УДК 004.942

Р.І. Дзінько, А.М. Дзінько, О.І. Лісовиченко

ДОСЛІДЖЕННЯ ЕФЕКТИВНОСТІ ВИКОНАННЯ АЛГЕБРАЇЧНИХ ОПЕРАЦІЙ НАД МАТРИЦЯМИ НА ГРАФІЧНОМУ ПРОЦЕСОРІ

Анотація: В роботі розглядається застосування GPGPU підходу до виконання операцій над матрицями (додавання, множення, множення на скаляр). Виконується порівняльний аналіз швидкості виконання операцій на ЦП та ГП однієї цінової категорії.

Ключові слова: матриця, GPGPU, cuda, ATI Stream, обчислювальне ядро, графічний процесор, центральний процесор.

Вступ

При виконанні проектування гнучких виробничих систем часто виникають задачі, основою математичного апарату для вирішення яких, використовуються матриці. Серед таких задач поширеними є: пряма задача кінематики, обернена задача кінематики, календарне планування, оперативне планування.

Часто для вирішення таких прикладних задач потрібні числові матриці великої розмірності. Операції над такими матрицями потребують значних обчислювальних ресурсів. Матриці допускають наступні алгебраїчні операції:

- додавання двох матриць однакового розміру;
- множення двох матриць підходящого розміру (матрицю, що має n стовпців можна множити тільки на матрицю, що має n рядків);
- множення матриці на скаляр,
- пошук оберненої матриці до заданої.

Особливістю даних операцій з точки зору їх виконання на обчислювальних ресурсах є можливість їх паралельного виконання, оскільки дані операції складаються з незалежних алгебраїчних операцій над числами, що є елементами матриці.

Паралельне виконання алгоритмів, що реалізують алгебраїчні операції над матрицями можливе з використанням багатоядерних процесорів чи багатопроцесорних систем.

На сьогоднішній день популярним стає використання графічних процесорів (ГП) для виконання обчислень загального призначення (GPGPU): технології nVidia CUDA та ATI Stream [4,5].

GPGPU (графічний процесор для обчислень загального призначення) – це техніка використання графічного процесора, що, зазвичай, використовується в обчисленнях для комп'ютерної графіки, яка дає змогу

© Р.І. Дзінько, А.М. Дзінько, О.І. Лісовиченко, 2012

 ISSN 1560-8956

виконувати обчислення загального призначення, які, зазвичай, виконуються на центральному процесорі [1].

Особливістю графічних процесорів є те, що вони проектуються для виконання великої кількості паралельних обчислень. Таким чином, при однаковій вартості один центральний процесор (ЦП), що може виконувати паралельно 4 потоки рівний вартості графічного процесора, що може виконувати від 20 до 50 паралельних потоків обчислень.

Проте, незважаючи, на апаратні можливості виконання паралельних обчислень, графічні процесори мають відмінну від класичних процесорів програмну модель, що називається “одна інструкція – масив даних”. Така програмна модель вимагає паралельного виконання лише однієї інструкції на заданому масиві даних, що унеможливлює використання умовних конструкцій та вимагає відповідної адаптації алгоритмів (рис. 1).

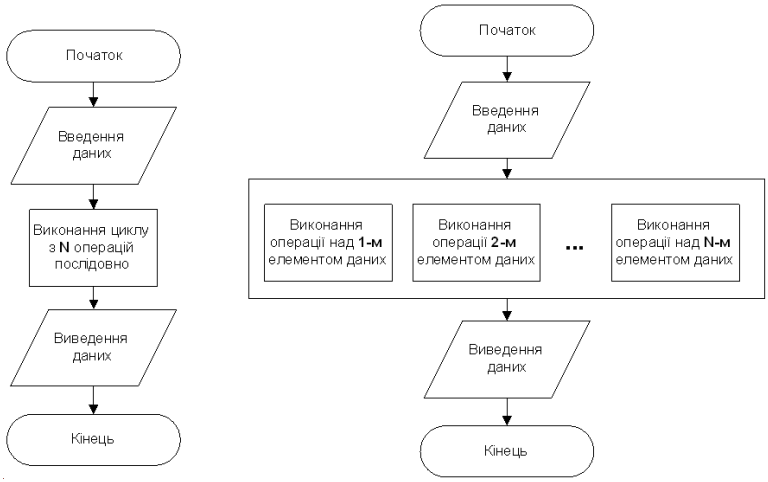

Рис. 1 – Виконання довільного алгоритму на:а) ЦП; б) ГП

При розробці алгоритму, який повинен виконуватись на графічному процесорі, слід враховувати особливість даної програмної моделі, а також те, що при вказаному підході далеко не всі алгоритми виконуватимуться ефективніше на ГП, ніж на ЦП.

