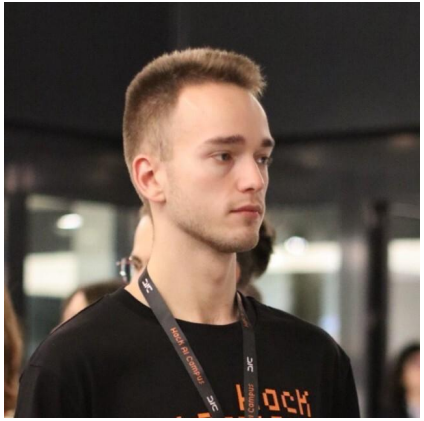
P4-Enabled Container Migration in Kubernetes

Mentors: Radostin Stoyanov, Davide Scano
Contributor: Stanislav Kosorin

Google
Summer of Code

Stanislav Kosorin

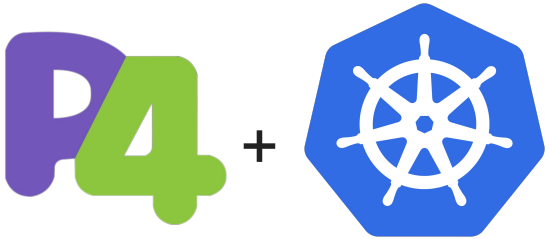
- CS Master's Student at the Technical University of Berlin
- Part-Time Software Engineer at Cresta
- Interested in Distributed Systems and Networking
- Github Profile: https://github.com/stano45
- Contact: kosorin.com/contact



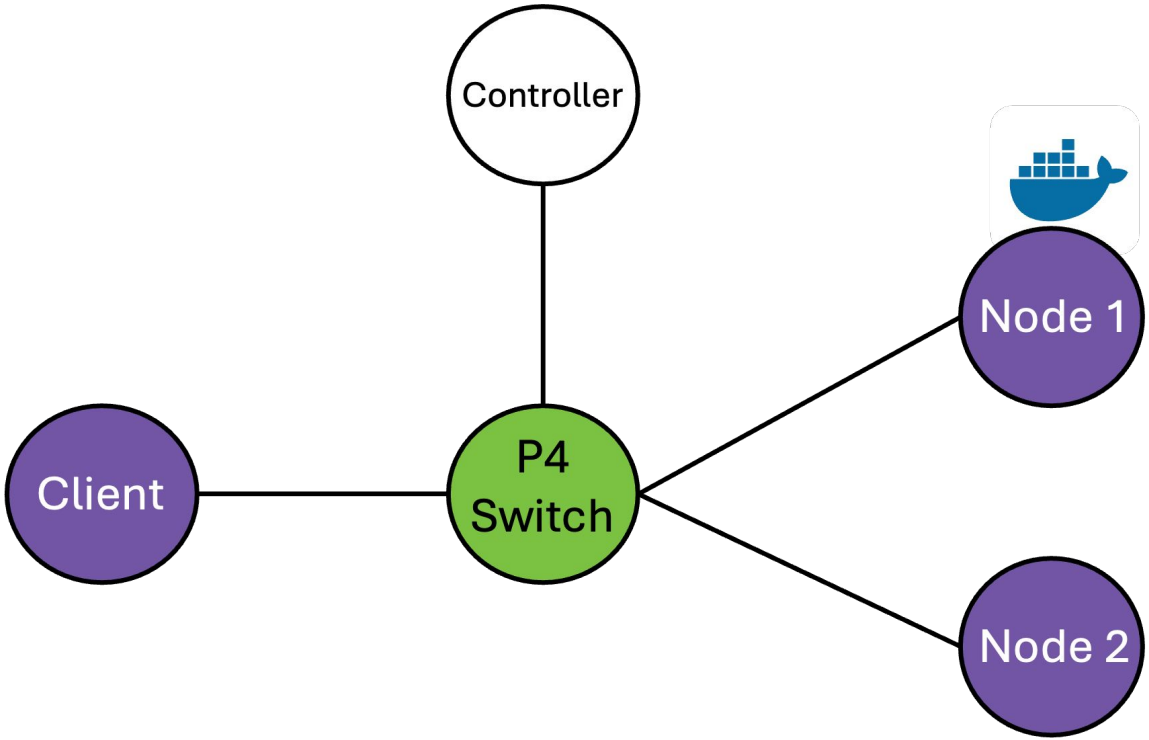
P4-Enabled Container Migration in Kubernetes

