МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ВІННИЦЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ

Інформаційна технологія паралельно-ієрархічного оброблення зображень у гетерогенному програмно-апаратному забезпеченні

Доповідач:
Кулик Олександр Олександрович, 1КН-15мн
Науковий керівник:
д. т. н., професор Яровий Андрій Анатолійович

- **Мета дослідження**: підвищення швидкодії процесу паралельно-ієрархічної обробки плямоподібних зображень на основі гетерогенного програмно-апаратного забезпечення.
- **Об'єкт дослідження:** процес паралельного оброблення плямоподібних зображень.
- **Предмет дослідження:** технології паралельно-ієрархічного оброблення зображень у гетерогенному програмно-апаратному середовищі.

Задача дослідження

- 1) здійснити аналіз предметної області високопродуктивних паралельних обчислень
- 2) здійснити класифікацію методів оброблення зображень;
- 3) здійснити аналіз математичних моделей паралельно-ієрархічного оброблення зображень
- 4) здійснити комп’ютерне моделювання процесу паралельно-ієрархічного оброблення зображень в гетерогенному програмно-апаратному середовищі;
- 5) розробити структурну організацію та здійснити програмну реалізацію паралельно-ієрархічного оброблення зображень у гетерогенному програмно-апаратному забезпеченні.

Наукова новизна

- Удосконалено моделі паралельно-ієрархічного перетворення зображень шляхом комбінованого використання технологій NVIDIA CUDA та OpenMP, що забезпечило підвищення швидкодії обробки плямоподібних зображень великої роздільної здатності.
- Удосконалено моделі організації високопродуктивного обчислювального процесу динамічного оброблення плямоподібних зображень у гетерогенній платформі, що забезпечило підвищення швидкодії обробки зображень до 78% при застосуванні двох GPU-пристроїв та до 173% при застосуванні чотирьох GPU-пристроїв порівняно із реалізацією на одному GPU- пристрої.
- Розроблено моделі та структурну організацію високопродуктивної паралельно-ієрархічної системи на основі використання технології Multi-GPU Programming, що дозволило здійснювати оброблення та класифікацію зображень великої роздільної здатності з підвищеною продуктивністю.

Практичне значення отриманих результатів

- Розроблена інформаційна технологія дозволяє здійснювати оброблення та класифікацію плямоподібних зображень великої роздільної здатності з підвищеною швидкодією.
- Удосконалено алгоритми ПІ перетворення для комбінованого використання технологій NVIDIA CUDA, OpenMP та Multi-GPU Programming у гетерогенній платформі, які забезпечують підвищену швидкодію обробки зображень.

Актуальність

- Останнім часом спостерігається тенденція до зростання потреби у проведенні високопродуктивних обчислень, аналізу та обробки великих обсягів інформації (задачі прогнозування погоди та змін клімату, моделювання складних фізичних процесів, розпізнавання зображень, синтез мови людини тощо).
- Одними з найбільш перспективних технологій для проведення високопродуктивних обчислень є технології паралельних обчислень.
- Це в свою чергу приводить до підвищення актуальності методів, що здатні використовувати усі переваги концепції паралелізму.
- Паралельно-ієрархічне перетворення належить до таких методів.

Класифікація методів розпізнавання зображень

Паралельно-ієрархічне перетворення

- ПІ перетворення – це принцип паралельного оброблення інформації, метою якого є досягнення максимально можливої алгоритмічної та схемотехнічної швидкодії.
- ПІ перетворення застосовується для виділення характерних ознак зображень, їх кодування, класифікації та скорочення розмірності при виконанні обчислень.
- Додатковою перевагою ПІ перетворення є його орієнтованість на паралельну реалізацію.

Математична модель прямого паралельно-ієрархічного перетворення

$$\Phi_{t=2}^{k}\left[\mathrm{T}\left(G\left(\bigcup_{S=1}^{S}\left(\bigcup_{i=1}^{n}a_i\right)\right)\right)\right]=\bigcup_{t=2}^{k}a_{11}^{t}$$

- T – операція транспонування, G – операція G-перетворення, S – операція зсуву.

