ФИЗИКА ЭЛЕМЕНТАРНЫХ ЧАСТИЦ И АТОМНОГО ЯДРА
2025. Т. 56, вып. 6. С. 1990–1991

SCHEDULING ML AND HPC JOBS WITH SHOC PLATFORM OVER KUBERNETES

D. A. Petrosyan [1, *]

[1] Institute for Informatics and Automation Problems, NAS RA, Yerevan, Armenia

High-Performance Computing (HPC) and Machine Learning (ML) workloads are central to scientific advancements, simulations, and AI-driven applications. While traditional schedulers like Slurm have been widely adopted for their precise resource management and advanced scheduling capabilities in tightly coupled, resource-intensive environments, the rise of containerization and cloud-native technologies has introduced new paradigms in workload orchestration.

Kubernetes, a leading container orchestration platform, offers dynamic provisioning, auto-scaling, and seamless integration with cloud environments. These features make Kubernetes well-suited for flexible and scalable workloads. However, it was not originally designed with traditional HPC workloads in mind, presenting challenges such as hardware-specific granularity and dependency-aware scheduling.

The Shoc (Serverless HPC Over Cloud) platform addresses these challenges by extending Kubernetes' capabilities to support diverse and complex HPC and ML workflows. This paper explores the architecture and features of the Shoc platform, demonstrating its ability to provide an efficient, serverless experience for scheduling ML and HPC jobs over Kubernetes, effectively bridging the gap between traditional schedulers and modern, cloud-native environments.

Высокопроизводительные вычисления (HPC) и задачи машинного обучения (ML) играют ключевую роль в научных исследованиях, моделировании и приложениях на основе искусственного интеллекта. Несмотря на широкое применение традиционных планировщиков, таких как Slurm, благодаря их точному управлению ресурсами и передовым возможностям планирования в условиях тесно связанных и ресурсоемких сред рост контейнеризации и облачно-нативных технологий привнес новые подходы к управлению задачами.

Kubernetes — ведущая платформа для оркестрации контейнеров — предоставляет динамическое выделение ресурсов, автомасштабирование и бесшовную интеграцию с облачными средами. Эти возможности делают Kubernetes подходящей для гибких и масштабируемых задач. Однако изначально Kubernetes не была разработана с учетом традиционных HPC-задач, что создает такие проблемы, как управление на уровне специфики оборудования и планирование с учетом зависимостей.

Платформа Shoc (Serverless HPC Over Cloud) решает эти проблемы, расширяя возможности Kubernetes для поддержки разнообразных и сложных процессов HPC и ML. Рассматриваются архитектура и функции платформы Shoc, демон-

* E-mail: davit.petrosyan@iiap.sci.am

стрируется ее способность обеспечивать эффективное бессерверное управление задачами ML и HPC на базе Kubernetes, успешно преодолевая разрыв между традиционными планировщиками и современными облачно-нативными подходами.

PACS: 44.25.+f; 44.90.+c