ISSN 2522-9818 (print)
ISSN 2524-2296 (online) *Innovative technologies and scientific solutions for industries. 2023. No. 1 (23)*

УДК 004.75 DOI: https://doi.org/10.30837/ITSSI.2023.23.072

О. Мамчич, М. Волк

ОЦІНЮВАННЯ ЕНЕРГЕТИЧНИХ ВИТРАТ У ПРОЦЕСІ ВИКОРИСТАННЯ МОБІЛЬНИХ ПРИСТРОЇВ ДЛЯ ХМАРНИХ ОБЧИСЛЕНЬ

Сучасні обчислювальні задачі потребують зростання обчислювальних потужностей. Це викликає необхідність створення та виробництва нового обладнання для хмарних обчислень. Одночасно з цим кількість персональних мобільних пристроїв уже вимірюється мільярдами, і навіть часткове їх залучення могло б зменшити вимоги до виробництва. Крім того, мобільне апаратне забезпечення є більш енергоефективним, що сприяє значному заощадженню енергії. У статті досліджено питання якісної та кількісної оцінки ефективності використання мобільних пристроїв для обчислень порівняно з традиційними стаціонарними рішеннями. **Мета роботи** – обґрунтувати таку гіпотезу: обчислення в хмарі на основі мобільних пристроїв суттєво зменшує використання енергії, ніж обчислення на стаціонарному обладнанні. Для цього показано, що обчислення на мобільному графічному процесорі є більш енергетично ефективним, ніж обчислення на стаціонарному процесорі. Для визначення якісної переваги проаналізовано публічні джерела та бенчмарки. На основі досліджених даних обчислено показники ефективності для різноманітних мобільних і стаціонарних графічних процесорів. Аргументовано, що здебільшого мобільні рішення витрачають суттєво менше енергії порівняно з стаціонарними рішеннями. Для обчислення кількісної переваги проведено експеримент на основі двох платформ: мобільної та стаціонарної. Одну й ту саму обчислювальну задачу було реалізовано за допомогою Apple Metal та NVidia CUDA. На основі цієї задачі обчислено показники енергетичної ефективності мобільного й стаціонарного графічного професора. За результатами дослідження визначено суттєву перевагу мобільного графічного процесора в перерахунку на енергетичну ефективність. Цей результат є релевантним, оскільки платформи вийшли в один рік із різницею у кілька місяців, отже, їх можна вважати ровесницями одна одній. Надані підходи не враховують споживання всіх інших частин системи, крім графічних процесорів. Це означає, що споживання материнської плати, блоку живлення тощо можуть схилити перевагу на користь мобільного процесора ще більше. Але для розподілених обчислень дуже важливим є мережне з'єднання, що може споживати суттєву кількість енергії на мобільному пристрої. Подальші дослідження стосуватимуться більш всеосяжного обліку споживання енергії різними підсистемами комп'ютера.

Ключові слова: обчислювальна хмара; хмара смартфонів; енергетична ефективність; хмарні обчислення; CUDA; Metal; апаратне прискорення.

Вступ

У сучасному світі якість обчислювальних потужностей є значною проблемою. Вимоги до швидкості та масштабування обчислювальної техніки зростають, що пов'язано із збільшенням обсягів даних, які необхідно обробляти, поширенням штучного інтелекту, машинного навчання та інших високотехнологічних проєктів.

Водночас створюється значний тиск на виробників комп'ютерів та інших пристроїв, що мають забезпечити необхідну продуктивність і ефективність у різноманітних умовах. Щоб відповісти на ці виклики, багато компаній зосереджуються на розробленні новітніх технологій, таких як обчислювальні хмари, що здатні забезпечити якісну швидкість і масштабованість. Збільшується також попит на енергоефективні та екологічні рішення, оскільки споживання енергії та викиди відходів з обчислювальної техніки можуть завдати негативного впливу на довкілля. Загалом, зростання вимог до обчислювальної потужності є важливою тенденцією, що визначає розвиток сучасного світу й зумовлює величезний прогрес у різних галузях науки й техніки.

Збільшення попиту на обчислювальні хмари – один із найбільш помітних напрямів у світі IT-індустрії. Обчислювальні хмари містять значну кількість обчислювальних ресурсів, таких як процесори, пам'ять, сховища даних і мережні ресурси, що надаються користувачам через інтернет. Зростання популярності хмарних обчислень зумовлене багатьма факторами, зокрема безперервним збільшенням обсягів даних, розвитком інтернет-технологій та обсягів складних обчислювальних задач у різних галузях. Хмарні обчислення, як зручний і економічний спосіб розв'язання задач, стають дуже ефективним інструментом.

За останні роки ринок хмарних обчислень значно зріс, пропонуючи більші обсяги обчислювальних ресурсів за доступнішими цінами. Водночас поширення хмарних обчислень сприяє розвитку нових технологій та інновацій, що полегшують життя бізнесу та користувачів.

Сучасний стан наукових досліджень та технологій в промисловості. 2023. №1 (23)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

У контексті зростання попиту на обчислювальну потужність енергетична ефективність обчислювальних систем також є дуже важливою проблемою з погляду економії ресурсів і збереження довкілля. Інтернет, соціальні мережі, штучний інтелект, системи автоматизації, наукові дослідження тощо потребують значної кількості обчислювальних ресурсів. Але водночас зростають і витрати енергії, потрібної для живлення цих систем. Це може створювати проблеми зі збільшенням викидів вуглецю та забрудненням довкілля. Розробники обчислювальної техніки й програмного забезпечення працюють над створенням більш енергоефективних систем, над використанням відновлювальних джерел енергії та вдосконаленням технологій охолодження.