Математична модель паралельно-ієрархічного оброблення зображень на основі нормуючого рівняння

$$d = \frac{\left(a_1\right)^{j+1}}{\left(\sum_{t=2}^{k} a_{11}^{t}\right)^{j}} + \frac{\left(a_2\right)^{j+1}}{\left(\sum_{t=2}^{k} a_{11}^{t}\right)^{j}} + \cdots + \frac{\left(a_{N-1}\right)^{j+1}}{\left(\sum_{t=2}^{k} a_{11}^{t}\right)^{j}} + \frac{\left(a_N\right)^{j+1}}{\left(\sum_{t=2}^{k} a_{11}^{t}\right)^{j}}.$$

- d – коефіцієнт нормуючого рівняння, a_{ij} – значення хвостових елементів зображення.

Паралелізм

- Під паралелізмом, як правило, розуміють концепцію обчислень, згідно з якою обрана задача розбивається на підзадачі, які виконуються одночасно. Паралелізм може бути реалізований у різних формах:

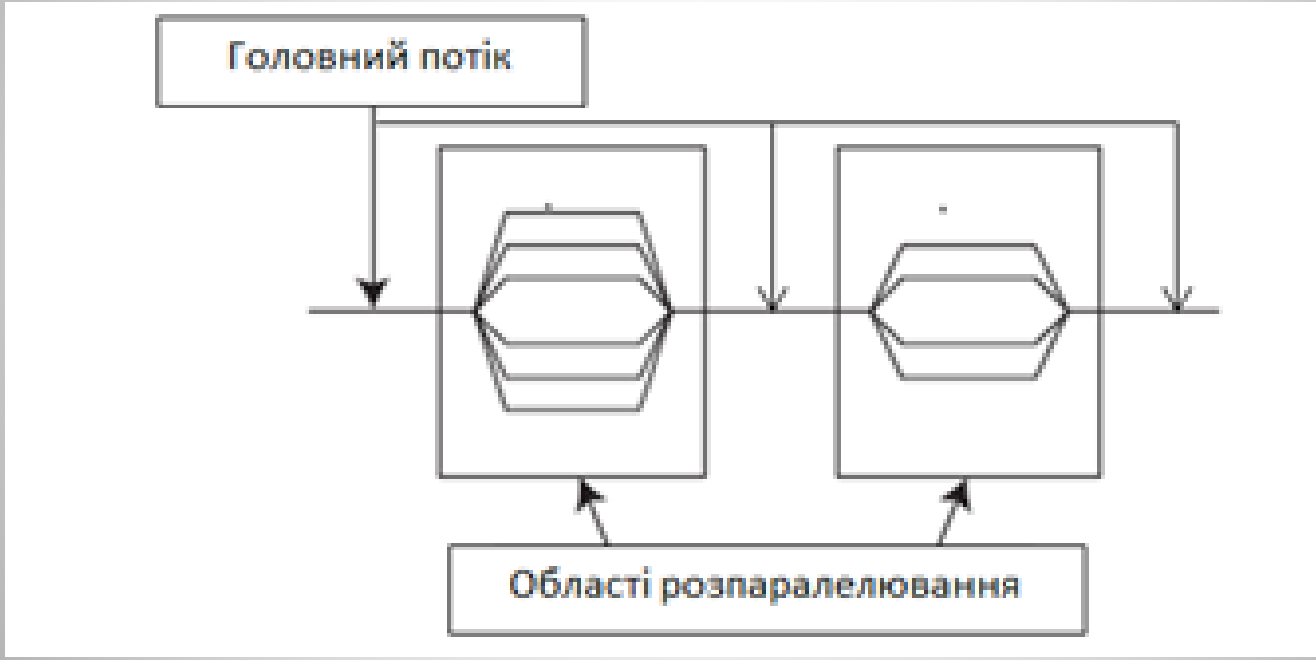
  - багатопоточність (виконання декількох потоків одночасно на одному CPU);
  - обчислення на графічних картах (GPGPU), у тому числі на декількох GPU одночасно (Multi-GPU);
  - паралельні обчислення на основі розподілених систем та інші.

