HPS

jonathan.decker@uni-goettingen.de

Jonathan Decker, Sören Metje, Prof. Dr. Julian Kunkel

Running Kubernetes Workloads on Rootless HPC Systems using Slurm

IARIA

Table of contents

Presenter Resume

- PhD student with focus on HPC
- Supervised by Prof. Dr. Julian Kunkel
- Interested in combining Kubernetes and HPC
- Working on free LLM services
  for research and education in Germany
  - See `https://kisski.de`

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN
IN PUBLICA COMMODA SEIT 1737
GWDG
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen

Kubernetes on HPC Systems

- Kubernetes is well established for Cloud Computing
- High-Performance Computing (HPC) systems offer compute power
  - Typically run batch schedulers such as Slurm
- How to run Kubernetes workloads on HPC?
  - Cannot replace Slurm
  - Cannot break existing security model, e.g., no root access
  - Should run workloads across multiple nodes
  - Should support all Kubernetes features
  - Should not impose performance overhead
  - Should be easy to operate and well maintained

Running Kubernetes Workloads on HPC

- Various approaches exist, including
  - WLM-Operator, Singularity as K8s CRI
  - Bridge Operator, submit Slurm jobs in K8s
  - High-Performance Kubernetes (HPK), K8s on Apptainer
  - Kube-Slurm, Slurm controls K8s on same nodes
  - Slinky, Slurm operator in K8s cluster
  - Kind-Slurm-Integration (KSI), K8s on rootless Podman (*our approach*)
- Need categorization for comparison
  - Wickberg of SchedMD defines 4 models:
    Under, Distant, Adjacent, Over
  - From perspective of Slurm

Integration Models

- **Under**
  - Slurm cluster runs within K8s cluster via one or more pods
  - Slinky
- **Distant**
  - Nodes are either part of K8s or Slurm cluster, nodes may be moved
  - No open source
- **Adjacent**
  - K8s and Slurm cooperate via some tool but can be used individually
  - WLM-Operator, Bridge Operator, HPK
- **Over**
  - Entire K8s environment in Slurm job, removed upon job completion
  - KSI
- Focus on Adjacent and Over

Viable Approaches (for Our Use Case)

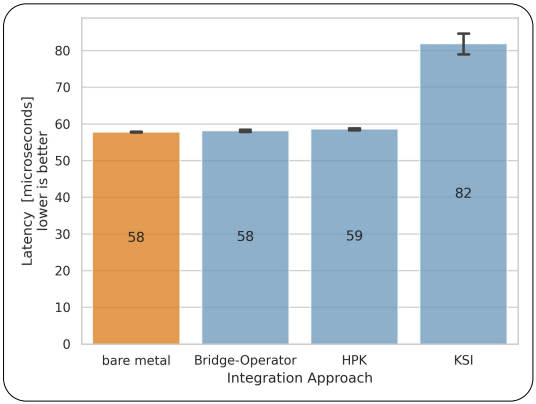
- Bridge Operator via Adjacent
  - K8s control plane outside of Slurm job
  - BridgeJob CRD, converted and send to Slurm API
  - Supports Kubeflow
  - No actual containers under Slurm
- HPK via Adjacent
  - K8s control plane in single Apptainer container
  - Virtual Kubelet to represent Slurm cluster as single node
  - Each pod submitted as Apptainer Slurm job
  - Advanced network features not supported, e.g., services
- KSI via Over
  - No external components, only rootless dependencies
  - Utilizes Kind via rootless Podman
  - Does not support multi-node, could be enabled via Kilo or Liqo
- WLM-Operator was defunct, project archived end of 2020

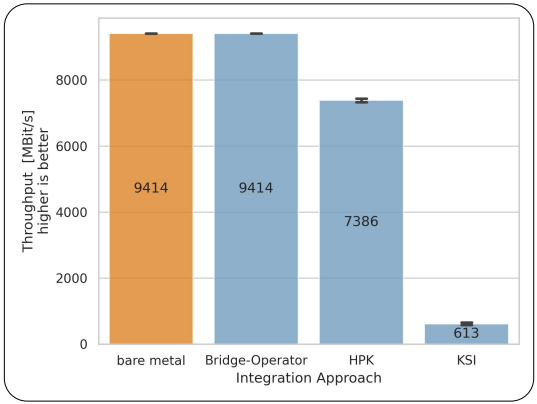
Benchmarking Approach

- Factors
  - Startup time
  - CPU compute performance
  - Memory throughput
  - Storage throughput
  - Network latency
  - Network bandwidth
- Deployment
  - 2 CentOS Stream 9 nodes
  - 2 CPUs each
  - SSD storage
  - 10Gb Ethernet
- Baseline is bare metal, i.e., only Slurm

Performance

- Significant differences in startup time and network





- KSI performance significantly slower
- Caused by rootless Podman networking via slirp4netns

Evaluation Overview

- Bridge Operator
  - Accepts Slurm jobs via a CRD in Kubernetes
  - Effectively bare metal performance
  - Very limited Kubernetes features available
- HPK
  - Runs containers via Apptainer
  - Close to bare metal performance
  - Advanced features not supported, e.g., services, kubectl exec
- KSI
  - Runs K8s via Kind on rootless Podman
  - Significantly reduced network performance
  - All K8s features supported, except for multi-node

Open Problems

- Trade off: Performance vs Features
  - Some overhead is expected
  - Cannot have all features without performance hit
- Only prototype implementations available
  - Unmaintained or incomplete
  - No standard solution available
- Investigating other approaches
  - k3d and Usernetes

Conclusion

- Contributions
  - Design and implementation of KSI
  - Comparison of solutions for rootless Kubernetes under Slurm
- Takeaways
  - Current solutions trade features for performance
  - No definitive solution yet, only prototypes
- Contact
  - Jonathan Decker jonathan.decker@uni-goettingen.de
  - Georg-August-Universität Göttingen, Germany https://uni-goettingen.de/