- **Motivation:** Kubernetes supports container checkpointing
- Create container checkpoint in one pod, and restore it in another pod
- **Problem:** Container IP might change during migration, networking adjustments are not automatically handled
- **Idea:** Use a P4-based switch, insert match-action table entries

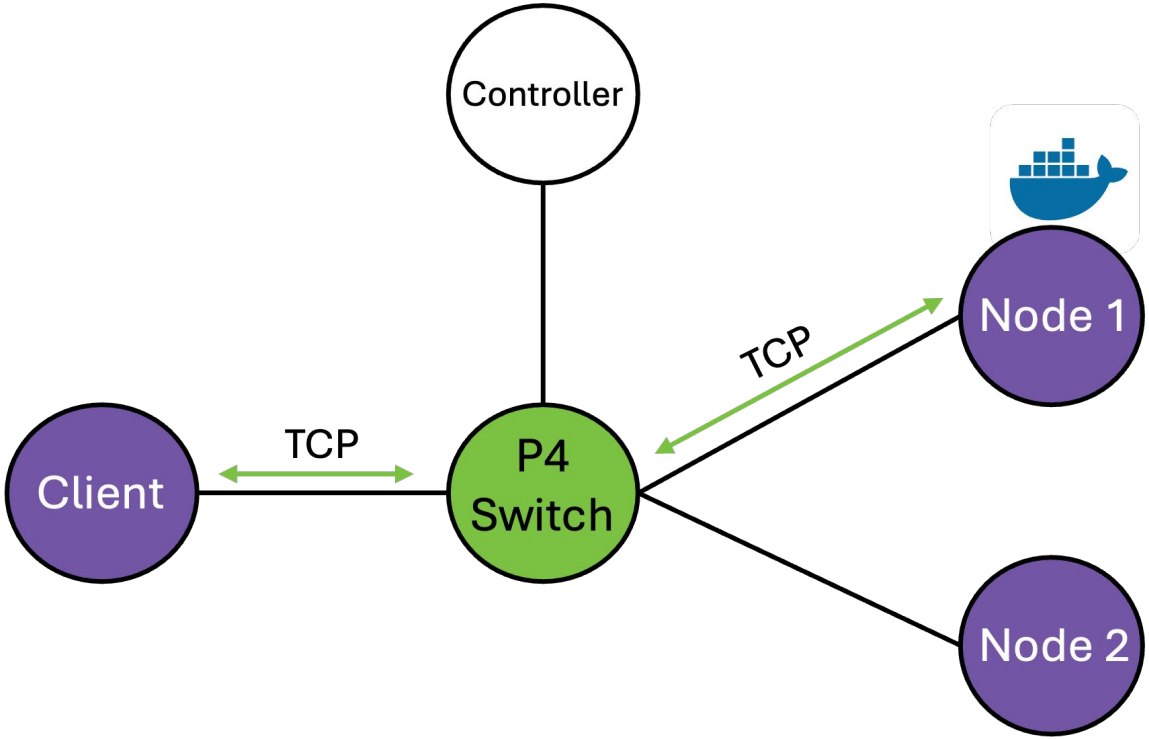
=> **keep the TCP connection alive**



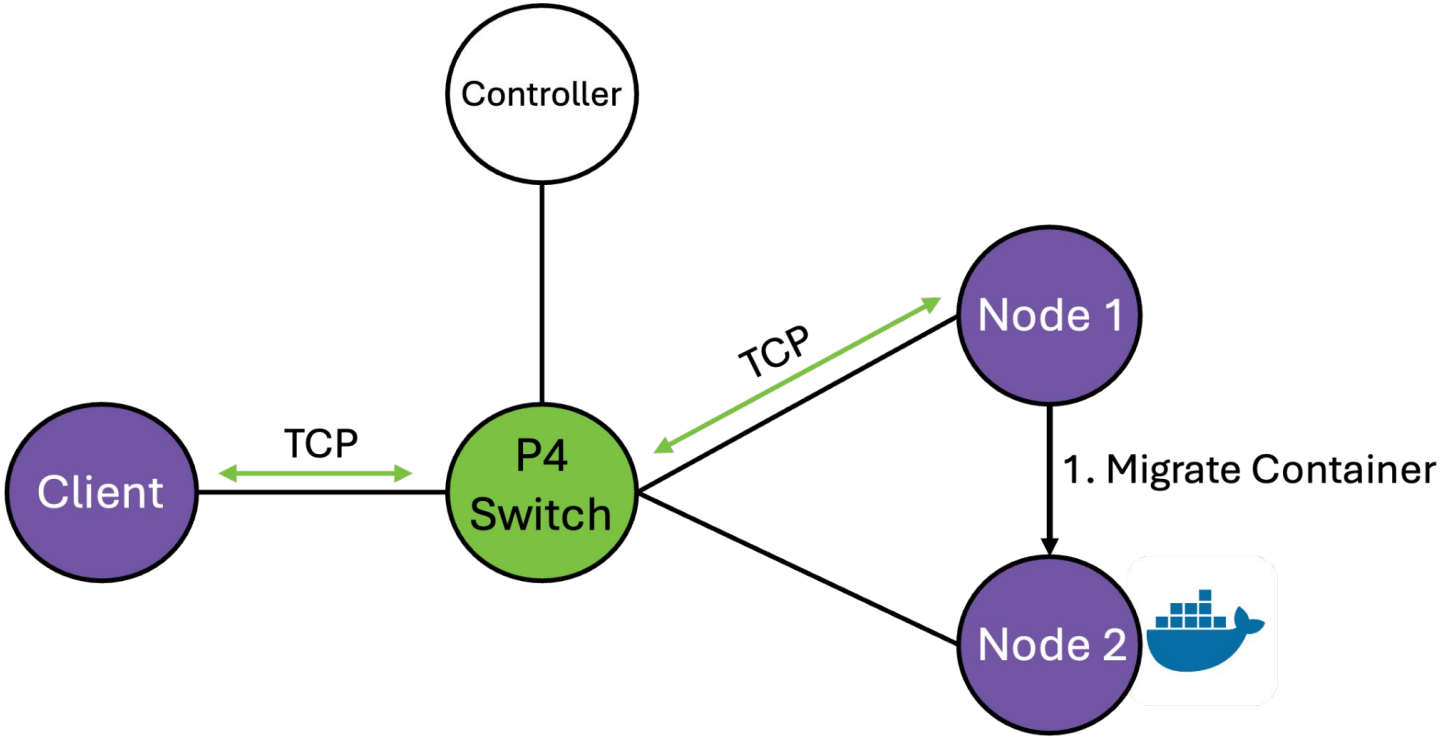
P4-Enabled Container Migration in Kubernetes



