УДК 004.89

О. В. Олещук, О. Є. Попель, М. Б. Копитчук
Одеський національний політехнічний університет

ДОСЛІДЖЕННЯ ЕНЕРГЕТИЧНОЇ ЕФЕКТИВНОСТІ ГРАФІЧНИХ ПРИСКОРЮВАЧІВ ФІРМИ NVIDIA

Детально розглянуто енерговитрати під час виконання обчислень на CPU і GPU. Подано математичний апарат для обчислення питомої енергетичної ефективності й проведено експерименти, що дають змогу порівняти енергетичну ефективність CPU і GPU, а також виявити її залежність від певних параметрів програмної реалізації.

Ключові слова: зелені технології, паралельні обчислення, GPU загального призначення, технологія CUDA.

STUDY OF ENERGY EFFICIENCY FIRM NVIDIA GRAPHICS ACCELERATORS

Energy consumption of calculations in CPU and GPU is considered in the article. The mathematical apparatus for calculating the share of energy efficiency and made a series of experiments that allow to compare the energy efficiency of CPU and GPU, as well as identify its dependence on a number of parameters of program implementation.

Key words: green technology, parallel computing, general purpose GPU, technology CUDA.

Постановка проблеми

У сучасному світі спостерігається кілька важливих взаємопов'язаних тенденцій. По-перше, зростає кількість мобільних пристроїв, поширюється сфера їх використання. Технології, що раніше були доступні лише для суперкомп'ютерів та високопродуктивних стаціонарних обчислювальних машин, поступово стають доступними для переносних пристроїв. Одним з таких досягнень є значне зростання потужності графічного процесора. По-друге, все актуальнішою стає проблема енергозбереження. І для переносних пристроїв, що не мають постійного доступу до електричної мережі, ця проблема ще більше загострюється. І, по-третє, об'єм даних, що повинен бути оброблений, збільшується швидкими темпами. Отже, доводиться одночасно розв’язувати дві взаємовиключні задачі: підвищувати продуктивність обчислень і знижувати енерговитрати [1].

Аналіз літературних джерел

Перспективним напрямом для вирішення поставлених завдань є перенесення основних обчислень з центрального процесора (CPU) на графічне ядро (GPU) з використанням технології CUDA як програмної платформи [2]. Раніше були проведені дослідження деяких показників енерговитрат і обчислювальної потужності графічного ядра. Зокрема, в [3] розглянуто показники енергетичної ефективності, які давали змогу відповісти на питання: як довго здатна пропрацювати обчислювальна система в автономному режимі, виконуючи певні обчислення?; яка необхідна потужність джерел живлення; яким буде енергоспоживання за максимального навантаження CPU і GPU?, яка частина енергії витрачається на роботу операційної системи і периферійних пристроїв, а яка – безпосередньо на обчислювальні операції? В [4] розглянуто питання можливого підвищення

продуктивності у разі переходу з CPU на GPU безвідносно економії або втрат енергії. Якщо об'єднати підсумки цих двох досліджень, виникає запитання: яка ціна виконання однієї обчислювальної операції з енергетичного погляду?

Мета

Для відповіді на поставлені вище питання потрібно ввести поняття питомої енергетичної ефективності. Під нею розумітимемо показник витрат електроенергії, необхідних для виконання певного заздалегідь заданого набору обчислювальних операцій. Також потрібні математичні формули, що пов'язують введене поняття з іншими параметрами обчислювальної системи і надають змогу визначити питому енергетичну ефективність експериментально.

Модель для тестування питомої енергетичної ефективності

Як обчислювальна модель, на якій проводилося тестування питомої енергетичної ефективності, використано повнозв'язний персептрон Розенблатта (рис. 1). Цей різновид нейронних мереж [5] має такі важливі властивості, як високий природний паралелізм, висока обчислювальна складність за порівняно невеликої кількості вхідних і вихідних даних, а також представлення даних у форматі дійсних чисел з одинарною точністю. Саме для вирішення таких завдань ідеально підходить архітектура графічного ядра і, відповідно, технологія CUDA.

