K2S-Gate: A Secure Kubernetes-to-Slurm Gateway with Unified Service Catalogs for Web-based AI and HPC Applications

Gukhua Lee(ghlee@kisti.re.kr), Joon Woo and Taeyoung Hong

Korea Institute of Science and Technology Information (KISTI), Daejeon, Republic of Korea

1 Introduction

Problem Statement

- No **unified gateway** for **web applications** across cloud-native and HPC
- Traditional HPC: limited support for cloud-like interactive applications
- No secure, consistent access path spanning both environments

Goal

- Develop a **Kubernetes-to-Slurm Gateway (K2S-Gate)** ⇢ bridge Kubernetes and Slurm
- Provide Unified Service Catalog ⇢ **one-click** deploy of AI & HPC applications
- Establish a **Secure Access Layer** ⇢ safe, consistent web access to services

2 Related Works

Objective	Slinky	SUNK	OOD	K2S-Gate
1. Kubernetes-Slurm gateway	◐	◐	△	●
2. Unified service catalogs (AI and HPC applications)	✕	✕	◐	●
3. Secure access layer (spanning both environments)	✕	✕	◐	●

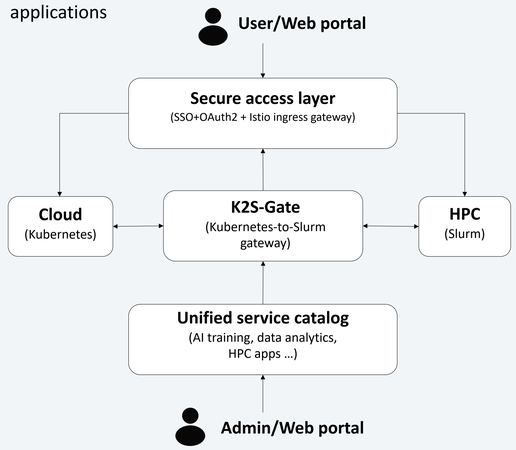
● **Full support** ◐ **Partial support** △ **Limited or indirect support** ✕ **Not addressed**

- **Slinky** ◐**1** ⇢ Runs Slurm as a Kubernetes operator (infra-level convergence), not a bidirectional gateway
- **SUNK** ◐**1** ⇢ Co-schedules Slurm & Kubernetes workloads on shared node pools, but scoped to "Slurm on Kubernetes" and proprietary
- **OOD (Open OnDemand)** △**1** ⇢ Treats Kubernetes as just one of several job-submission back ends; doesn't bridge Slurm and Kubernetes to each other ◐**2** ⇢ Offers an interactive web-app catalog (Jupyter, RStudio, VNC), but scoped to HPC only-no cloud-native app integration ◐**3** ⇢ Provides secure web access via per-user NGINX + authentication, but limited to the HPC portal not spanning both environment