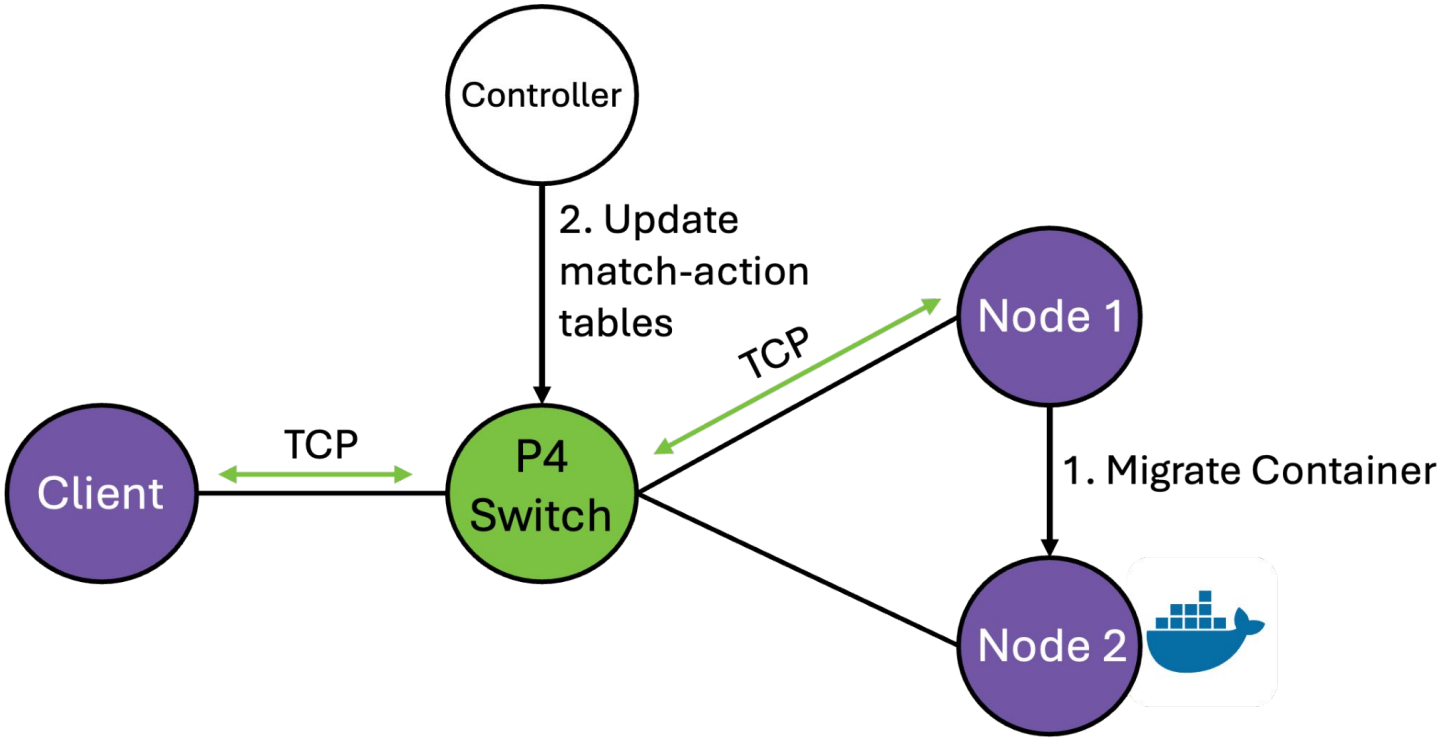
P4-Enabled Container Migration in Kubernetes



