWeave’s core API and orchestration is Kubernetes
- Separating orchestration entirely means two separate pools of compute to maintain and operate
- Without Slurm, customers lose out on industry best solutions for HPC
# SUNK Features

<table>
<thead>
<tr>
<th>Configuration &amp; Deployment</th>
<th>Kubernetes Integration</th>
<th>State Management</th>
</tr>
</thead>
</table>
| - Deployable in suite of **helm chart** related tools including popular gitops workflows | - Inherits the **core Kubernetes features**  
  ○ HA of control plane services  
  ○ Dynamic node scaling  
  ○ Resource management with request and limits  
  ○ Shared filesystem via PersistentVolumeClaim resources | - **Dynamic nodes** with two way syncing of state between k8s and slurm |
| - **Easy tracking & configuration** of prolog and epilog scripts | - **K8s scheduler** for scheduling native kubernetes workloads via Slurm scheduler  
  ○ Serverless/Bursty workloads | - **Automatic topology** generation |
| - **Support for s6** scripts and services | | - **Support for Pyxis** container execution |
| - Configurable authentication schemes including Ldap through companion OpenLdap helm chart or third party solution (Authentik, GAuth, etc.) | | - **GRES support** and auto-identification |
SUNK Implementation Overview

Services *containerized in Kubernetes*
SUNK Implementation Overview

Services **containerized in Kubernetes**

Slurm components as **Pods**

- Controller
- Accounting
- Rest API
- Slurm DB
- Login Nodes
SUNK Implementation Overview

Services **containerized in Kubernetes**

Slurm components as **Pods**

Configuration as **ConfigMaps** and **Secrets**

- Single source of truth mounted across all components
- Slurm config
- Topology
- Prolog/Epilog Scripts
SUNK Implementation Overview

Services **containerized in Kubernetes**

Slurm components as **Pods**

Configuration as **ConfigMaps** and **Secrets**

**Nodesets** maintaining compute
- A CRD for scheduling slurmd containers on Kubernetes nodes
- Tracks status of the compute nodes within slurm
SUNK Implementation Overview

Services **containerized in Kubernetes**

Slurm components as **Pods**

Configuration as **ConfigMaps** and **Secrets**

**Nodesets** maintaining compute

**Slurm Syncer** reconciling state

- Sync the state of Kubernetes nodes into slurm and vice-versa
- Interact with Slurm’s REST API
SUNK Implementation Overview

Services **containerized in Kubernetes**

Slurm components as **Pods**

Configuration as **ConfigMaps** and **Secrets**

**Nodesets** maintaining compute

**Slurm Syncer** reconciling state

Staying consistent with the **Operators**

- Reconciles node changes from the Slurm and Kubernetes sides
SUNK Implementation Overview

Services **containerized in Kubernetes**

Slurm components as **Pods**

Configuration as **ConfigMaps and Secrets**

**Nodesets** maintaining compute

**Slurm Syncer** reconciling state

Staying consistent with the **Operators**

Schedule from both sides

- Allows scheduling of Kubernetes workloads on shared pool of compute with Slurm
SUNK Implementation Overview

Services **containerized in Kubernetes**

Slurm components as **Pods**

Configuration as **ConfigMaps** and **Secrets**

**Nodesets** maintaining compute

**Slurm Syncer** reconciling state

Staying consistent with the **Operators**

**Schedule** from both sides

Expose prometheus **Metrics**
Nodesets

- A cross between daemonset and statefulset
- Kubernetes Status fields based on state in Slurm, node status, and pod affinity
- Schedules pods running slurmd taking into account:
  - Node affinity
  - Tenancy of other workloads
  - Kubernetes and Slurm node conditions
- Protected rolling updates and scale up/down considering Drain/Active conditions

<table>
<thead>
<tr>
<th>NAME</th>
<th>DESIRED</th>
<th>FEASIBLE</th>
<th>CURRENT</th>
<th>READY</th>
<th>UP-TO-DATE</th>
<th>RUNNING</th>
<th>DRAIN</th>
</tr>
</thead>
<tbody>
<tr>
<td>slurm-a40</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>slurm-h100</td>
<td>32</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>0</td>
</tr>
</tbody>
</table>
Leveraging **Slurm’s REST API**, the syncer keeps the **state of compute consistent** between kubernetes and slurm.

Reconciles pod **annotations, labels, and conditions** to match drain and running states in Slurm.

Pod conditions such as **drained/cordoned** nodes originating from **K8s** get pushed to slurm.

Efficiently caches the state from slurm to avoid overwhelming requests.
Scheduler

Controller acting as a **kubernetes scheduler**

Allows compute resources associated with a Slurm Cluster to be **allocated to k8s workloads dynamically**

Pods schedule according to node assignments matching Slurm placeholder jobs scheduled from within Slurm

Dynamically switches compute between training and inference workloads
SUNK in Practice

Declarative management with Helm

Compatible with GitOps pattern
Ex: syncing with ArgoCD

Slurm Prometheus Metrics exporter & Grafana Dashboards

Phalanx HPC verification workloads to assure health of cluster resources
Inflection AI’s ultra-performant cluster smashes MLPerf record

In a joint submission with NVIDIA, CoreWeave delivered record-breaking performance results on the MLPerf™ benchmark in June 2023.