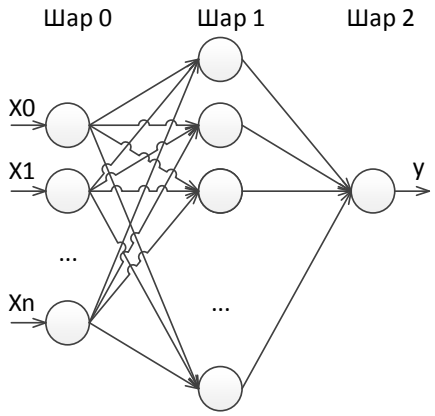

Рис. 1. Модель нейронної мережі, що реалізується

Для порівняльного аналізу питомої енергетичної ефективності ця нейронна мережа була реалізована на центральному процесорі, а потім – на графічному ядрі за допомогою технології CUDA. Нейронна мережа завжди була тришаровою і завжди мала один вихідний нейрон. Для виявлення залежності питомої енергетичної ефективності від особливостей виконаних обчислень у ході експерименту змінювали певні параметри. Такі параметри, як I – кількість входів нейронної мережі та H – кількість нейронів у прихованому шарі нейронної мережі, є власне параметрами нейронної мережі. Також інтерес становить величина R, яка відповідає кількості одночасно виконуваних потоків на графічному ядрі.

Використано ще два додаткові параметри:

C_0 – кількість наборів вхідних значень, що обчислюються у вигляді єдиного пакета даних;

C_1 – кількість повторень обчислень між вимірами рівня заряду акумулятора.

Параметри C_0 та C_1 необхідні для врахування впливу особливостей програмної реалізації на питому енергетичну ефективність, зокрема, впливу особливостей взаємодії між CPU і GPU,

Оскільки у складі CPU і GPU є засоби підвищення продуктивності за рахунок одночасного виконання множини операцій, такі як конвеєрна і матрична організація GPU, система передбачення розгалужень у CPU тощо, то як одиницю обчислень є сенс використовувати доволі великий блок операцій. А оскільки в експериментах тестувалися різні реалізації нейронних мереж, то за одиничний блок операцій приймали один розрахунок значення вихідного нейрона. Але оскільки в ході експери-

ментів тестувалися нейронні мережі з різною кількістю вхідних та прихованих нейронів, то безпосередньо порівняти можливо результати тільки тих експериментів, в яких параметри I та H збігалися. Параметри R, C_0 та C_1 характеризують лише особливості технічної реалізації, тому допустиме безпосереднє порівняння результатів експериментів, що відрізняються тільки значеннями цих параметрів.

Показники енергетичної ефективності

У [3] розглянуто низку показників, що прямо або побічно відображають витрати енергії на виконання обчислень. Найінформативнішим з них є параметр b_p – енергетична ефективність обчислень, що характеризує витрати безпосередньо на обчислення, які тестуються. Фізичний зміст цієї величини – це зменшення заряду акумулятора за одиницю часу в разі стовідсоткового завантаження обчислювача. Одиницею вимірювання цієї величини є %/хв, тобто у показнику b_p як складової частини міститься час. Для усунення часової складової можна врахувати продуктивність обчислювальної системи, тобто кількість операцій, виконаних за одиницю часу.

Тоді поняття "питома енергетична ефективність" визначатиметься як

$$b_0 = \frac{b_p}{P},$$

де P – кількість одиничних блоків операцій, які виконуються за одиницю часу.

Приблизно величину P можна обчислити за формулою

$$P = \frac{K \cdot C}{T_H},$$

де K – кількість вимірювань заряду акумулятора; T_H – гіпотетично максимально можливий час роботи від акумулятора за розряду від 100 до 0 %.

Вираз $K{\cdot}C$ визначає загальну кількість обчислень виходів нейронної мережі, виконаних протягом одного експерименту.

Величину T_H можна обчислити за формулою

$$T_H = \frac{100}{Q_1 - Q_K}, \tag{1}$$

де T – інтервал часу виконання одиничного експерименту; Q_1, Q_K – рівні заряду акумулятора на початку і в кінці експерименту відповідно.

Але оскільки в обчисленні T_H за формулою (1) беруть участь одиничні значення двох крайніх реєстрацій рівня заряду, то отриманий результат чутливий до одиничних викидів.

Для точнішого визначення величини P можна врахувати, що обчислювальна потужність CPU і GPU постійна, а значить, зростання кількості виконаних операцій з плином часу є лінійним, причому вільний член лінійної регресії дорівнює нулю. Тоді

$$M_k = P \cdot t_k + \varepsilon_M,$$

де M_k – кількість операцій, виконана до моменту k-го вимірювання; t_k – момент часу k-го вимірювання; ε_M – похибка кількості виконаних операцій.