За результатами досліджень компанії IDC, 2021 р. світова кількість відвантажених смартфонів становила 1,24 млрд одиниць [1]. З огляду на те, що смартфони зазвичай використовуються протягом кількох років, загальна кількість пристроїв у використанні може бути вищою. Наприклад, за інформацією *Statista*, на початку 2021 р. у світі було близько 3,8 млрд активних користувачів смартфонів [2]. Смартфон – це обчислювальний пристрій, що більшу частину часу не є активним, і якщо залучити до обчислювальних задач навіть незначну кількість смартфонів, це могло б створити надзвичайно великий обсяг обчислювальної потужності без виробництва нових чипів.

Згідно з проєктом *green500* [3], найбільш енергоефективні 10 суперкомп'ютерів у світі мають енергетичну ефективність у 38–65 GFLOPS обчислювальної потужності на ват потужності. За офіційною інформацію, мобільний *SoC Apple A14 Bionic*, що використовується в *iPhone* 12, має номінальну обчислювальну потужність 1000 GFLOPS за умови номінального споживання 6 Вт, що оцінює ефективність у 167 GFLOPS/W. Звичайно, таким чином маємо загальну оцінку, яку не можна застосовувати для практичних задач, але це порушує питання про доцільність використання смартфонів як енергоефективних обчислювальних систем. Ця робота досліджує доцільність залучення для обчислення наявних мобільних обчислювальних пристроїв у контексті зменшення витрат енергії для обчислення. Метою статті є обґрунтування такої гіпотези: обчислення в хмарі на основі мобільних пристроїв використовує суттєво менше енергії, ніж обчислення на стаціонарному обладнанні.

Аналіз наявних робіт

Існує дуже обмежена кількість робіт щодо практичного розподіленого обчислення на смартфонах і побудови обчислювальних хмар на основі смартфонів. Наприклад, у праці [4] автори будують розподілену систему обчислення в хмарі. Ця система складається з серверного вебзастосунку, який керує розподілом задачі, та клієнтського браузерного вебзастосунку, що обчислює. Референсною задачею є навчання нейронної мережі. Обчислення виконується в середовищі браузера, що інтерпретує мову *JavaScript*. З одного боку, таке рішення є гнучким і не залежить від платформи, з іншого – ефективність обчислення засобами *JavaScript* дуже низька через особливості інтерпретації та неможливість доступу до засобів апаратного прискорення. Але це рішення доводить концептуальну можливість використовувати мобільні пристрої для обчислення.

Проєкт *World Community Grid* [5] є практичною реалізацією розподіленої системи обчислення із залученням смартфонів. Від застосовує більш розвинену систему керування на сервері: оркестрування, дублювання, гнучке планування. На клієнтському застосунку для обчислення використовується *OpenCL*, що дає змогу застосовувати деякі засоби апаратного прискорення. Цей проєкт не ставить за мету вивчення або максимізацію енергетичної ефективності, однак це один із розвинених представників у галузі.

Під час пандемії COVID-19 долучення персональних смартфонів для обчислення дослідницьких задач [6] зробило вагомий внесок у вивченні вірусу. Такі проєкти, як *BOINC* [7], *Folding@home* [8], *DreamLab, Rosetta@home* [9], активно долучали смартфони ентузіастів до своєї обчислювальної інфраструктури, щоб вивчити особливості вірусу та створити вакцини.

Також існує досить багато робіт щодо рендерингу графіки на смартфонах [10]. Але завдання рендерингу є вузькою та особливою. Для неї існує дуже розвинена інфраструктура, що передбачає засоби апаратного прискорення, програмні рішення тощо. Деякі проєкти (*DreamLab*) підтримують розподілений рендеринг на смартфонах. Але ця інфраструктура погано підходить для загальних задач обчислення.

Задача балансування навантаження [11] з урахуванням витрат на передачу даних нагальна не тільки в контексті хмар на основі смартфонів, але є їх невід'ємною частиною.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2023. No. 1 (23)

Задача обчислення на смартфоні спожитої енергії є також малодослідженою. У роботі [12] наводиться приклад евристичної моделі розподілу задач у гетерогенній хмарі. Ця модель бере до уваги як вартість обчислення в гетерогенній хмарі, так і вартість комунікації. Мета дослідження полягає в мінімізації часу виконання, але автори зауважують, що метод ефективний і для мінімізації енергетичних витрат.

В іншій роботі [13] описано підхід до обчислення споживання енергії в процесі обчислення в розподіленій системі та подано можливість прогнозування енергетичних витрат на іншій системі на основі даних попереднього виконання.

Ця стаття є продовженням і розвитком раніше опублікованих тез [14].

Опис методу дослідження

Завдання цієї роботи – довести гіпотези щодо переваг графічних смартфонів над стаціонарними графічними процесорами у сфері енергетичної ефективності та кількісне оцінювання цих переваг. В ідеальних умовах для надійного порівняння продуктивності стаціонарного апаратного забезпечення та мобільного обладнання необхідний доступ до всього обладнання та єдиного еталонного тесту, що може надійно вимірювати як продуктивність, так і енергоспоживання для різноманітних завдань. І якщо першу вимогу можна виконати орендою відповідного обладнання в хмарах, то другу – ні.

Мобільні та стаціонарні комп’ютери мають дуже різні завдання та способи використання. Апаратне забезпечення стаціонарних комп’ютерів розроблено для тривалої роботи в умовах пікового навантаження. Воно може тривалий час працювати під великим навантаженням без змін у продуктивності. Для цього стаціонарне обладнання забезпечено відносним потужним електричним живленням. Саме обладнання досить велике, має значну власну теплоємність та площу поверхні, що дозволяє встановлювати потужну та ефективну конвекційну систему охолодження або навіть рідинну систему охолодження. Це означає, що стаціонарний комп’ютер споживає стільки енергії, скільки потрібно для підтримки максимальної продуктивності, а ефективна система охолодження запобігає тротлінгу – процесу вимушеного зниження продуктивності процесора через перегрів. Навпаки, мобільні пристрої мають акумулятор обмеженої ємності порівняно з серверними/настільними джерелами живлення, а їх система охолодження пасивна й обмежена відтоком тепла через відносно малий корпус мобільного пристрою. Це призводить до відносно швидкого (декілька хвилин) перегріву пристрою.