- Кожний з наведених способів має свої переваги та недоліки. Доцільність їхнього застосування залежить від наявних ресурсів та поставленої задачі.

Багатопоточність

- Під багатопоточністю зазвичай розуміють виконання декількох потоків одночасно на одному CPU.
- Робота багатопоточної програми починається з ініціалізації та виконання головного потоку (процесу), який у міру необхідності створює і виконує паралельні потоки, передаючи їм необхідні дані.
- Перевагами багатопоточності є можливість більш ефективно використовувати обчислювальну потужність CPU, а також можливість забезпечити повноцінну взаємодію між паралельними потоками.
- До недоліків варто віднести варто віднести складність програмування, що виникає з потреби синхронізувати та узгоджувати роботи потоків. Окрім того, темпи зростання обчислювальної потужності CPU є відносно повільними.

Принципова схема паралельної програми



GPGPU - технології

- GPGPU – це техніка використання графічного процесору відеокарти для виконання неграфічних обчислень.
- Однією з головних переваг GPU є велика кількість ядер (декілька тисяч), що дозволяє в повній мірі реалізувати переваги паралелізму. Більш того, спостерігається тенденція до швидкого росту обчислювальних потужностей GPU.
- Недоліками GPGPU є потреба в значній модифікації існуючих алгоритмів для їх виконання на GPU. Окрім того, організація GPGPU-обчислень та переміщення даних для обробки з CPU на GPU та навпаки є досить трудомістким процесом, що дещо звужує коло задач, для яких доцільно використовувати GPGPU.

NVIDIA CUDA

- CUDA – це розроблена NVIDIA програмно-апаратна архітектура паралельних обчислень на основі мови програмування C.
- Важливою особливістю CUDA є наявність додаткових бібліотек, які дозволяють значно спростити паралельне програмування на GPU. Прикладом такої бібліотеки є CUDA Thrust
- Значною перевагою CUDA є підтримка технології Multi-GPU Programming, яка дозволяє одночасне використання декількох графічних карт при виконанні CUDA-обчислень.

Комп’ютерне моделювання паралельно-ієрархічного оброблення зображень у гетерогенному програмно-апаратному забезпеченні

- Задача: виконання елементарних арифметичних дій над матрицями.
- Початковими даними є числові масиви розмірністю від 16384 до 67108864 елементів, що відповідає кількості пікселів в зображеннях розмірністю від 128x128 до 8192x8192 пікселів.
- В тестовій програмі проводилась робота з 20 парами числових масивів, що відповідає набору з 20 зображень. Кількість операцій, що виконувалась в процесі моделювання приймала значення 250, 500 та 1000 операцій на одне зображення (пару числових масивів).
- Загалом з точки зору паралельних обчислень моделювання було проведене в чотирьох режимах:

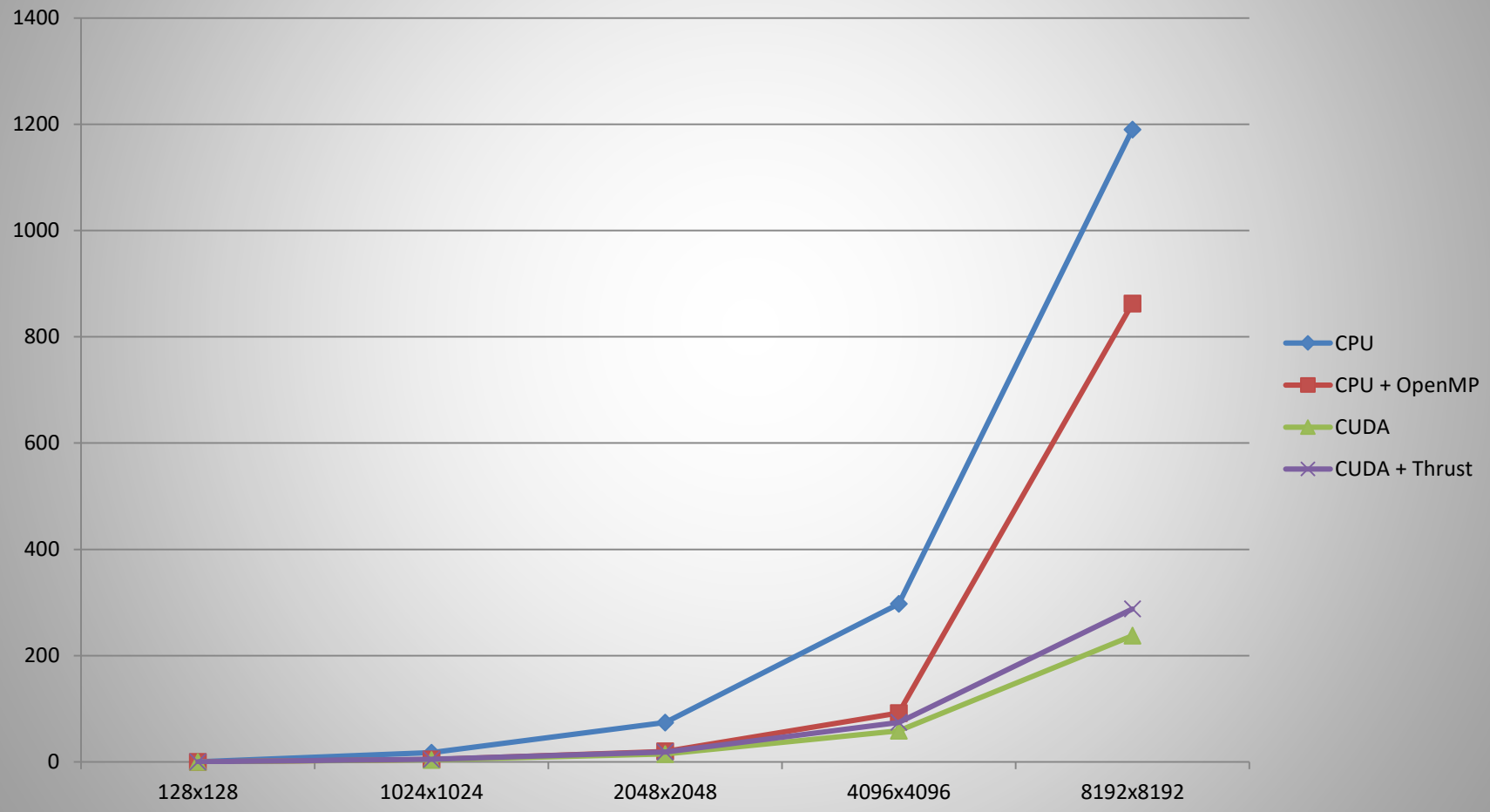
  1) Виконання на CPU без застосування паралельних обчислень.

  2) Виконання на CPU з застосуванням багатопоточності (OpenMP) на рівні набору зображень, тобто декілька зображень оброблялись одночасно в паралельному режимі