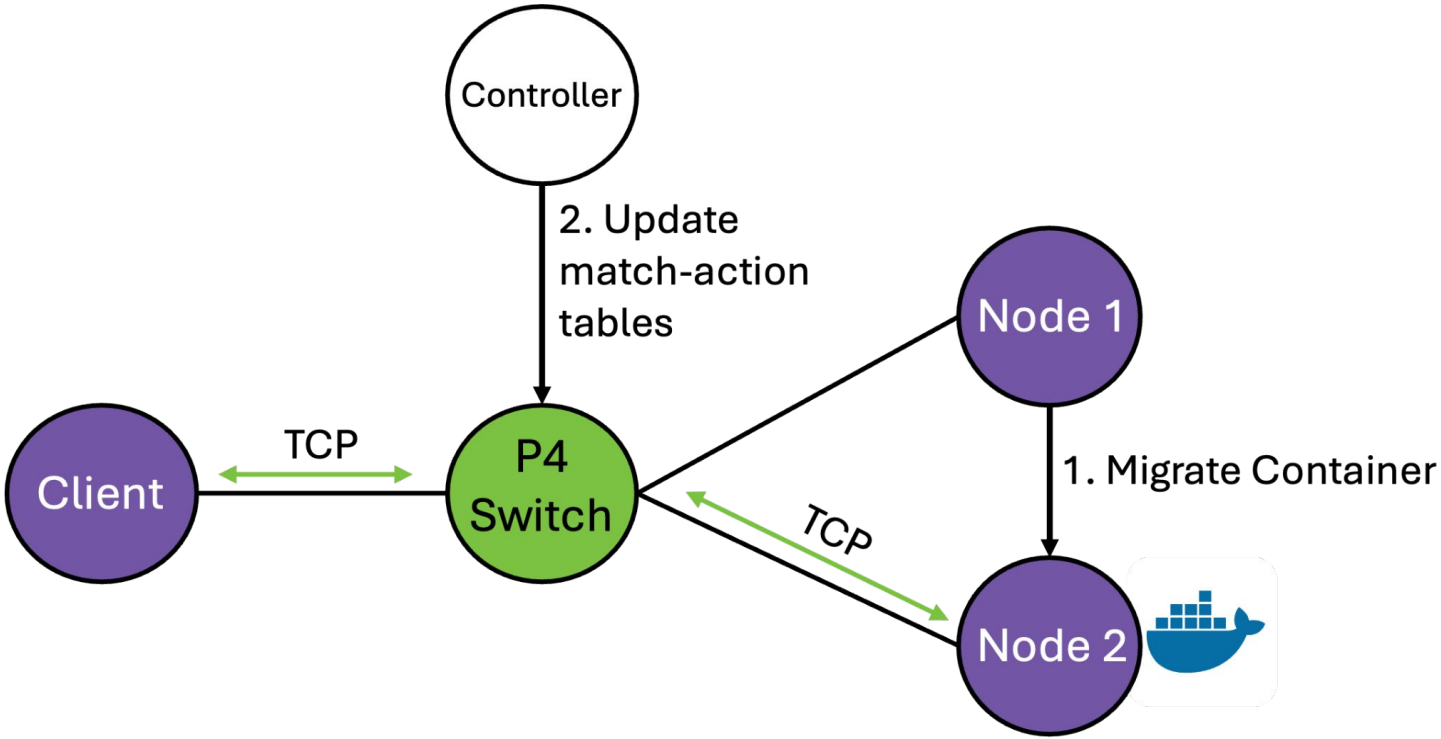
P4-Enabled Container Migration in Kubernetes