Режим роботи мобільного пристрою потребує також іншої стратегії керування частотою та живленням мобільного процесора. На відміну від стаціонарного, він розрахований на короткі періоди навантаження. Навіть коли користувач упродовж тривалого часу користується браузером, навантаження також має форму короткотривалих піків: сторінка завантажується, обробляється, а потім навантаження знижується, поки не буде нової активності користувача. Це робить виправданим більш агресивний режим бусту частоти, поки процесор холодний, ніж на стаціонарному обладнанні. Стаціонарні процесори також застосовують можливість тимчасового саморозгону, поки вони холодні [15], але в мобільних процесорів цей процес значно агресивніший. Це призводить до швидкого перегріву мобільного процесора під навантаження та динамічного зниження частоти або тротлінгу. Отже, на відміну від стаціонарного, мобільний процесор за умови сталого навантаження має короткий період пікової продуктивності й потім довготривалий період зниженої на 15–25% продуктивності. Типова продуктивність мобільного процесора з плином часу показана на рис. 1.

Рис. 1. Продуктивність мобільного процесора під навантаженням

Така різниця в роботі мобільних і стаціонарних процесорів робить складним і водночас непотрібним створення тестів, що порівнювали б їх енергетичну ефективність під тривалими великими сталими навантаженнями. Навіть використання 3D-бенчмарків ускладнене тим, що їх мобільні версії адаптовані для мобільних пристроїв. Тому порівнювати результати тесту на основі однієї бібліотеки для мобільного й стаціонарного процесора не має сенсу:

Сучасний стан наукових досліджень та технологій в промисловості. 2023. №1 (23)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

обчислювальні задачі фактично різні. Аналізуючи публічні бенчмарки, ми прийшли до висновку, що тести для стаціонарних комп'ютерів майже повністю ізольовані та не стосуються мобільних. Тобто бенчмарки для стаціонарних комп'ютерів дуже репрезентативні для порівняння стаціонарних рішень між собою, а мобільні – для порівняння мобільних рішень між собою. Кожна група тестів адаптувалася саме для цих задач, і тому надійних публічних результатів порівняння продуктивності стаціонарних і мобільних процесорів не існує.

Тому ми зробили іншу стратегію порівняння на основі різних тестів. Для кожної групи (стаціонарні та мобільні графічні процесори) ми знаходимо бенчмарки, що досягають максимального навантаження впродовж часу та вимірюють в абсолютних одиницях. Також для кожної групи обираємо бенчмарки, що досягають максимального споживання енергії впродовж тривалого часу. Для кожної моделі графічного процесора обираємо 2–3 бенчмарки з максимальною продуктивністю та беремо медіану, також обираємо 2–3 бенчмарки з максимальним споживанням та беремо медіану. Цей підхід не можна розглядати як метод оцінювання енергетичної ефективності конкретної моделі графічного процесора, але метод оцінювання однаковий, і для вагомої вибірки він даватиме приблизне розуміння енергоефективності стаціонарних і мобільних графічних процесорів.

Усі задачі поділимо на три групи: задачі на обчислення з рухомою точкою з одинарною точністю (*single precision*), половинною точністю (*half precision*) та подвійною точністю (*double precision*). Референсними графічними процесорами оберемо пристрої різних виробників (*NVidia GeForce, AMD Radeon, Adreno, Mali, Apple Buinic*) одного й того самого періоду (2019–2021). Результати виміру наведені в табл. 1 і 2.

Таблиця 1. *Енергетична ефективність стаціонарних графічних процесорів*

	Half Precision GFLOPS/W	**Single Precision GFLOPS/W**	**Double Precision GFLOPS/W**
GeForce RTX 2050 (M)	268.889	134.444	4.200
GeForce RTX 2060 Max-Q (M)	140.015	70.000	2.185
GeForce RTX 2060 (M)	102.400	51.200	1.600
GeForce RTX 2070 Max-Q (M)	136.500	68.250	2.138
GeForce RTX 2070 (M)	115.391	57.704	1.800
GeForce RTX 2070 Super Max-Q (M)	138.250	69.125	2.163
GeForce RTX 2070 Super (M)	122.870	61.443	1.922
GeForce RTX 2080 Max-Q (M)	161.125	80.588	2.525
GeForce RTX 2080 (M)	124.800	62.413	1.953
GeForce RTX 2080 Super Max-Q (M)	165.875	82.950	2.588
GeForce RTX 2080 Super (M)	127.800	63.900	2.000
GeForce RTX 2060	80.638	40.319	1.263
GeForce RTX 2060 (12 GB)	77.632	38.816	1.211
GeForce RTX 2060 Super	82.069	41.034	1.280
GeForce RTX 2070	85.314	42.657	1.331
GeForce RTX 2070 Super	84.302	42.149	1.316
GeForce RTX 2080	93.660	46.828	1.465
GeForce RTX 2080 Super	89.212	44.604	1.396
GeForce RTX 2080 Ti	96.057	48.029	1.504
GeForce RTX 3050 Laptop	66.600	66.600	1.031
GeForce RTX 3050 Ti Laptop	81.523	81.523	1.262
GeForce RTX 3060 Laptop	136.750	136.750	2.138
GeForce RTX 3070 Laptop	152.095	152.095	2.371
Geforce RTX 3070 Ti Laptop	139.238	139.238	2.171
GeForce RTX 3080 Laptop	165.043	165.043	2.574
Geforce RTX 3080 Ti Laptop	162.696	162.696	2.539
GeForce RTX 3050	69.985	69.985	1.092
GeForce RTX 3060	74.929	74.929	1.171
GeForce RTX 3060 Ti	80.985	80.985	1.265
GeForce RTX 3070	92.336	92.336	1.441

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) *Innovative technologies and scientific solutions for industries. 2023. No. 1 (23)*