3 Architecture

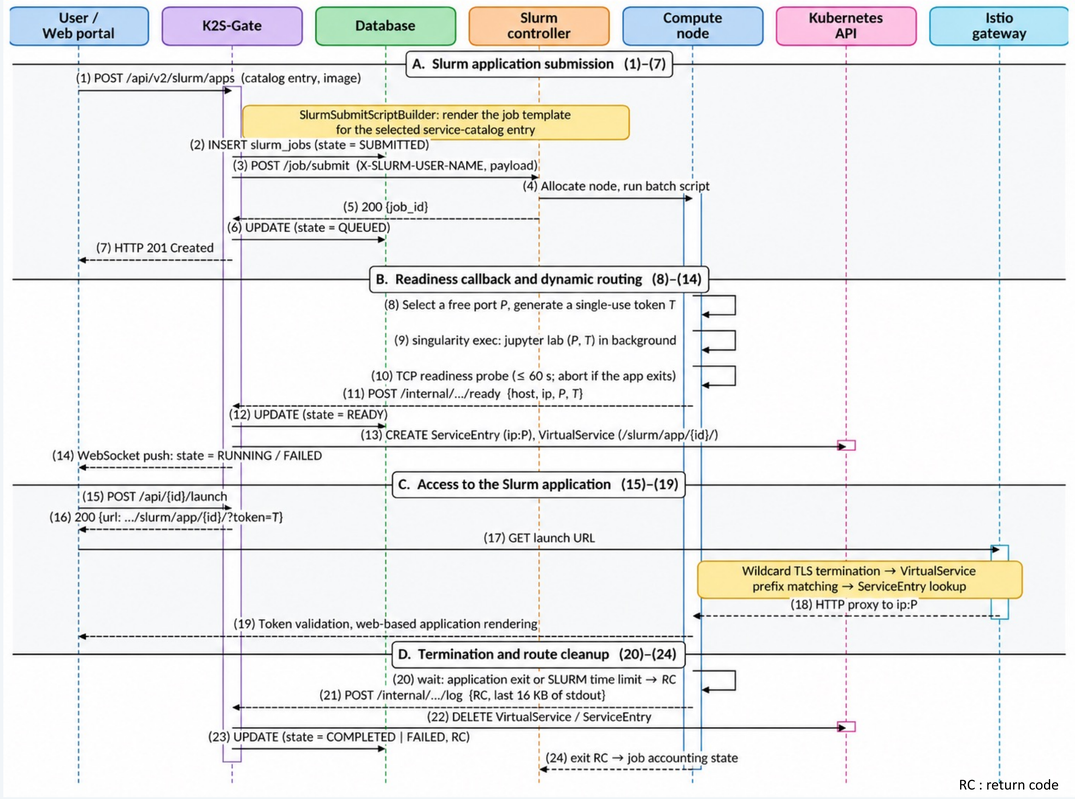
Key Components

- **K2S-Gate** provides secure, on-demand deployment of and access to web-based applications by coupling Slurm job execution with Kubernetes-native routing
- **Unified Service Catalog is** a set of predefined, administrator-authored templates (Deployment, Service, and VirtualService YAML manifests) that enables the deployment of AI and HPC applications across heterogeneous environments
- **Secure Access Layer** combines single sign-on (SSO) through an OAuth2/OIDC-based Identity Provider(IdP) with an Istio ingress gateway, providing secure and unified access to web-based applications

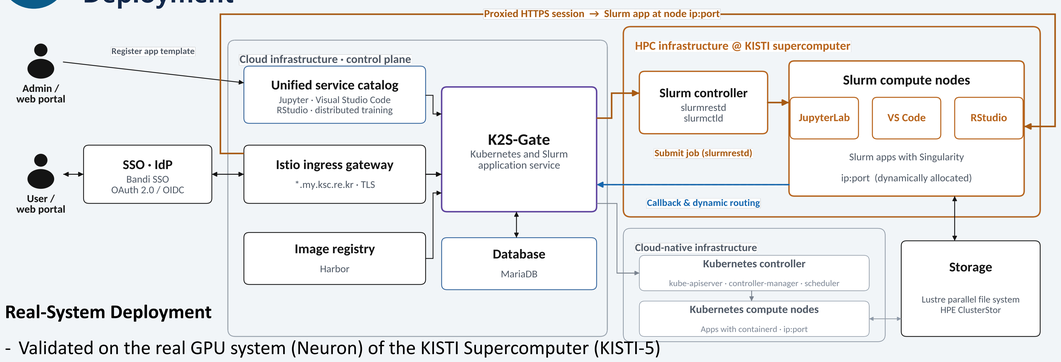


4 Method

Sequence Diagram of K2S-Gate



5 Deployment



Real-System Deployment

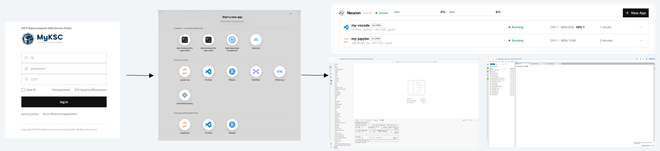
- Validated on the real GPU system (Neuron) of the KISTI Supercomputer (KISTI-5)
- HPC Infrastructure: 78 Nodes with 338 GPUs (8.9 PF)

6 Contribution

What's New

- A job-to-service transformation gateway that exposes HPC batch workloads as dynamically accessible services
- A cross-platform mechanism for bridging Kubernetes and Slurm environments
- A secure, gateway-driven access architecture for web-based applications

Implementation



7 Conclusion and Future Work

Conclusion

- K2S-Gate unifies cloud–HPC execution through a job-to-service transformation, validated in production on KISTI-5 via MyKSC (KISTI supercomputing web portal).

Future Work

- Planned deployment on the KISTI-6 supercomputer (Hangang) with extended platform capabilities

  Hangang ⇢ GPU Partition: 8,496 GPUs (GH200)

  CPU Partition: 1,600 CPUs (AMD)



[Acknowledgement]

This research was supported by Korea Institute of Science and Technology Information(KISTI) (No. (KISTI) K26L1M2C2).