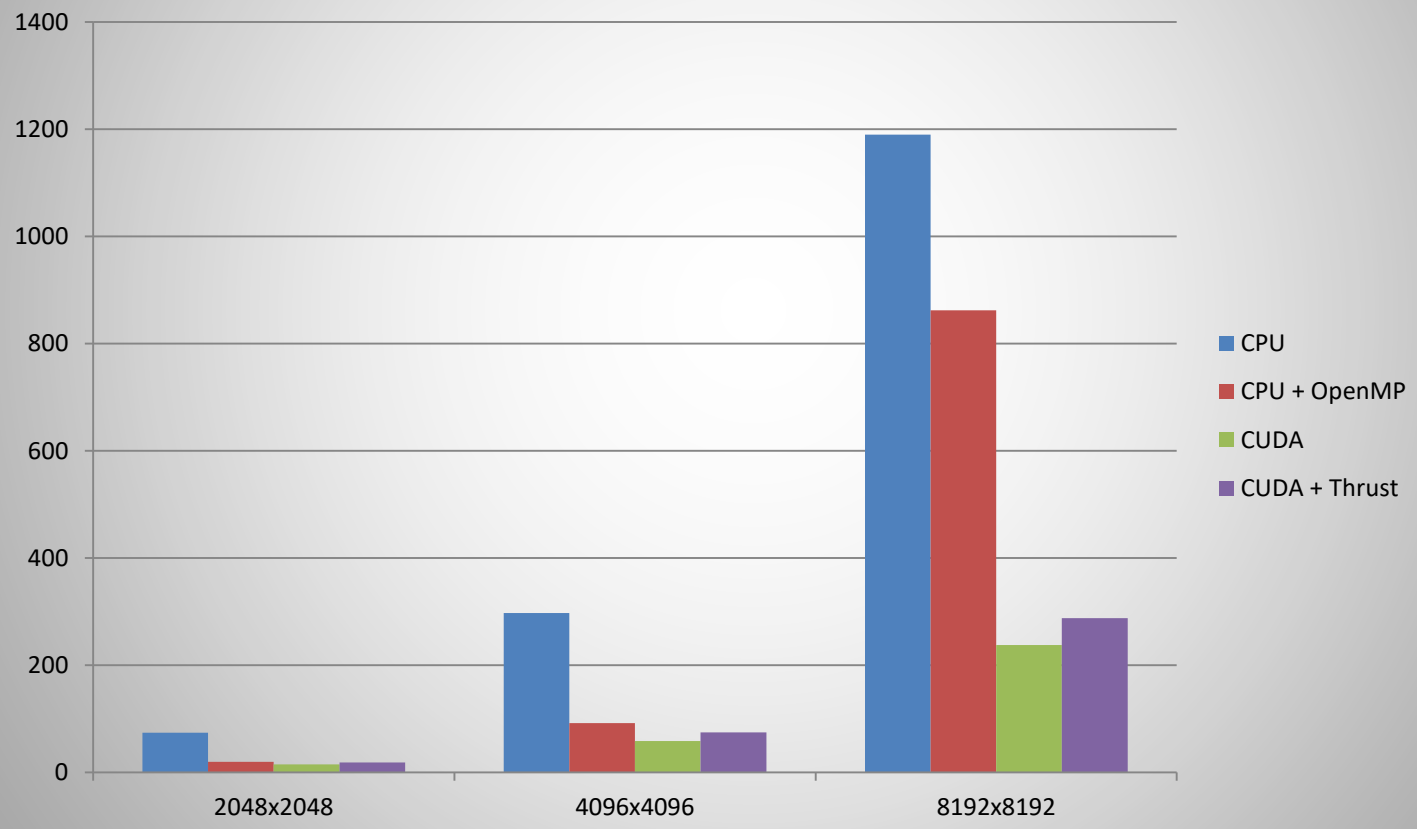
  3) Виконання на GPU на основі GPGPU-технології NVIDIA CUDA.