Кінець табл. 1

	Half Precision GFLOPS/W	Single Precision GFLOPS/W	Double Precision GFLOPS/W
GeForce RTX 3070 Ti	74.828	74.828	1.172
GeForce RTX 3080	93.025	93.025	1.453
GeForce RTX 3080 (12 GB)	87.551	87.551	1.366
GeForce RTX 3080 Ti	97.429	97.429	1.523
GeForce RTX 3090	101.660	101.660	1.589
GeForce RTX 3090 Ti	88.882	88.882	1.389
Radeon RX 6400 (Navi 24 XL)	134.528	67.264	4.204
Radeon RX 6500 XT (Navi 24 XT)	102.035	51.018	3.188
Radeon RX 6600 (Navi 23 XL)	135.303	67.636	4.227
Radeon RX 6600 XT (Navi 23 XT)	125.497	62.751	3.922
Radeon RX 6650 XT (Navi 23 KXT)	119.922	59.961	3.748
Radeon RX 6700 (Navi 22 XL)	129.023	64.514	4.032
Radeon RX 6700 XT (Navi 22 XT)	114.909	57.457	3.591
Radeon RX 6750 XT (Navi 22 KXT)	106.496	53.248	3.328
Radeon RX 6800 (Navi 21 XL)	129.332	64.664	4.040
Radeon RX 6800 XT (Navi 21 XT)	138.240	69.120	4.320
Radeon RX 6900 XT (Navi 21 XTX)	153.600	76.800	4.800
Radeon RX 6950 XT (Navi 21 KXTX)	141.221	70.609	4.412
Radeon RX 6300M (Navi 24 XML)	250.800	125.200	7.824
Radeon RX 6500M (Navi 24 XM)	199.400	99.600	6.224
Radeon RX 6600S (Navi 23)	195.375	97.625	5.601
Radeon RX 6700S (Navi 23)	158.400	79.200	4.950
Radeon RX 6600M (Navi 23)	175.500	78.000	4.875
Radeon RX 6650M (Navi 23)	147.500	73.750	4.609
Radeon RX 6800S (Navi 23)	184.300	92.200	5.765
Radeon RX 6650M XT (Navi 23)	166.167	83.083	5.193
Radeon RX 6700M (Navi 22 XM)	157.104	78.556	4.904
Radeon RX 6800M (Navi 22 XTM)	162.414	81.241	5.077
Radeon RX 6850M XT (Navi 22 XTM)	160.182	80.061	5.004
MIN	66.600	38.816	1.031
MAX	268.889	165.043	7.824
AVG	127.892	78.738	2.902
MEDIAN	127.800	73.750	2.185

Таблиця 2. *Енергетична ефективність стаціонарних графічних процесорів*

	Half Precision GFLOPS/W	Single Precision GFLOPS/W	Double Precision GFLOPS/W
Adreno 610	182.00	91.00	22.67
Adreno 640	265.85	132.92	33.23
Adreno 650	347.38	173.68	43.42
Adreno 660	382.14	191.07	47.76
Adreno 680	292.57	146.29	36.57
Adreno 730	570.06	285.03	71.26
Mali-G57 MP2	384.00	192.00	96.00
Mali-G78 MP20	396.53	198.27	49.59
Apple A12	126.79	63.39	15.87
Apple A15	416.67	208.33	52.08
MIN	126.79	63.39	15.87
MAX	570.06	285.03	96.00
AVG	336.40	168.20	46.85
MEDIAN	364.76	182.38	45.59

Сучасний стан наукових досліджень та технологій в промисловості. 2023. №1 (23)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

Нехай E – енергетична ефективність; E_{HP}, E_{SP}, E_{DP} – енергетична ефективність обчислення з половинною, одинарною та подвійною точністю; E_{MIN}, E_{MAX}, E_{AVG} – мінімальна, максимальна й середня енергетичні ефективність; E^M, E^S – енергетична ефективність мобільних і стаціонарних графічних процесорів відповідно. Тоді зростання ефективності B обчислюється за формулами:

$B_{HP,min} = E^M_{HP,min} / E^S_{HP,min}$ $B_{HP,max} = E^M_{HP,max} / E^S_{HP,max}$ $B_{HP,avg} = E^M_{HP,avg} / E^S_{HP,avg}$	$B_{SP,min} = E^M_{SP,min} / E^S_{SP,min}$ $B_{SP,max} = E^M_{SP,max} / E^S_{SP,max}$ $B_{SP,avg} = E^M_{SP,avg} / E^S_{SP,avg}$	$B_{DP,min} = E^M_{DP,min} / E^S_{DP,min}$ $B_{DP,max} = E^M_{DP,max} / E^S_{DP,max}$ $B_{DP,avg} = E^M_{DP,avg} / E^S_{DP,avg}$

Отже, мобільний графічний процесор у середньому витрачає в 2,1 раза менше енергії на обчислення з одинарною точністю, в 2,6 раза менше за умови обчислення з половинною точністю та в 16,2 раза менше, якщо обчислення здійснюється з подвійною точністю. Результат можна побачити на рис. 2.