P4-Enabled Container Migration in Kubernetes

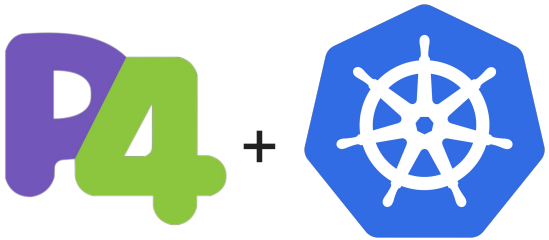


P4-Enabled Container Migration in Kubernetes

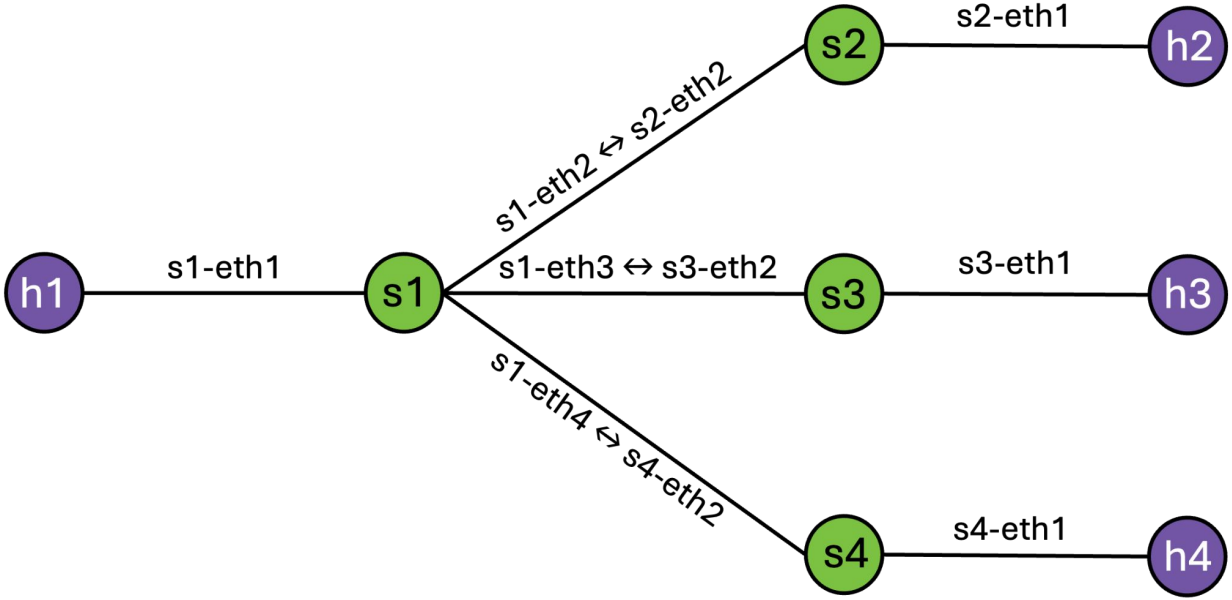


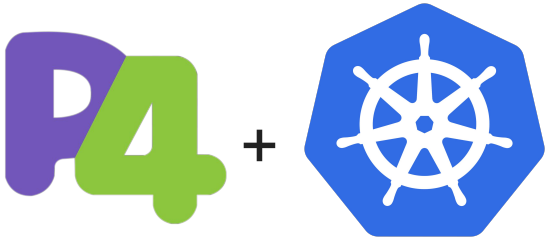
Process Migration

- **CRIU =** Checkpoint/Restore In Userspace
- Create a checkpoint of a Linux process, **including sockets**
- Restore process with sockets directly in *ESTABLISHED* state
- The IP address changes => rewrite address in the checkpoint file