  4) Виконання на GPU на основі GPGPU-технології NVIDIA CUDA з застосуванням бібліотеки CUDA Thrust

Графічне представлення моделювання обробки зображень при 1000 операціях на одне зображення



Графічне представлення моделювання обробки зображень при 1000 операціях на одне зображення



Відносний приріст швидкодії від реалізації обробки зображень: на основі NVIDIA CUDA в порівнянні з послідовною реалізацією на CPU

Розмір зображення (масиву)	**Приріст швидкодії (рази) – 250 операцій**	**Приріст швидкодії (рази) – 500 операцій**	**Приріст швидкодії (рази) – 1000 операцій**
128x128	1,01	1,32	1,9
1024x1024	3,05	3,87	4,62
2048x2048	3,4	4,29	5,01
4096x4096	3,44	4,34	5,05
8192x8192	3,86	4,77	5,12

Реалізація ПІ оброблення зображень на основі Multi-GPU Programming

- Алгоритм ПІ перетворення є орієнтованим на паралельне виконання, у тому числі на основі NVIDIA CUDA. Проте його реалізація на основі системи з двома GPU лише засобами Multi-GPU Programming виявилась недоцільною через синхронність окремих функцій.
- Комбіноване застосування OpenMP та Multi-GPU Programming для створення двох незалежних потоків, кожен з яких працює зі своїм GPU та набором зображень дозволило обійти це обмеження та досягти значного збільшення швидкодії за рахунок використання 4 GPU.