Постановка задачі

Дано прямокутні матриці A та B розмірності $n \times m$ та алгебраїчні операції, операндом яких дана матриця може виступати: додавання двох матриць однакового розміру, множення двох матриць підходящого розміру (матрицю, що має n стовпців можна множити тільки на матрицю, що має n рядків), множення матриці на скаляр, пошук оберненої матриці.

Необхідно розробити програми, що реалізують виконання даних операцій над матрицями на ЦП та ГП та виконати порівняльний аналіз швидкості виконання алгоритмів.

Додавання двох матриць однакового розміру

Додаванням двох матриць однакового розміру $A + B$ є знаходження матриці C, всі елементи якої рівні попарно сумі всіх елементів матриці A та матриці B з відповідними індексами: $c_{ij} = a_{ij} + b_{ij}$.

Дана операція може бути представлена як алгоритм для ЦП у вигляді двох вкладених циклів, де на кожному кроці внутрішнього циклу обчислюється сума відповідних елементів двох вхідних матриць A та B, а результат присвоюється відповідній комірці матриці C (рис. 2. а).

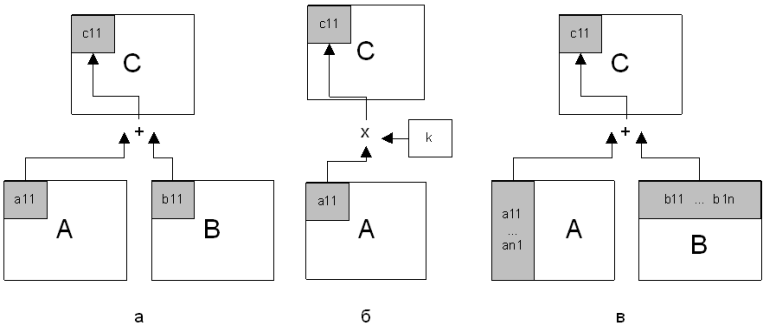

Рис. 2 – Операції над матрицями на ЦП (а – додавання матриць, б – множення матриці на скаляр, в – множення матриць)

Як бачимо з рис. 2.а, частина алгоритму, що включає вкладені цикли, може виконуватись паралельно, оскільки кожна алгебраїчна операція додавання може виконуватись незалежно від іншої.

В залежності від кількості ядер в процесорі, можна розділити матриці A та B на сегменти (наприклад, для 4-ядерного процесора, матриці можна розділити на 4 частини розмірності $\frac{n}{2} \times \frac{m}{2}$. Кожен паралельний потік буде заповнювати свою частину результуючої матриці.

Додавання матриць за допомогою ГП потребує створення так званого обчислювального ядра, що являє собою функцію на мові програмування С та виконується безпосередньо на ГП. Всі реалізації алгоритмів на ГП в даній статті наведені з використанням технології CUDA від компанії-виробника графічних процесорів nVidia [4].

Блок-схема алгоритму виконання операцій над матрицями зображена на рис. 3. Блок-схема є універсальною, а з точки зору реалізації різницею в операціях буде тільки обчислювальне ядро.

Особливістю обчислювального ядра є те, що він виконується на обчислювальній сітці. В даному випадку обчислювальною сіткою виступає сітка розмірності $n \times m$, що відповідає розмірності матриць-операндів.