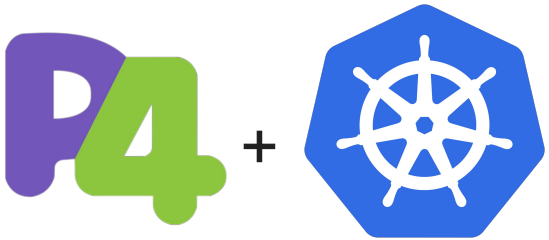
Process Migration: Topology



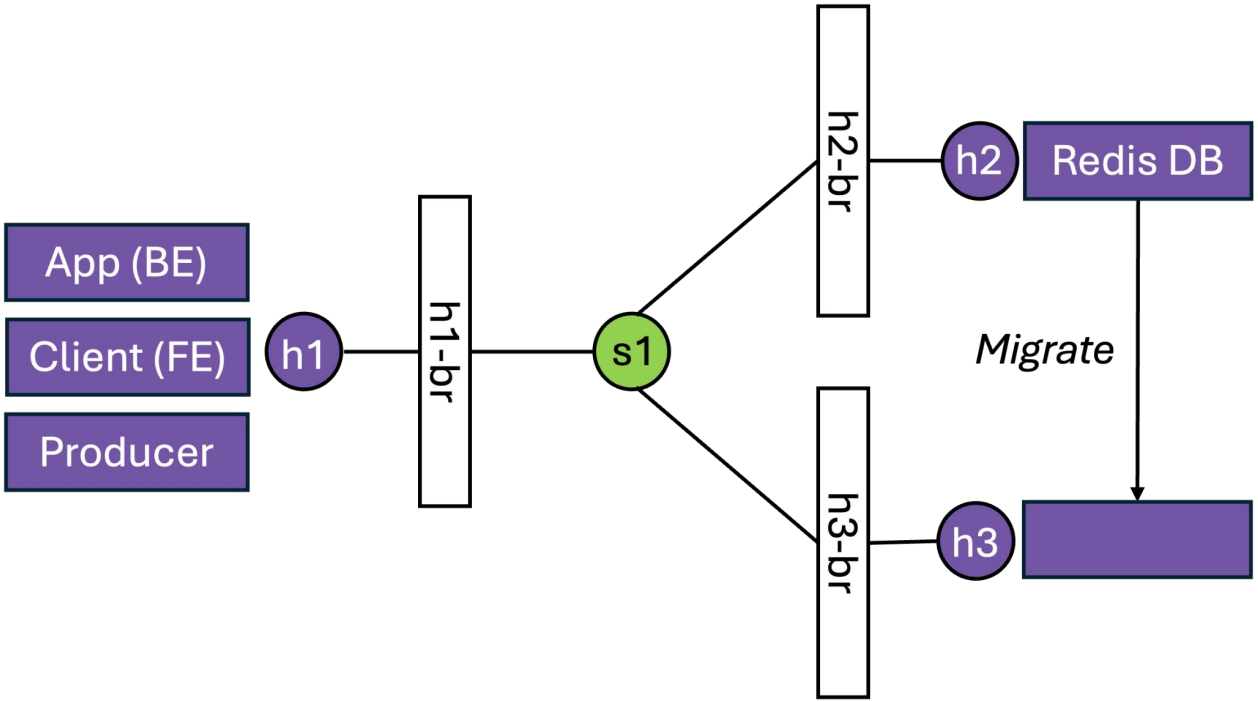


Container Migration

- Podman and Docker use CRIU to create container checkpoints
- **Idea:** Each pod in a separate network
- Migrate containers between pods

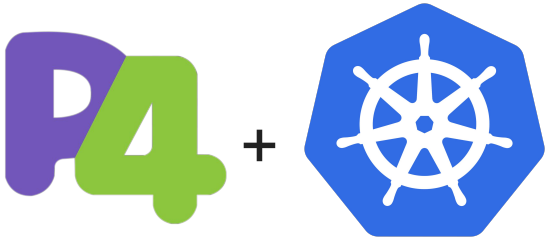


Container Migration: Demo



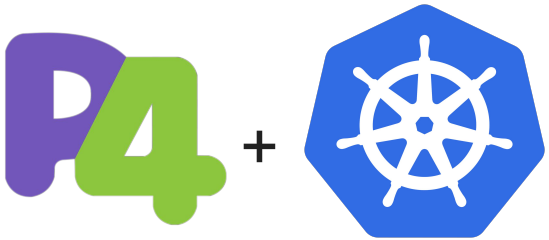
Container Migration: Demo

https://drive.google.com/file/d/1LKIvduw0_MY8WDp9iDVyUm09PqHz2f-P/view?usp=drive_link



P4-Enabled Container Migration in Kubernetes

- Container C/R is a beta feature in Kubernetes (kubelet API)
- **CRI-O** (Container Runtime Interface) uses CRIU to create container checkpoints
- **Problem 1:** Container migration is a manual process
- **Problem 2:** BMv2 switch in Kubernetes?