Алгоритм ПІ оброблення зображень з використанням двох графічних карт

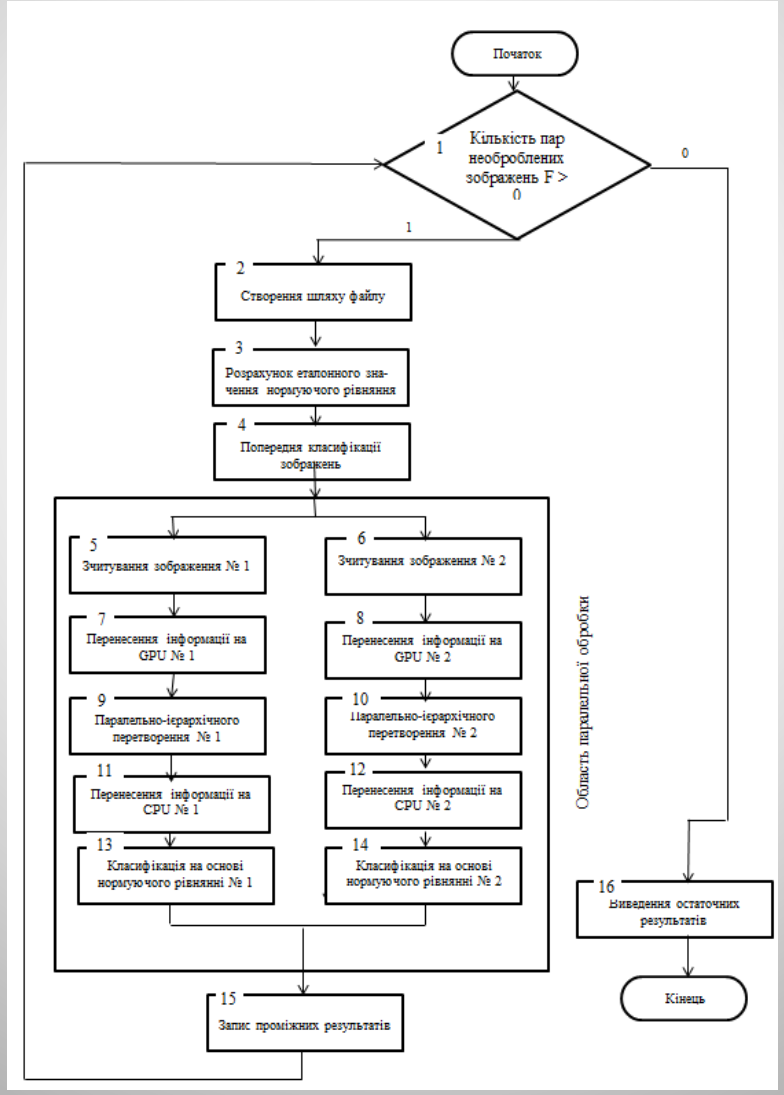