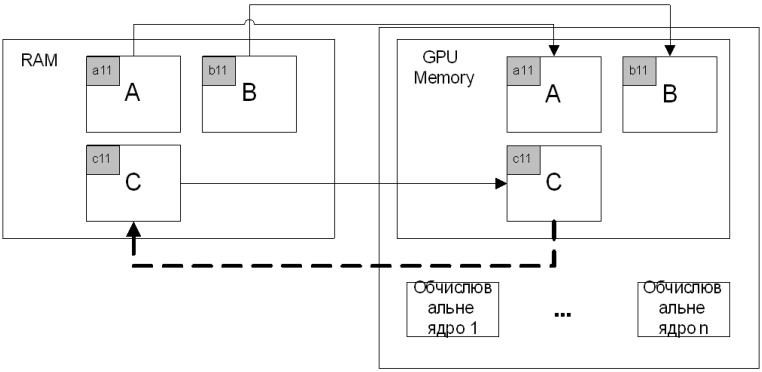

Рис. 3 – Виконання алгоритмів, що реалізують операції над матрицями на ГП (додавання матриць, множення матриці на скаляр, множення матриць)

Алгоритм, що реалізований в обчислювальному ядрі, виконується строго паралельно, де кожен потік виконує операцію на клітинці обчислювальної сітки, що в нашому випадку відповідає клітинці елементам матриць. Його можна представити у вигляді наступних кроків (рис. 3):

- визначення рядка обчислювальної сітки, що обраний в даний момент (є рядком матриці – операнда;
- визначення стовпця обчислювальної сітки, що обраний в даний момент (є стовпцем матриці – операнда;
- виконання операції додавання над елементами матриці, що визначаються індексами визначеними в попередніх пунктах та запис результату у відповідну комірку матриці-результату.

Реалізацію даного алгоритму було від тестовано на ГП nVidia 8800, що підтримує паралельне виконання до 20 потоків обчислень.

Множення матриці на скаляр

Операцією множення матриці A на скаляр k є знаходження матриці C, елементи якої отримуються множенням елементів початкової матриці A на число k: $c_{ij} = a_{ij} \times k$.

Дана операція може бути представлена як алгоритм для ЦП у вигляді двох вкладених циклів, де на кожному кроці внутрішнього циклу обчислюється сума відповідних елементів двох вхідних матриць A та B, а результат присвоюється відповідній комірці матриці C (рис. 2,б).

Алгоритм для виконання операції множення матриці на скаляр для ГП лід виконати у вигляді обчислювального ядра. Його можна представити у вигляді наступних кроків:

- визначення рядка обчислювальної сітки, що обраний в даний момент (є рядком матриці – операнда;

- визначення стовпця обчислювальної сітки, що обраний в даний момент (є стовпцем матриці – операнда;
- виконання операції множення елемента матриці, що визначається індексами визначеними в попередніх пунктах та запис результату у відповідну комірку матриці-результату.

Множення матриці на матрицю

Операцією множення матриці A на скаляр k є знаходження матриці C, елементи якої отримуються шляхом сумування попарних добутків відповідних елементів стовпця та рядка матриць - операндів, а результат формує елемент вихідної матриці із такими ж індексами: $c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$

Таким чином, у порівнянні з реалізаціями двох попередніх операцій, алгоритм представлятиме не 2, а 3 вкладених цикли, де у внутрішньому циклі обчислюється елемент результуючої матриці. Алгоритм для виконання на ЦП представлений на рис. 1. в.

Алгоритм для виконання операції множення матриць для ГП слід виконати у вигляді обчислювального ядра. Його можна представити у вигляді наступних кроків (рис. 3):

- визначення рядка обчислювальної сітки, що обраний в даний момент (є рядком матриці – операнда;
- визначення стовпця обчислювальної сітки, що обраний в даний момент (є стовпцем матриці – операнда;
- виконання циклу обчислення кожного елементу результуючої матриці (4).

Виконання операції множення елемента матриці, що визначається індексами визначеними в попередніх пунктах та запис результату у відповідну комірку матриці-результату.

Дослідження ефективності виконання алгоритмів

Для того, щоб отримати уявлення про ефективність виконання алгоритмів, виконаємо програми, що реалізують описані в попередніх розділах статті алгоритми на ЦП та ГП. Для того, щоб пересвідчитись в правильності проведених результатів, проведемо обчислення на квадратних матрицях різних розмірів. Час роботи алгоритмів наведений у мілісекундах в таблиці 1.