Величину M_k можна визначити за формулою

$$M_k = C \cdot k. \tag{2}$$

Тоді, використовуючи метод найменших квадратів, точніше значення P визначимо за формулою

$$P = \frac{\bar{t} \cdot \overline{M}}{\overline{t^2}}. \tag{3}$$

Враховуючи (2) і виконавши необхідні спрощення на підставі формули (3), отримаємо

$$P = \frac{C(K+1)}{2} \cdot \frac{\bar{t}}{\overline{t^2}}.$$

Як і у випадку з параметром b_p, значення величини b_0 завжди є негативним, що відображає витрати енергії зі зростанням кількості операцій. Що менше абсолютне значення величини b_0, то менше витрат енергії припадає на одну обчислювальну операцію.

Результати, наведені в табл. 1, дають змогу з'ясувати, як співвідносяться витрати енергії, необхідні для виконання одного розрахунку виходу нейронної мережі при 1024 нейронах у прихованому шарі і 32 вхідних нейронах.

Як видно, реалізації, основані на технології CUDA, забезпечують значно вищу продуктивність. Порівняно з реалізацією на CPU продуктивність зростає в 7,8 та 18,2 разу за 128 і 512 потоків відповідно. І крім цього, обидві реалізації на GPU є енергетично ефективнішими з розрахунку енерговитрат за одиницю часу (рис. 2).

У підсумку виграш у питомій енергетичній ефективності, яка визначається двома цими факторами, у разі використання GPU як обчислювача виявляється ще вищим. А саме: за 128- та 512-потокової реалізації на GPU виконання однієї обчислювальної операції виявляється економічнішим у 11,2 та 24,0 рази відповідно.

Таблиця 1

Питома енергетична ефективність, якщо I = 32, H = 1024

Обчислювач	C_0	C_1	R	$P \cdot 10^{-3}$, оп./хв	$b_0 \cdot 10^6$, %/оп.
CPU	1	10000	–	67,8	-12,53
GPU	10000	100	128	526,4	-1,12
GPU	10000	100	512	1235,2	-0,52

І хоча, згідно з результатами, показаними на рис. 2, використання меншої кількості потоків графічного ядра видавалось енергетично ефективнішим, оскільки потребувало менших енерговитрат за одиницю часу, у підсумку під час розрахунку енерговитрат на операцію виявилося, що зниження кількості потоків не веде до реальної економії, оскільки зі зменшенням кількості потоків обчислювальна потужність знижується значно більше, ніж енергоспоживання.

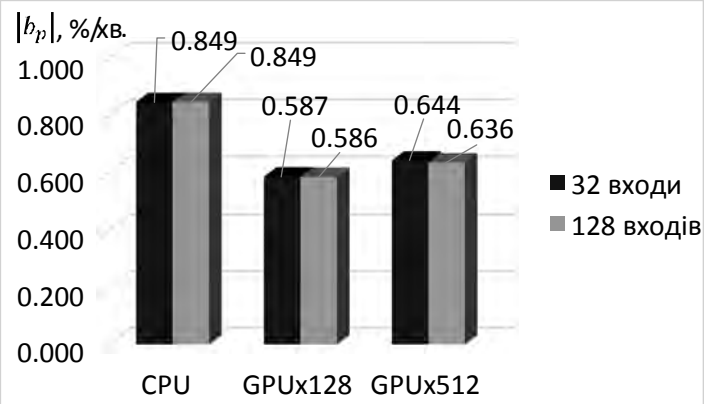

Рис. 2. Залежність енергетичної ефективності від кількості входів і кількості потоків

Дані табл. 2 показують, що технологія CUDA зберігає колосальну перевагу в енергетичній ефективності й за іншої конфігурації нейронної мережі. В цьому випадку збільшено кількість входів нейронної мережі до 128, а кількість нейронів у прихованому шарі знижено до 512. Виграш у продуктивності у цьому випадку досяг 36,4 разу, а у питомій енергетичній ефективності – в 44,1 разу.

Нейронна мережа, що складається з 32 вхідних і 1024 прихованих нейронів (табл. 3), як така, що являє собою базову конфігурацію в описуваних експериментах, була досліджена ретельніше, і на її основі можна зробити висновки про вплив розміру пакета даних на питому енергетичну ефективність.