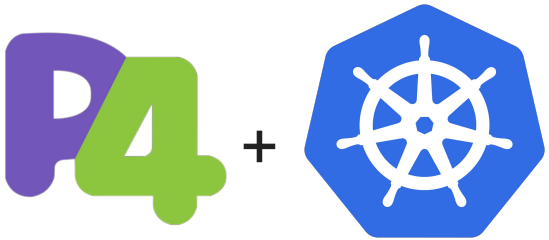


P4-Enabled Container Migration in Kubernetes

Problem 1: Container migration is a manual process

Solution: write a script to:

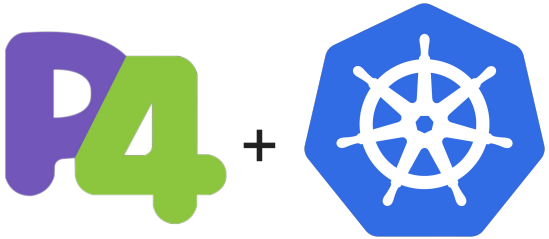
1. Call the kubelet checkpoint API
2. Build an OCI container image from the checkpoint
3. Push the image to a (local) container registry
4. Edit the deployment manifest to use the new image
5. Re-apply the manifest
6. Wait for Kubernetes to create the container



P4-Enabled Container Migration in Kubernetes

Problem 2: BMv2 switch in Kubernetes?

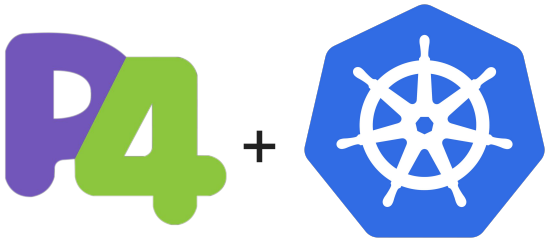
- **Solution 1:** Use MetalLB to route traffic to a specific node, extend kube-proxy to use BMv2, load-balance traffic **between nodes**
- **Solution 2:** Build a CNI plugin (similar to kube-router) to load-balance traffic **between pods on each node** using BMv2



P4-Enabled Container Migration in Kubernetes

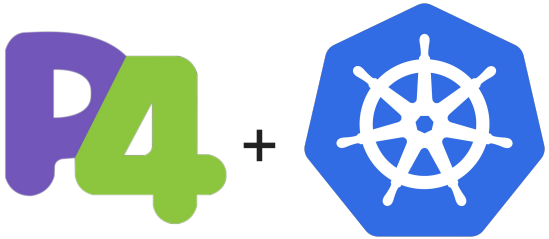
Load-balance traffic **between pods on each node** using BMv2:

1. Containerize the BMv2 switch and the controller
2. Write a custom CNI plugin to:
    a. attach container interfaces to BMv2 switch as ports
    b. update match-action tables
3. Deploy everything as a DaemonSet

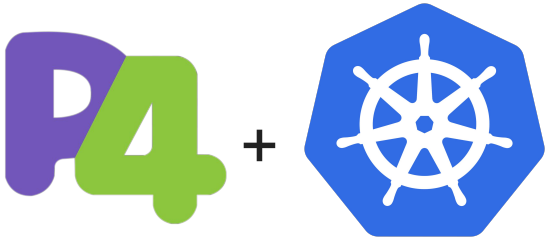


Future Work

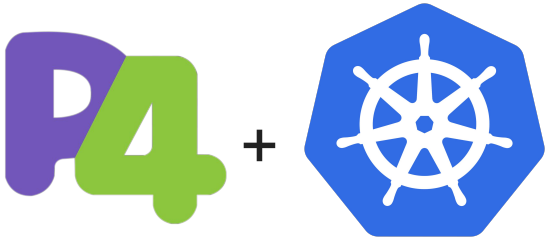
- Deploy load balancer on a **Tofino2** switch, test container migration in a **multi-node** Kubernetes cluster
- Deploy load balancer on **SmartNICs**, such as Nvidia BlueField
- Research ways to optimize the migration process in Kubernetes



Thank you!



Questions?



Links and Further Reading

- Github Repo
- GSoC Project Page
- Kubelet Checkpoint API
- Forensic Container Checkpointing
- Kube-router
- MetalLB
- Tofino2
- CRIU - Migrating TCP Connections