**УДК 004.457**

# РАЗРАБОТКА ИНФРАСТРУКТУРЫ ДЛЯ АДАПТИВНОГО ВЫДЕЛЕНИЯ РЕСУРСОВ ПОД ЗАДАЧИ МАШИННОГО ОБУЧЕНИЯ

**Дунаев М.В.** (ИТМО),
**Научный руководитель – кандидат технический наук, старший научный сотрудник Ходненко И.В.** (ИТМО)

**Введение.** В современных системах машинного обучения, где модели становятся всё более ресурсоёмкими, ключевой задачей является рациональное использование вычислительных мощностей [1]. Высокая стоимость оборудования и возросшие требования к производительности диктуют необходимость масштабируемой архитектуры, позволяющей эффективно распределять нагрузку и избегать простоев, особенно при коллективной разработке [2].

**Основная часть.** Одним из наиболее гибких подходов к созданию такой инфраструктуры является контейнеризация. Она обеспечивает быструю подготовку окружения, изоляцию процессов и удобство развертывания моделей [3]. В предлагаемом решении используется «лизинг» Docker-контейнеров: для каждого разработчика создаётся индивидуальная среда с необходимыми библиотеками, а по окончании работы контейнер освобождает ресурсы.

При этом система мониторинга (Prometheus, Grafana) отслеживает загрузку GPU и CPU в реальном времени, что позволяет оперативно перераспределять ресурсы. Подобная архитектура не только повышает общий коэффициент использования оборудования, но и упрощает масштабирование: при росте числа задач легко добавить новые узлы. Реализованная модель доказала эффективность на практике, позволяя параллельно работать над проектами нескольким специалистам [4].

**Выводы.** Разработанный подход к контейнеризации и автоматическому лизингу ресурсов даёт возможность повысить производительность, упростить обслуживание систем машинного обучения и обеспечить отказоустойчивость. В дальнейшем планируется расширить функциональность за счёт интеграции с облачными платформами и внедрить механизмы предсказательной аналитики для ещё более гибкого управления ресурсами.

**Список использованных источников**:

1. Jauro F. et al. Deep learning architectures in emerging cloud computing architectures: Recent development, challenges and next research trend // Applied Soft Computing. – 2020. – Т. 96. – С. 106582.

2. Elzeki O. M., Rashad M. Z., Elsoud M. A. Overview of scheduling tasks in distributed computing systems // International Journal of Soft Computing and Engineering. – 2012. – Т. 2. – № 3. – С. 470–475.

3. Lwakatare L. E. et al. Large-scale machine learning systems in real-world industrial settings: A review of challenges and solutions // Information and software technology. – 2020. – Т. 127. – С. 106368.

4. Chen Q. et al. Design of an adaptive GPU sharing and scheduling scheme in container-based cluster // Cluster Computing. – 2020. – Т. 23. – С. 2179–2191.