Таблиця 2

Питома енергетична ефективність за I = 128, H = 512

Обчислювач	C_0	C_1	$P \cdot 10^{-3}$, оп./хв	$b_0 \cdot 10^6$, %/оп.
CPU	1	10000	25,3	-31,64
GPU	10000	100	920,5	-0,72

Якщо $C_0 = 1$, тобто у випадку, коли кожен набір входів нейронної мережі розглядався як окремий пакет даних, виникали істотні накладні витрати на взаємодію між CPU і GPU, і, як наслідок, виграш у продуктивності порівняно з реалізацією розрахунків на CPU становив усього 3,8 разу, а у питомій енергетичній ефективності – 3,2 разу.

Об'єднавши дані у пакети по 1000 вхідних наборів, вдалося підвищити продуктивність майже до максимуму, досяжного у цій серії експериментів, однак накладні витрати в цьому випадку все ще були великі, і через значне навантаження на CPU енерговитрати за одиницю часу залишалися на вищому рівні, ніж у разі реалізації обчислень тільки на CPU. Але загалом за рахунок істотного приросту продуктивності питома енергетична ефективність цієї реалізації на GPU виявилася вищою у 22 рази.

Таблиця 3

Питома енергетична ефективність, якщо I = 128, H = 1024

Обчислювач	C_0	C_1	R	$P \cdot 10^{-3}$, оп./хв	$b_0 \cdot 10^6$, %/оп.
CPU	1	10000	–	12.5	-68.20
GPU	1	100000	512	47.4	-21.41
GPU	1000	1000	512	309.7	-3.11
GPU	10000	100	128	154.9	-3.78
GPU	10000	100	512	310.3	-2.05

Підвищивши C_0 до 10 тис. і знизивши кількість потоків до 128, вдалося знизити енергоспоживання за одиницю часу, але при цьому істотно зменшилася продуктивність. Взагалі для цієї реалізації продуктивність вища у 12.4 разу, питома енергетична ефективність – у 18 разів порівняно з реалізацією на CPU. Найкращим рішенням за всіма показниками виявилося об'єднання даних у пакети по 10 тис. вхідних наборів зі збереженням кількості потоків – 512. При цьому продуктивність зросла у 24,9 разу, а питома енергетична ефективність – у 33,3 разу.

Висновок

Виконуючи обчислення, доцільно звертати увагу на витрати електроенергії, потрібні для цього. Як було показано, на енерговитрати істотно впливає вибір апаратного обчислювача. У цьому разі як можливі альтернативи розглядалися CPU та GPU. Споживання енергії у процесі обчислень можна порівняти за двома основними показниками: енергетичною ефективністю обчислень і питомою енергетичною ефективністю.

Енергетична ефективність обчислень характеризує енерговитрати за одиницю часу в разі максимально повного завантаження обчислювача за вибраних значень параметрів математичної моделі й відповідної програмної реалізації. За цим показником ефективність реалізацій на CPU і GPU відрізнялася на десятки відсотків [3], причому як в один, так і в інший бік. Це свідчить про залежність якості програмної реалізації від правильного підбору певних параметрів. Особливе значення у цьому випадку має об'єднання обчислень в обробку великих масивів даних.

Питома енергетична ефективність у багатьох випадках надає ціннішу інформацію, оскільки вона відображає енерговитрати у розрахунку на одну обчислювальну операцію. Оскільки цей показник враховує не тільки витрати енергії за одиницю часу, а ще й безпосередньо пов'язаний з продуктивністю обчислювача, то програмні реалізації на базі технології CUDA за питомою енергетичною ефективністю виявляються в рази вигіднішими від аналогічних рішень на CPU. Зокрема, у серії проведених досліджень вдалося досягти економії енергії у понад тридцять разів.

1. Зеленая ИТ-инженерия. В 2-х томах. Том 1. Принципы, компоненты, модели / под ред. В. С. Харченко. – Министерство образования и науки Украины, Национальный аэрокосмический университет им. Н. Е. Жуковского "ХАИ". – 2014. – 594 с. 2. CUDA Zone. http://www.nvidia.ru/object/cuda_home_new_ru.html (Last access: 18.07.2015). 3. Oleshchuk O., Popel O., Kopytchuk M. Study of energy efficiency of CPU and GPU // 14th International Conference "Research and Development in Mechanical Industry" RaDMI, 2014, Topola, Serbia. 4. Олещук О. В. Моделювання повнозв'язної нейронної мережі з використанням технології CUDA / Олещук О. В., Попель О. Є., Копитчук М. Б. // Вісник Національного університету "Львівська політехніка". – 2012. – № 747. – С. 131–139. 5. Каллан Р. Основные концепции нейронных сетей; пер. с англ. – М.: Издательский дом "Вильямс", 2001. – 288 с.