Newsletter

CloudSkin

July, 2025

Dynamic In-node Group-aware Scheduling for Multi-tenant Machine Learning Service on Kubernetes

This week, I have the pleasure to attend the **IEEE Cloud 2025 conference** in Helsinki and present our research article entitled **"Dynamic In-node Group-aware Scheduling for Multi-tenant Machine Learning Services on Kubernetes"**.

In this paper, we presented an in-node group-aware scheduling mechanism for multiple ML services, where each service scheduling contains a group resource selection and container resource assignment. We also provide a dynamic resource controller (DRC) to dynamically

by Peini Liu and Jordi Guitart (Barcelona Supercomputing Center)

reallocate the resource allocation for containers using the mechanism, which monitors the group changes and acts with the real containers' system cgroups adaptation. Through the dynamic resource controller implementation, workload involving multiple machine learning services shows significant performance improvement (up to 44%).

In addition, we also have the honour of receiving the **Best Paper Award from the prestigious conference**. Please feel free to contact us if you are interested in our research work.

cloudskin.eu
@cloudskin2023
github.com/cloudskin-eu

Funded by the European Union