УДК 519.6:004.42:004.31

ОПТИМІЗАЦІЯ ПРОДУКТИВНОСТІ ПАРАЛЕЛЬНИХ АЛГОРИТМІВ РОЗВ'ЯЗАННЯ СЛАР НА GPU

Жульковський О.О. к.т.н., доц.[1]; Жульковська І.І., к.т.н., доц.[2]
[1] Дніпровський державний технічний університет;
[2] Університет митної справи та фінансів
[1] Україна, Кам'янське; [2] Україна, Дніпро
[1] olalzh@ukr.net; [2] inivzh@gmail.com

Анотація. *Досліджено ефективність реалізації алгоритмів розв'язання СЛАР на графічних процесорах із використанням JCuda. Проаналізовано вплив архітектури CUDA, методів матричних розкладів та перестановок матриць на продуктивність. Порівняльні тести CPU і GPU показали прискорення GPU-реалізацій у 5-7 разів.*
Ключові слова: *GPU-прискорення; СЛАР; CUDA; JCuda; матричні розклади.*

Вступна частина. Зі зростанням обсягів даних та складності моделей у сучасних обчислювальних системах постає потреба у прискоренні розв'язання систем лінійних алгебраїчних рівнянь (СЛАР), які є базовою складовою більшості задач чисельного моделювання. Використання графічних процесорів (GPU) як обчислювальних прискорювачів відкриває можливості для значного підвищення продуктивності обчислень, особливо при роботі з великими розрідженими матрицями. Архітектура CUDA від NVIDIA дозволяє реалізувати масово-паралельні алгоритми, а бібліотека JCuda забезпечує доступ до цих можливостей із мови Java, поєднуючи зручність високорівневого програмування з потужністю GPU [1].

Актуальність роботи зумовлена необхідністю створення ефективних алгоритмів розв'язання СЛАР, здатних забезпечити баланс між швидкодією, точністю та оптимальним використанням пам'яті при гібридних CPU/GPU-обчисленнях.

Мета дослідження – підвищення ефективності паралельних обчислень шляхом оптимізації реалізації матричних розкладів (Холецького, QR, LU) на GPU з архітектурою CUDA, дослідження впливу перестановок матриць на швидкодію та порівняння продуктивності з традиційними CPU-реалізаціями.

Основна частина. Для оцінювання ефективності паралельних обчислень СЛАР використано бібліотеку JCuda, яка забезпечує зручний інтерфейс доступу до архітектури CUDA безпосередньо з мови Java. Такий підхід дозволяє реалізовувати високопродуктивні GPU-орієнтовані алгоритми у високорівневому середовищі програмування, зберігаючи при цьому гнучкість та зручність розробки. У роботі активно використовувалися бібліотеки cuBLAS, cuSOLVER та cuSPARSE, призначені для виконання операцій лінійної алгебри, що істотно спрощує розроблення наукових застосунків та підвищує стабільність обчислень завдяки оптимізованим функціям NVIDIA.

Дослідження охоплює три класичні методи матричних розкладів: LU, QR і Холецького, які є основними при розв'язанні СЛАР різної структури [2]. Для підвищення ефективності та зменшення обсягу проміжних даних застосовано дві стратегії перестановки матриць – AMD (Approximate Minimum Degree) та RCM (Reverse Cuthill–McKee). Ці методи мінімізують заповнення матриць у процесі факторизації, покращують локальність доступу до пам'яті й, відповідно, знижують затрати часу на обчислення. Такий підхід особливо ефективний для розріджених матриць, де структура ненульових елементів суттєво впливає на швидкодію.

Експерименти проведено на графічному процесорі NVIDIA GTX 1660 Ti (1536 CUDA-ядер, 6 ГБ VRAM) у порівнянні з процесором Intel Core i5-9300H. Досліджено вплив розмірності (від 500×500 до 10000×10000) та розрідженості матриць на продуктивність алгоритмів. Для кожного методу розкладу виконано серії тестів із різними варіантами перестановок, що дозволило оцінити вплив структури даних на ефективність обчислень.

Результати показали, що GPU-реалізації забезпечують прискорення у 5-7 разів порівняно з CPU, причому найвищі показники досягнуто для великих матриць ($>10^4$ елементів) і при використанні QR-розкладу, який краще масштабується при паралельних обчисленнях. Застосування перестановок AMD та RCM дозволило додатково зменшити обсяг проміжних даних і підвищити балансування потоків, що забезпечило додаткове прискорення на 10-15 %.

Отримані результати підтверджують доцільність використання GPU для задач із великою кількістю невідомих і значною обчислювальною складністю. Для малих систем витрати на передавання даних між CPU та GPU стають суттєвими, що знижує загальну ефективність паралельної обробки. Тому перспективним напрямом подальших досліджень є гібридні CPU/GPU-рішення, які дозволяють комбінувати сильні сторони обох архітектур і адаптувати обчислення до конкретних типів задач.

Висновки. Проведене дослідження підтвердило доцільність застосування GPU для прискорення розв’язання СЛАР. Застосування JCuda у поєднанні з методами матричних розкладів дозволяє досягти значного підвищення продуктивності без зниження точності результатів.

Подальші дослідження передбачають: оптимізацію LU-розкладу для GPU; застосування методів змішаної точності для підвищення енергоефективності; розробку адаптивних гібридних стратегій CPU/GPU для масштабованих систем.

Отже, GPU-прискорення є перспективним інструментом для розв’язання СЛАР великої розмірності, що відкриває нові можливості у високопродуктивних обчисленнях [3, 4].

Отримані результати можуть бути використані для вдосконалення програм чисельного моделювання, наукових симуляцій і прикладних інженерних задач [5].

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] Kawtikwar, A., & Nagi, R. (2024). HyLAC: Hybrid linear assignment solver in CUDA. *Journal of Parallel and Distributed Computing, 187*, 104838. doi: 10.1016/j.jpdc.2024.104838

[2] Zhulkovskyi, O., Zhulkovska, I., Ulianovska, Yu., Kosuhina, O., & Riabovolenko, V. (2025). Performance improvement of parallel algorithms for solving SLAEs via GPU computing. *Informatics and Mathematical Methods in Simulation, 15*(3), 327–336. doi: 10.15276/imms.v15.no3.327

[3] Masciari, E., & Napolitano, E. V. (2024). Sustainability and high performance computing. *Lecture Notes in Computer Science*, 237–242. doi: 10.1007/978-3-031-78093-6_21

[4] Bouras, M., & Idrissi, A. (2023). A survey of parallel computing: Challenges, methods and directions. *Modern Artificial Intelligence and Data Science. Studies in Computational Intelligence*, 1102, 67–81. doi: 10.1007/978-3-031-33309-5_6

[5] Zhulkovskii, O., Panteikov, S., & Zhulkovskaya, I. (2022). Information-modeling forecasting system for thermal mode of top converter lance. *Steel in Translation*, 52(5), 495–502. doi: 10.3103/s0967091222050138

Zhulkovskyi O.O., Zhulkovska I.I.

Optimization of the Performance of Parallel Algorithms for Solving SLAEs on GPU

Abstract. *The efficiency of implementing algorithms for solving systems of SLAEs on graphics processors using the JCuda library has been investigated. The impact of CUDA architecture, matrix decomposition methods, and matrix permutation strategies on computational performance was analyzed. Comparative tests on CPU and GPU demonstrated a 5–7× speedup for GPU-based implementations.*

Keywords: *GPU acceleration; SLAE; CUDA; JCuda; matrix decompositions.*