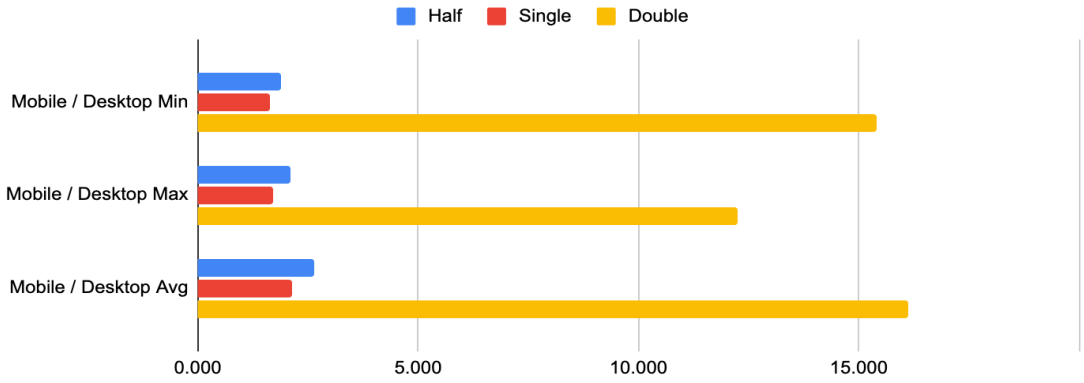

Рис. 2. Зростання ефективності мобільних процесорів порівняно зі стаціонарними

Відрив у 2,6 раза в обчисленнях із половинною точністю може бути трохи завищеним, бо для багатьох стаціонарних графічних процесорів ефективність такого обчислення майже не відрізняється від обчислення з одинарною точністю. Це може бути наслідком особливостей самого тесту. У графічних процесорів *NVidia* є два основних типи операції з половинною точністю: одинарна й пакетна [16]. У першому випадку в одній інструкції – одна операція з половинною точністю, в другому випадку – дві операції в одній інструкції. Тривалість виконання операції з половинною та одинарною точністю у графічних процесорів *NVidia* однакова (у тактах), тому одинарні операції половинної точності мають таку саму або трохи вищу швидкість $(\sim 5\%)$ через зменшення навантаження на шину даних. Пакетні або упаковані операції ефективніші, і за один і той самий час графічний процесор може виконувати більше таких операцій. Ми не можемо знати деталі реалізації деяких бенчмарків, але однакова ефективність виконання операції з одинарною та половинною точністю швидше за все зумовлена реалізацією одинарних операцій з половинною точністю.

Різниця між ефективністю обчислення з подвійною точністю пояснюється тим, що стаціонарні графічні процесори традиційно мають значно меншу ефективність обчислень із подвійною точністю. Це є наслідком того, що подвійна точність нечасто потрібна в комп'ютерній графіці, фізиці, кодуванні відео тощо. Усе інше добре виконує центральний процесор за допомогою розширених векторних інструкцій. У мобільних пристроях центральний і графічний процесор мають доступ до спільної пам'яті, тому високоефективне оброблення даних із подвійною точністю може бути дуже легко перенесене на графічний процесор, щоб таким чином спростити центральний процесор.

Експериментальне дослідження

Оцінювання ефективності на основі публічних джерел і бенчмарків може якісно визначити стан

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2023. No. 1 (23)

речей, але не підходить для кількісного оцінювання. Вхідні дані не повні та засмічені: ми не виконуємо одні й ті самі задачі на мобільному й стаціонарному процесорі, навіть не знаємо деталей реалізації. Тому для вимірювання та порівняння ефективності зробимо власний бенчмарк.

Для обох платформ (мобільної та стаціонарної) реалізуємо синтетичний бенчмарк. Референсною задачею є помноження матриць великої розмірності з додатковими операціями, що не змінюють результат, але роблять задачу більш інтенсивною на обчислення, а не на пам'ять.

Мобільними платформами оберемо *iPhone 12 mini* 2020 р. Реалізація референсної задачі виконана за допомогою *Metal*, бо *Metal* є найкращим інструментом для використання апаратного прискорення в інфраструктурі *Apple* [17]. Вимірювання споживання енергії напряму є неможливим, тому ми повністю зарядимо батарею, поставимо на виконання задачу та дочекаємось відключення мобільного пристрою через розрядку акумулятора. Ємність і стан акумулятора – відомі, тому можна розділити загальну кількість операцій на витрачену енергію акумулятора, і це будуть ті самі GFLOPS/W. Ядро обчислення *Metal* наведене у фрагменті 1.

Фрагмент коду 1. Ядро обчислення *Metal*

```
kernel void ck_mat_prod_s(device float* inputA                [[buffer(PTMatProdCKIndex_InputA)]],
                          device float* inputB                [[buffer(PTMatProdCKIndex_InputB)]],
                          device float* output                [[buffer(PTMatProdCKIndex_Output)]],
                          constant PTMatProdCKUniforms &uniforms    [[buffer(PTMatProdCKIndex_Uniforms)]],
                          uint2 gid                           [[thread_position_in_grid]])
{
  float sum = 0.0;
  int dim = uniforms.size.x;

  device const float4* lineA = (device const float4*)(inputA + gid.x * uniforms.size.x);
  device const float4* lineB = (device const float4*)(inputB + gid.y * uniforms.size.x);

  int lineLength = dim / 4;

  const int kPayloadTimes = 32;

  for (int i = 0; i < lineLength; ++i)
  {
    float4 va = lineA[i];
    for (int j = 0; j < kPayloadTimes; ++j)
    {
      va = va * uniforms.am;
      va = va * uniforms.im;
    }
    float4 vb = lineB[i];
    for (int j = 0; j < kPayloadTimes; ++j)
    {
      vb = vb * uniforms.am;
      vb = vb * uniforms.im;
    }
    sum += dot(va, vb);
  }

  output[gid.x + gid.y * dim] = sum;
}
```

Штучне додаткове навантаження здійснюється за допомогою додаткових 32 циклів по 4 множення матриці на вектор (вектором є кожні чотири елементи рядка та стовпця матриць, які ми помножуємо). На результат це не впливає, бо матриці am та im є зворотними одна до одної, але такий спосіб запобігає оптимізації зайвого коду компілятором.

Кількість корисних операцій обчислюється таким чином. Кожний елемент добутку матриць є скалярним добутком рядка першої матриці на стовпець другої. Тому чистий скалярний добуток мав би складність $N+N-1$, де N – довжина рядка/стовпця. Додаткове баластне навантаження додає по 4×32 матричних добутки на кожні 4 елементи скалярного добутку. Отже, можемо стверджувати, що на кожен елемент N відбувається додатково 32 добутки матриці на вектор. Добуток матриці 4×4 на вектор з 4 елементів займає 28 операцій. Тому кінцева складність обчислення кожного члена добутку матриць буде $N+N-1+28N$, нехай $30N$ для спрощення.

