Интеллектуальное прогнозирование времени выполнения вычислительных задач GROMACS для оптимизации работы планировщика задач суперкомпьютера "сHARISMa" НИУ ВШЭ

М.Т. Саликова, П.С. Костенецкий, Р.М. Мишенин
НИУ ВШЭ

Современные суперкомпьютеры являются незаменимыми инструментами для решения сложных научных и промышленных задач. Эффективное планирование потока вычислительных задач суперкомпьютера требует информации о времени, за которое может завершиться каждая из ожидающих запуска задач. Нередко пользователи устанавливают задаче временной лимит с большим запасом, что снижает эффективность работы планировщика. Разработка метода предсказания времени выполнения задач Gromacs на суперкомпьютере позволит более точно выстраивать поток задач, оптимально загружая вычислительные ресурсы суперкомпьютера.

Известные подходы предсказания длительности выполнения вычислительных задач обычно основываются на обобщенных метриках [1, 2], таких как число задействованных ядер и объем памяти, но не всегда учитывают специфические особенности отдельных приложений. Например, на время выполнения задач молекулярной динамики в пакете Gromacs влияют такие параметры, как количество атомов, временной шаг и общее количество шагов моделирования. Чем сложнее структура и выше плотность сетки, тем больше вычислительных ресурсов требуется, что влияет на общую длительность расчета.

Для осуществления исследования потребовалось агрегировать данные о задачах молекулярной динамики Gromacs, выполненных на суперкомпьютере сHARISMa. Датасет включает более двух тысяч записей, каждая из которых содержит такие параметры, как количество атомов, временной шаг, общее число шагов моделирования, число задействованных ядер, объем используемой памяти и фактическое время выполнения задачи. Для того чтобы увеличить пригодность для использования в прогнозных моделях, были учтены только данные о задачах, помеченных системой анализа HPC TaskMaster [3] как эффективные. Перед анализом датасет был подвергнут предварительной обработке, в том числе очистке от аномалий и нормализации, что позволило минимизировать влияние шумов и обеспечить корректность дальнейших вычислений.

На текущем этапе работы разработан алгоритм для статистического анализа, использующий ключевые параметры задач Gromacs для формирования прогноза времени выполнения. Следующей целью исследования является внедрение модели машинного обучения для повышения точности предсказания. После завершения этапа обучения модель будет протестирована и оценена ее эффективность в реальных условиях. В дальнейшем модель будет интегрирована в систему HPC TaskMaster для оптимизации потока задач суперкомпьютера.

Литература

1. Guo J. et al. Machine Learning Predictions for Underestimation of Job Runtime on HPC System // Supercomputing Frontiers, SCFA'18. Vol. 10776. Springer, 2018. P. 179–198. Lecture Notes in Computer Science. DOI: 10.1007/978-3-319-69953-0_11.
2. Gaussier E. et al. Improving backfilling by using machine learning to predict running times // Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. ACM, 2015. Article 64. P. 1–10. DOI: 10.1145/2807591.2807646.
3. Kostenetskiy P. et al. Enhancement of the Data Analysis Subsystem in the Task-Efficiency Monitoring System HPC TaskMaster for the сHARISMa Supercomputer Complex at HSE University // Parallel Computational Technologies, PCT'2024. Vol. 2241. Springer, 2024. P. 49–64. Communications in Computer and Information Science. DOI: 10.1007/978-3-031-73372-7_4.