ФИЗИКА ЭЛЕМЕНТАРНЫХ ЧАСТИЦ И АТОМНОГО ЯДРА
2025. Т. 56, вып. 6. С. 2007–2008

ML-BASED OPTIMUM NUMBER OF CUDA STREAMS FOR THE GPU IMPLEMENTATION OF THE TRIDIAGONAL PARTITION METHOD

M. Veneva [1,*], *T. Imamura* [1,**]

[1] RIKEN Center for Computational Science (R-CCS), Minatojima-minami-machi, Chuo-ku, Kobe, Hyogo, Japan

We present a heuristic for finding the optimum number of CUDA streams by using tools common to the modern AI-oriented approaches and applied to the parallel partition algorithm. A time complexity model for the GPU realization of the partition method is built. Further, a refined time complexity model for the partition algorithm being executed on multiple CUDA streams is formulated. Computational experiments for different SLAE sizes are conducted, and the optimum number of CUDA streams for each of them is found empirically. Based on the collected data, a model for the sum of the times for the nondominant GPU operations (that take part in the stream overlap) is formulated using regression analysis. A fitting nonlinear model for the overhead time connected with the creation of CUDA streams is created. Statistical analysis is done for all the built models. An algorithm for finding the optimum number of CUDA streams is formulated. Using this algorithm, together with the two models mentioned above, predictions for the optimum number of CUDA streams are made. Comparing the predicted values with the actual data, the algorithm is deemed to be acceptably good.

Представлена эвристика для поиска оптимального количества потоков CUDA с использованием инструментов, общих для современных подходов, ориентированных на ИИ, и применяемых к алгоритму параллельного разбиения. Построена модель временной сложности для реализации метода разбиения на GPU. Далее сформулирована уточненная модель временной сложности для алгоритма разбиения, выполняемого на нескольких потоках CUDA. Проведены вычислительные эксперименты для различных размеров СЛАУ, и найдено эмпирически оптимальное количество потоков CUDA для каждого из них. На основе собранных данных с помощью регрессионного анализа сформулирована модель для суммы времен для недоминирующих операций GPU (которые принимают участие в перекрытии потоков). Создана подходящая нелинейная модель для накладных расходов, связанных с созданием потоков CUDA. Для всех построенных моделей выполнен статистический анализ. Сформулирован алгоритм для поиска оптимального количества потоков CUDA. С использованием этого алгоритма вместе с двумя

* E-mail: milena.p.veneva@gmail.com
** E-mail: imamura.toshiyuki@riken.jp

упомянутыми выше моделями сделаны прогнозы для оптимального количества потоков CUDA. Сравнивая прогнозируемые значения с фактическими данными, можно сказать, что алгоритм является приемлемо хорошим.

PACS: 02.60.−x; 02.70.−c; 02.60.Ed