Щоб було простіше, помножимо матриці розміром 1024×1024, тобто $N=1024$. Загалом у добутку буде $N\times N$ членів, кожний з яких має

Сучасний стан наукових досліджень та технологій в промисловості. 2023. №1 (23)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

складність $30N$ операцій. Кінцева складність обчислення матриці 1024×1024 – 32,211 млрд операцій із рухомою точкою. Кількість сервісних операцій не є суттєвою й значно менша ніж 1%.

Ємність батареї *iPhone 12 mini* – 8,57 Вт/год або 30852 Дж. Стан батареї – 84%, тобто можемо вважати, що ємність батареї – $P_M = 25915$ Дж. Нехай кількість обчислень матриць до повного розрядження – $K = 131279$. Тоді енергетичну ефективність обчислень можна визначити за формулою:

$$E_M = \frac{30N^3K}{P_M} = 163.179\cdot10^9 \text{ GFLOPS/W}.$$

Відповідно до табл. 2 значення E_M розташоване між піковою продуктивністю A12 (63.39 GFLOPS/W) та A15 (208.33 GFLOPS/W), ближче до A15, що є очікуваним.

Як стаціонарну платформу використаємо *GeForce RTX 3080* – відеокарта також 2020 р., тобто є ровесницею *iPhone 12 mini*. Референсною задачею застосовано такий самий добуток матриць, але реалізований за допомогою CUDA. Код ядра CUDA наведено у фрагменті коду 2.

Фрагмент коду 2. Ядро обчислення CUDA

```
__global__ void test_cuda_routine_kernel(const float* a, const float* b, float* c,
                                         const float* pam, const float* pim, int count, int dim)
{
   int index = blockIdx.x * blockDim.x + threadIdx.x;
   if (index >= count)
      return;

   int x = index % dim;
   int y = index / dim;

   float sum = 0.0f;
   const glm::vec4* va = (const glm::vec4*)(a + x * dim);
   const glm::vec4* vb = (const glm::vec4*)(b + y * dim);

   const glm::mat4* am = (glm::mat4*)pam;
   const glm::mat4* im = (glm::mat4*)pim;

   const int kPayloadTimes = 32;

   for (int i = 0; i < dim / 4; ++i)
   {
      glm::vec4 ea = va[i];
      for (int j = 0; j < kPayloadTimes; ++j)
      {
         ea = ea * (*am);
         ea = ea * (*im);
      }
      glm::vec4 eb = vb[i];
      for (int j = 0; j < kPayloadTimes; ++j)
      {
         eb = eb * (*am);
         eb = eb * (*im);
      }
      sum += glm::dot(ea, eb);
   }

   c[index] = sum;
}
```

Для вимірювання споживання енергії використовуємо застосунок GPU-Z. Він може знімати показники споживання енергії та зберігати у файл. Для вимірювання енергетичної ефективності знадобиться час виконання $T = 300$ с, кількість виконаних добутків матриць $K = 143994$ та середнє споживання енергії відеокартою $P_S = 182$ Вт (P_S обчислюється як різниця середнього споживання під час виконання тесту та споживання відеокарти в режимі спокою). Тоді енергетичну ефективність можна обчислити за формулою:

$$E_S = \frac{30N^3K}{P_ST} = 84.951\cdot10^9 \text{ GFLOPS/W}.$$

Знайдене значення продуктивності (84,951 GFLOPS/W) досить близьке до відповідного значення в табл. 1 (93,025 GFLOPS/W). Зростання

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2023. No. 1 (23)

ефективності обчислення дорівнює $B = EM/ES = 1{,}92$. Тобто на обчислення на мобільному пристрої було витрачено в 1,92 раза менше енергії за операцію, ніж на стаціонарному пристрої.

Зазначений метод враховує лише продуктивність графічного процесора та ігнорує фонове споживання системи, а також споживання інших складників системи. Стаціонарний комп'ютер споживає набагато більше, ніж відеокарта, тоді як споживання смартфона майже цілком складається зі споживання графічного процесора. Незважаючи на те, що ми не можемо кількісно оцінити загальне споживання в межах цього тесту, якісно це також необхідно оцінити на користь смартфона. Але в реальних задачах із розподіленого обчислення також суттєву частину енергії витрачатиме система мобільної передачі даних або підключення до локальної мережі. Натомість стаціонарний комп'ютер, особливо в датацентрі, має дротове підключення до мережі, і передача ним даних впливає на споживання енергії не так суттєво. Урахування та моделювання споживання всієї обчислювальної системи під час виконання конкретних обчислювальних задач є темою наступних наукових робіт, зокрема дисертації.

Висновки

У статті необхідно було оцінити, наскільки легітимною є гіпотеза, що використання мобільного пристрою для обчислення в хмарі вимагає суттєво менших витрат енергії, ніж таке саме обчислення на стаціонарному обладнанні, варіанти якого є основною частиною серверного обладнання у звичайному хмарному датацентрі. Для оцінювання застосовано публічні джерела й експеримент. Із джерел опрацьовано інформацію про продуктивність і споживання енергії поширеними моделями мобільних і стаціонарних графічних процесорів. Визначено, що мобільні графічні процесори мають суттєві переваги в енергоефективності, на відміну від стаціонарних, особливо в обчисленні з подвійною точністю. Але такий підхід не можна застосувати для обчислення кількісної оцінки переваги.

Проведено експеримент з обчислення енергоефективності на основі однієї і тієї самої синтетичної обчислювальної задачі. За результатом дослідження, мобільний графічний пристрій має також суттєву перевагу над стаціонарним за кількістю енергії за обчислювальну операцію. Це означає, що гіпотеза щодо переваг використання мобільних графічних процесорів для обчислення в хмарі, порівняно зі стаціонарним обладнанням, підтверджена й потребує подальших досліджень. Необхідно зауважити, що в статті поки йдеться про ефективність лише обчислювального складника, без урахування складника з передачі даних.