- **<11 minutes** to train GPT-3 training benchmark on a commercially available cluster
- **+3,500** NVIDIA Tensor Core H100 GPUs used
- **29x faster** than the next best competitor
- **4x bigger** than the next best competitor
Questions?

Slurm On Kubernetes to be open sourced in 2024

Looking for future partners & maintainers, reach out to sunk@coreweave.com if interested

Want to try it out on CoreWeave? Reach out to sales@coreweave.com

We’re Hiring!
https://www.coreweave.com/careers

Coming Soon
https://github.com/coreweave/sunk

Key Contributors
Peter Salanki
Andrew Senetar

VP of Engineering
Senior Infrastructure Engineer

Questions?

Slurm On Kubernetes to be open sourced in 2024

Looking for future partners & maintainers, reach out to sunk@coreweave.com if interested

Want to try it out on CoreWeave? Reach out to sales@coreweave.com

Coming Soon
https://github.com/coreweave/sunk

Questions?

Slurm On Kubernetes to be open sourced in 2024

Looking for future partners & maintainers, reach out to sunk@coreweave.com if interested

Want to try it out on CoreWeave? Reach out to sales@coreweave.com

We’re Hiring!
https://www.coreweave.com/careers

Coming Soon
https://github.com/coreweave/sunk

Questions?

Slurm On Kubernetes to be open sourced in 2024

Looking for future partners & maintainers, reach out to sunk@coreweave.com if interested

Want to try it out on CoreWeave? Reach out to sales@coreweave.com

We’re Hiring!
https://www.coreweave.com/careers

Coming Soon
https://github.com/coreweave/sunk

Questions?

Slurm On Kubernetes to be open sourced in 2024

Looking for future partners & maintainers, reach out to sunk@coreweave.com if interested

Want to try it out on CoreWeave? Reach out to sales@coreweave.com

Key Contributors
Peter Salanki
Andrew Senetar

VP of Engineering
Senior Infrastructure Engineer

Questions?

Slurm On Kubernetes to be open sourced in 2024

Looking for future partners & maintainers, reach out to sunk@coreweave.com if interested

Want to try it out on CoreWeave? Reach out to sales@coreweave.com

We’re Hiring!
https://www.coreweave.com/careers

Coming Soon
https://github.com/coreweave/sunk

Key Contributors
Peter Salanki
Andrew Senetar

VP of Engineering
Senior Infrastructure Engineer

Questions?

Slurm On Kubernetes to be open sourced in 2024

Looking for future partners & maintainers, reach out to sunk@coreweave.com if interested

Want to try it out on CoreWeave? Reach out to sales@coreweave.com

We’re Hiring!
https://www.coreweave.com/careers

Coming Soon
https://github.com/coreweave/sunk

Questions?

Slurm On Kubernetes to be open sourced in 2024

Looking for future partners & maintainers, reach out to sunk@coreweave.com if interested

Want to try it out on CoreWeave? Reach out to sales@coreweave.com

We’re Hiring!
https://www.coreweave.com/careers

Coming Soon
https://github.com/coreweave/sunk

Key Contributors
Peter Salanki
Andrew Senetar

VP of Engineering
Senior Infrastructure Engineer

Questions?

Slurm On Kubernetes to be open sourced in 2024

Looking for future partners & maintainers, reach out to sunk@coreweave.com if interested

Want to try it out on CoreWeave? Reach out to sales@coreweave.com

We’re Hiring!
https://www.coreweave.com/careers

Coming Soon
https://github.com/coreweave/sunk

Questions?

Slurm On Kubernetes to be open sourced in 2024

Looking for future partners & maintainers, reach out to sunk@coreweave.com if interested

Want to try it out on CoreWeave? Reach out to sales@coreweave.com

We’re Hiring!
https://www.coreweave.com/careers

Coming Soon
https://github.com/coreweave/sunk

Questions?

Slurm On Kubernetes to be open sourced in 2024

Looking for future partners & maintainers, reach out to sunk@coreweave.com if interested

Want to try it out on CoreWeave? Reach out to sales@coreweave.com

We’re Hiring!
https://www.coreweave.com/careers

Coming Soon
https://github.com/coreweave/sunk

Questions?

Slurm On Kubernetes to be open sourced in 2024

Looking for future partners & maintainers, reach out to sunk@coreweave.com if interested

Want to try it out on CoreWeave? Reach out to sales@coreweave.com

We’re Hiring!
https://www.coreweave.com/careers

Coming Soon
https://github.com/coreweave/sunk

Questions?

Slurm On Kubernetes to be open sourced in 2024

Looking for future partners & maintainers, reach out to sunk@coreweave.com if interested

Want to try it out on CoreWeave? Reach out to sales@coreweave.com

We’re Hiring!
https://www.coreweave.com/careers

Coming Soon
https://github.com/coreweave/sunk

Questions?
Thank You