На рисунку 4 наочно показаний час виконання операцій на графічному та центральному процесорах при виконанні операцій над маленькими (рис 4.а) та великими числовими матрицями (рис. 4.б).

Висновки

З наведеної в попередньому розділі таблиці та гістограм можна зробити декілька висновків про ефективність реалізації алгебраїчних операцій над матрицями на графічному процесорі:

 ISSN 1560-8956

Таблиця 1

Результати обчислення часу роботи алгоритмів на квадратних матрицях різних розмірів

Алгоритм	Розмір квадратної матриці				
	2	4	16	64	2000
Додавання (ЦП), мс	0.054	0.058	0.677	1.022	321.2
Додавання (ГП), мс	0.101	0.106	0.107	0.132	16.5
Множення на скаляр (ЦП), мс	0.061	0.068	0.778	1.203	364.4
Множення на скаляр (ГП), мс	0.111	0.121	0.123	0.164	19.2
Множення матриць ЦП, мс	0.401	0.611	0.925	1.934	1023.2
Множення матриць (ГП), мс	0.181	0.191	0.199	0.243	75.4

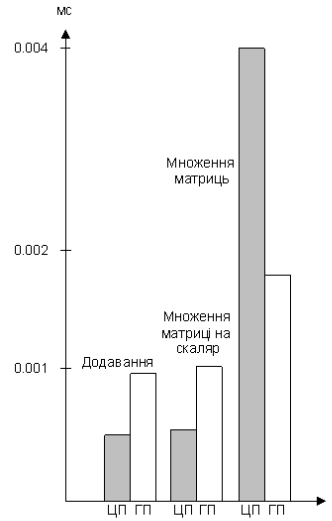

Рис 4. (а) Тест над матрицями розмірності 2x2

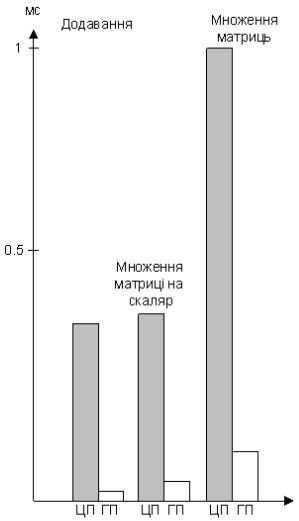

Рис 4. (б) Тест над матрицями розмірності 2000x2000

Рис. 4 – Гістограма часу виконання задач на графічному процесорі та центральному процесорі (а – для матриць розмірності 2x2, б – для матриць розмірності 2000x2000).

Всі алгоритми, що реалізують алгебраїчні операції над матрицями виконуються швидше на ГП, ніж на ЦП за рахунок можливості виконання обчислень паралельно. Це досягається за рахунок того, що обчислення є незалежними від проміжних результатів на різних етапах роботи алгоритмів.

Для матриць невеликого розміру час підготовки до власне виконання обчислювально складних етапів алгоритму при виконанні алго-

ритмів на ГП за рахунок виконання дій з підготування даних є високим, порівняно з власне процесор виконання обчислень. За рахунок цього тривалість виконання алгоритмів на ГП.

На великих розмірах матриць швидкість виконання алгоритмів значно підвищується і чим більше потокових ядер містить ГП, тим більший виграш в швидкості обчислень досягається

Отриманий аналіз дозволяє зробити висновок, що використання графічних процесорів для виконання алгебраїчних операцій над матрицями є ефективним та перспективним. Також це стосується і більш складних алгоритмів та прикладних задач, складовими яких є алгебраїчні операції над матрицями.

Література

1. GPGPU Review, Tobias Preis, European Physical Journal Special Topics 194, 87-119 (2011).
2. http://gpgpu.org/ - General-Purpose Computation Using Graphics Hardware.
3. GPGPU survey paper: John D. Owens, David Luebke, Naga Govindaraju, Mark Harris, Jens Kruger, Aaron E. Lefohn, and Tim Purcell. "A Survey of General-Purpose Computation on Graphics Hardware". Computer Graphics Forum, volume 26, number 1, 2007, pp. 80-113.

Отримано 13.02.2012 р.

 ISSN 1560-8956