Список літератури

1. Smartphone Market Share, Naliba Popal and Ryan Reith, Aug 2022. URL: https://www.idc.com/promo/smartphone-market-share/vendor
2. Number of smartphone subscriptions worldwide from 2016 to 2021, with forecasts from 2022 to 2027, *Published by Petroc Taylor*, Jan 18, 2023. URL: https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/
3. Green500 – *The 20th Green List was published Nov 15, 2022*, Dallas, TX. URL: https://www.top500.org/lists/green500/
4. Hamza moh. Salem, Distributed Computing System on a Smartphones-Based Network, *In book: Software Technology: Methods and Tools* (pp. 313–325), DOI:10.1007/978-3-030-29852-4_26, Oct 2019. DOI: https://www.doi.org/10.1007/978-3-030-29852-4_26
5. How World Community Grid Works, Sep 2021. URL: https://www.worldcommunitygrid.org/about/how.s
6. Manuel Delfino, Distributed Computing, *In book: Particle Physics Reference Library, Volume 2: Detectors for Particles and Radiation* (pp. 613–644), Sep 2020. DOI: https://www.doi.org/10.1007/978-3-030-35318-6_14
7. Berkeley Open Infrastructure for Network Computing Retrospect, *published by David P. Anderson*, 2021. URL: https://continuum-hypothesis.com/boinc_history.php
8. COVID-19 – What I Can Do, 2021. URL: https://foldingathome.org/diseases/infectious-diseases/covid-19/
9. "From the whole team at @UWproteindesign, THANK YOU!" *posted by Rosetta@Home on Twitter, June 2021*. URL: https://twitter.com/RosettaAtHome/status/1408533111793586178
10. Woong Seo, Sanghun Park, Insung Ihm, Efficient Ray Tracing of Large 3D Scenes for Mobile Distributed Computing Environments, *Published online, 2022 Jan 10, Department of Computer Science and Engineering, Sogang University*, Seoul 04107, Korea. DOI: https://www.doi.org/10.3390/s22020491

Сучасний стан наукових досліджень та технологій в промисловості. 2023. №1 (23)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

11. Himanshu Rai, Sanjeev Kumar Ojha, Alexey Nazarov, Cloud Load Balancing Algorithm, *Conference: 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, Dec 2020. DOI: https://www.doi.org/10.1109/ICACCCN51052.2020.9362810
12. A. Abdelmageed Elsadek, B. Earl Wells, Heuristic Model for Task Allocation in a Heterogeneous Distributed Computing System, *Conference: Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications, PDPTA 1996*, August 9–11, 1996, Sunnyvale, California, USA. URL: https://www.researchgate.net/publication/221132962_Heuristic_Model_for_Task_Allocation_in_a_Heterogeneous_Distributed_Computing_System
13. René Caspart, Sebastian Ziegler, Arvid Weyrauch, Precise Energy Consumption Measurements of Heterogeneous Artificial Intelligence Workloads, *Published online, 2022*. DOI: https://www.doi.org/10.48550/arXiv.2212.01698
14. Olexander Mamchych, Maksym Volk, Smartphone Based Computing Cloud and Energy Efficiency, 2022 *12th International Conference on Dependable Systems, Services and Technologies*, Athens, Greece. DOI: https://www.doi.org/10.1109/DESSERT58054.2022.10018740
15. How Intel Technologies Boost Your CPU's Performance, Thermal Velocity Boost, 2022, available at: https://www.intel.com/content/www/us/en/gaming/resources/how-intel-technologies-boost-cpu-performance.html
16. N.M Ho and W.F. Wong, Exploiting half precision arithmetic in Nvidia GPUs, *IEEE ICIAfS 2016, Department of Computer Science, National University of Singapore*, Singapore. DOI: https://doi.org/10.1109/HPEC.2017.8091072
17. Lars Gebraad, Andreas Fichtner, Seamless GPU Acceleration for C++-Based Physics with the Metal Shading Language on Apple's M Series Unified Chips, *Seismological Research Letters*, Feb 2023, USA. DOI: https://www.doi.org/10.1785/0220220241

References

1. Smartphone Market Share, Naliba Popal and Ryan Reith, Aug 2022. URL: https://www.idc.com/promo/smartphone-market-share/vendor
2. Number of smartphone subscriptions worldwide from 2016 to 2021, with forecasts from 2022 to 2027, *Published by Petroc Taylor*, Jan 18, 2023. URL: https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/
3. Green500 – *The 20th Green List was published Nov 15*, 2022, Dallas, TX. URL: https://www.top500.org/lists/green500/
4. Hamza moh. Salem, Distributed Computing System on a Smartphones-Based Network, *In book: Software Technology: Methods and Tools* (pp.313–325), DOI:10.1007/978-3-030-29852-4_26, Oct 2019. DOI: https://www.doi.org/10.1007/978-3-030-29852-4_26
5. How World Community Grid Works, Sep 2021. URL: https://www.worldcommunitygrid.org/about/how.s
6. Manuel Delfino, Distributed Computing, *In book: Particle Physics Reference Library, Volume 2: Detectors for Particles and Radiation* (pp. 613–644), Sep 2020. DOI: https://www.doi.org/10.1007/978-3-030-35318-6_14
7. Berkeley Open Infrastructure for Network Computing Retrospect, *published by David P. Anderson*, 2021. URL: https://continuum-hypothesis.com/boinc_history.php
8. COVID-19 - What I Can Do, 2021. URL: https://foldingathome.org/diseases/infectious-diseases/covid-19/
9. "From the whole team at @UWproteindesign, THANK YOU!" *posted by Rosetta@Home on Twitter*, June 2021. URL: https://twitter.com/RosettaAtHome/status/1408533111793586178
10. Woong Seo, Sanghun Park, Insung Ihm, Efficient Ray Tracing of Large 3D Scenes for Mobile Distributed Computing Environments, *Published online, 2022 Jan 10, Department of Computer Science and Engineering, Sogang University*, Seoul 04107, Korea. DOI: https://www.doi.org/10.3390/s22020491
11. Himanshu Rai, Sanjeev Kumar Ojha, Alexey Nazarov, Cloud Load Balancing Algorithm, *Conference: 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, Dec 2020. DOI: https://www.doi.org/10.1109/ICACCCN51052.2020.9362810
12. A. Abdelmageed Elsadek, B. Earl Wells, Heuristic Model for Task Allocation in a Heterogeneous Distributed Computing System, *Conference: Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications, PDPTA 1996, August 9–11, 1996,* Sunnyvale, California, USA. URL: https://www.researchgate.net/publication/221132962_Heuristic_Model_for_Task_Allocation_in_a_Heterogeneous_Distributed_Computing_System
13. René Caspart, Sebastian Ziegler, Arvid Weyrauch, Precise Energy Consumption Measurements of Heterogeneous Artificial Intelligence Workloads, *Published online*, 2022. DOI: https://www.doi.org/10.48550/arXiv.2212.01698
14. Olexander Mamchych, Maksym Volk, Smartphone Based Computing Cloud and Energy Efficiency, 2022 *12th International Conference on Dependable Systems, Services and Technologies*, Athens, Greece. DOI: https://www.doi.org/10.1109/DESSERT58054.2022.10018740
15. How Intel Technologies Boost Your CPU's Performance, Thermal Velocity Boost, 2022, available at: https://www.intel.com/content/www/us/en/gaming/resources/how-intel-technologies-boost-cpu-performance.html

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) *Innovative technologies and scientific solutions for industries. 2023. No. 1 (23)*

16. N. M Ho and W. F. Wong, Exploiting half precision arithmetic in Nvidia GPUs, *IEEE ICIAfS 2016, Department of Computer Science, National University of Singapore,* Singapore. DOI: https://doi.org/10.1109/HPEC.2017.8091072
17. Lars Gebraad, Andreas Fichtner, Seamless GPU Acceleration for C++-Based Physics with the Metal Shading Language on Apple's M Series Unified Chips, *Seismological Research Letters,* Feb 2023, USA. DOI: https://www.doi.org/10.1785/0220220241

Received 26.03.2023

Відомості про авторів / About the Authors

Мамчич Олександр Олександрович – Харківський національний університет радіоелектроніки, аспірант кафедри електронних обчислювальних машин, Харків, Україна; e-mail: oleksandr.mamchych@nure.ua; ORCID: https://orcid.org/my-orcid?orcid=0009-0001-6602-2929

Волк Максим Олександрович – доктор технічних наук, Харківський національний університет радіоелектроніки, професор кафедри електронних обчислювальних машин, Харків, Україна; e-mail: maksym.volk@nure.ua; ORCID: https://orcid.org/0000-0003-4229-9904

Mamchych Oleksand – Kharkiv National University of Radio Electronics, Postgraduate Student of the Department of Computing Machines, Kharkiv, Ukraine.

Volk Maksym – Doctor of Technical Sciences, Phd (Computer engineering), Kharkiv National University of Radio Electronics, Professor at Department of Computing Machines, Kharkiv, Ukraine.

ESTIMATION OF POWER CONSUMPTION OF MOBILE DEVICES IN CLOUD COMPUTING

Modern computing tasks require an increase in computing power. This necessitates the creation and production of new equipment for cloud computing. At the same time, the number of personal mobile devices is already measured in billions, and even their partial use could reduce production requirements. In addition, mobile hardware is more energy efficient, which contributes to significant energy savings. The article investigates the issue of qualitative and quantitative assessment of the efficiency of using mobile devices for computing compared to traditional stationary solutions. **The purpose of the work** is to substantiate the following hypothesis: computing in the cloud based on mobile devices significantly reduces energy consumption than computing on stationary equipment. For this purpose, we show that computing on a mobile GPU is more energy efficient than computing on a stationary processor. Public sources and benchmarks were analyzed to determine the qualitative advantage. On the basis of the studied data, efficiency indicators for various mobile and desktop GPUs are calculated. It is argued that in most cases, mobile solutions consume significantly less energy compared to desktop solutions. To calculate the quantitative advantage, an experiment was conducted on the basis of two platforms: mobile and desktop. The same computational task was implemented using Apple Metal and NVidia CUDA. Based on this task, the energy efficiency indicators of the mobile and stationary graphic professor were calculated. According to the results of the study, a significant advantage of the mobile GPU in terms of energy efficiency has been determined. This result is relevant because the platforms were released in the same year with a difference of several months, so they can be considered peers of each other. The approaches presented here do not take into account the consumption of all other parts of the system, except for the GPUs. This means that the consumption of the motherboard, power supply, etc. can tilt the balance in favor of the mobile processor even more. But for distributed computing, the network connection is very important, and it can consume a significant amount of power on a mobile device. Further research will focus on a more comprehensive accounting of the energy consumption of various computer subsystems.

Keywords: computing cloud; smartphone cloud; energy efficiency; cloud computing; CUDA; Metal; hardware acceleration.

Бібліографічні описи / Bibliographic descriptions

Мамчич О. О, Волк М. О. Оцінювання енергетичних витрат у процесі використання мобільних пристроїв для хмарних обчислень. *Сучасний стан наукових досліджень та технологій в промисловості*. 2023. № 1 (23). С. 72–82. DOI: https://doi.org/10.30837/ITSSI.2023.23.072

Mamchych, O, Volk, M. (2023), "Estimation of power consumption of mobile devices in cloud computing", *Innovative Technologies and Scientific Solutions for Industries*, No. 1 (23), P. 72–82. DOI: https://doi.org/10.30837/ITSSI.2023.23.072