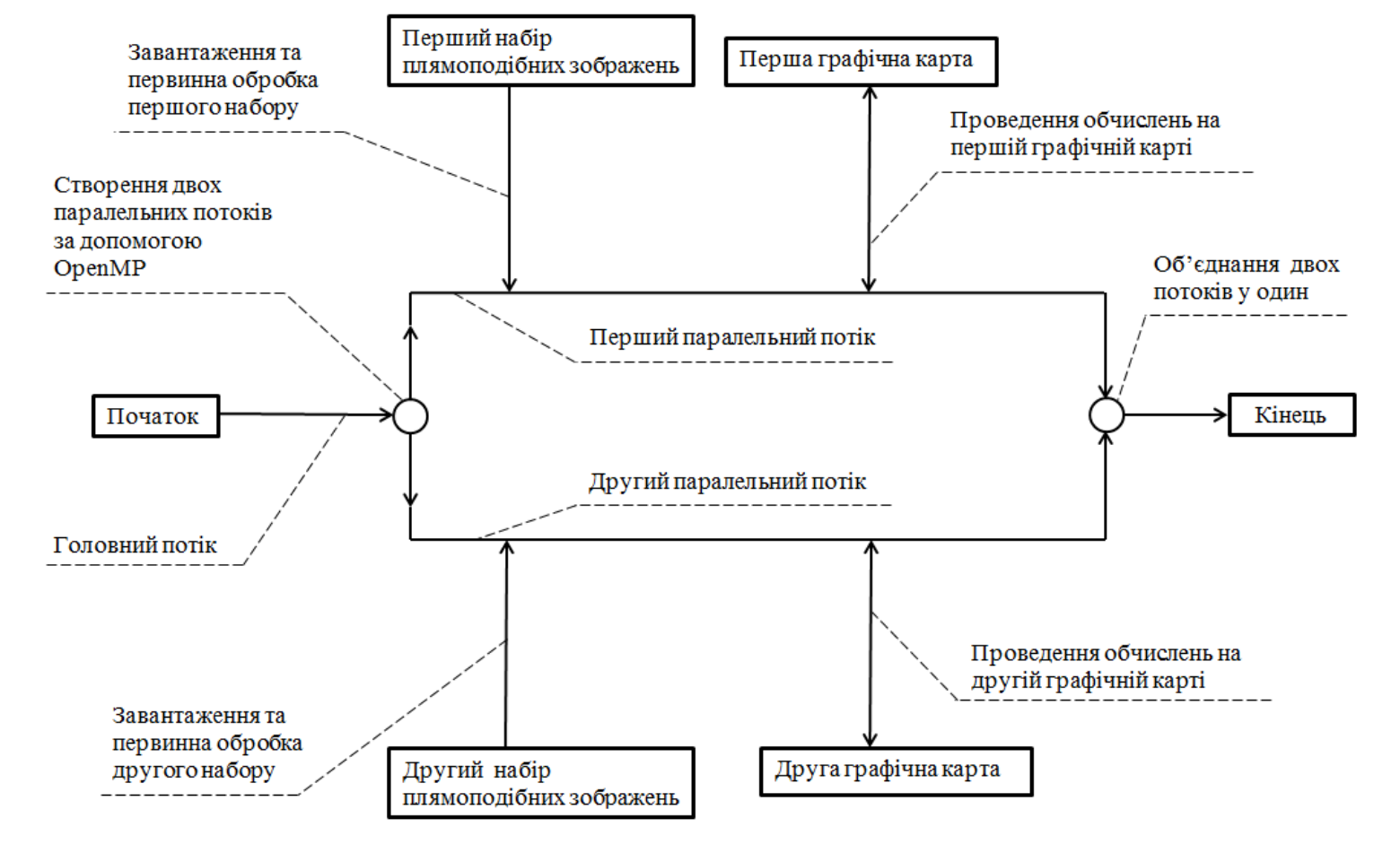
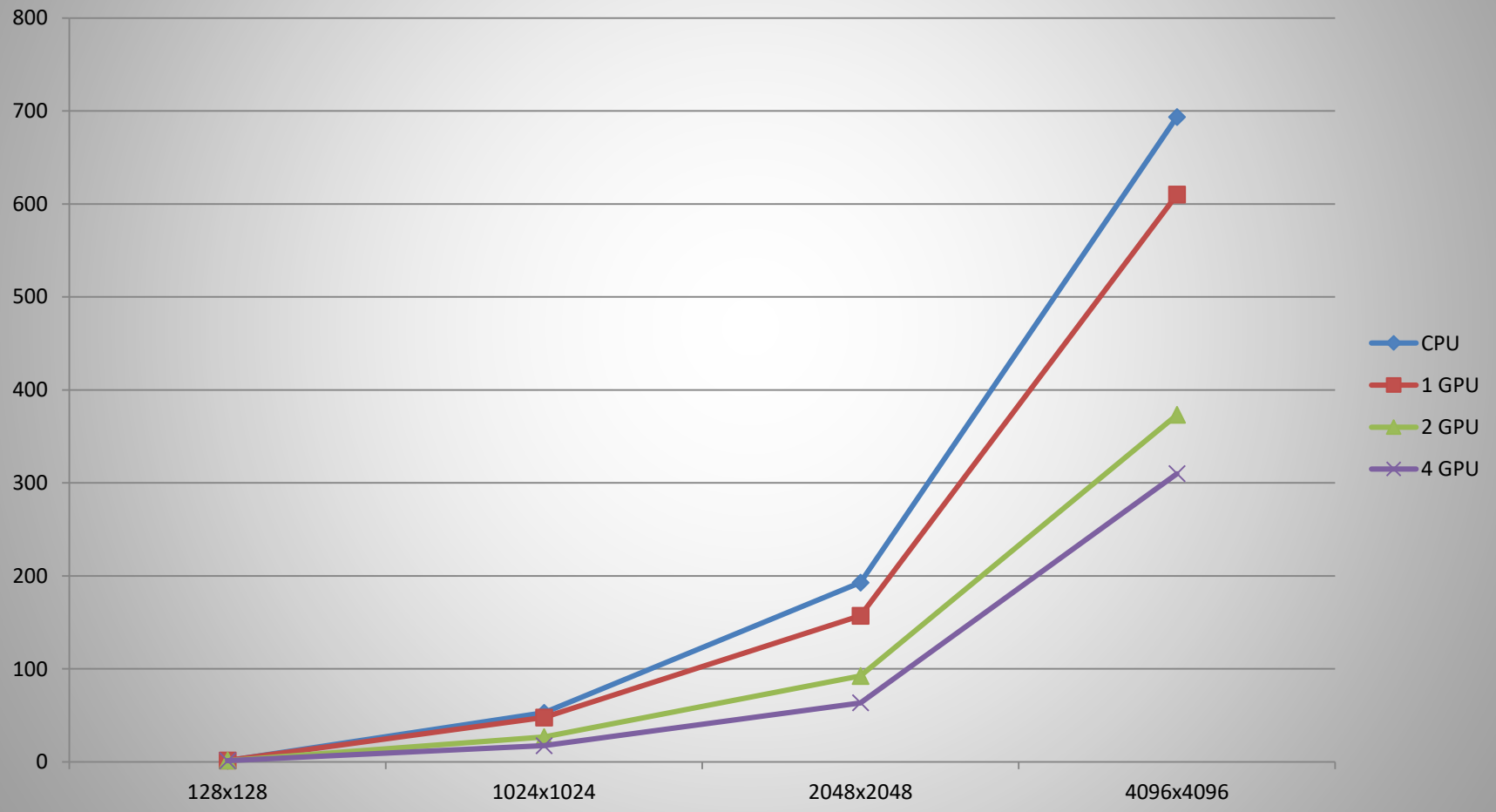


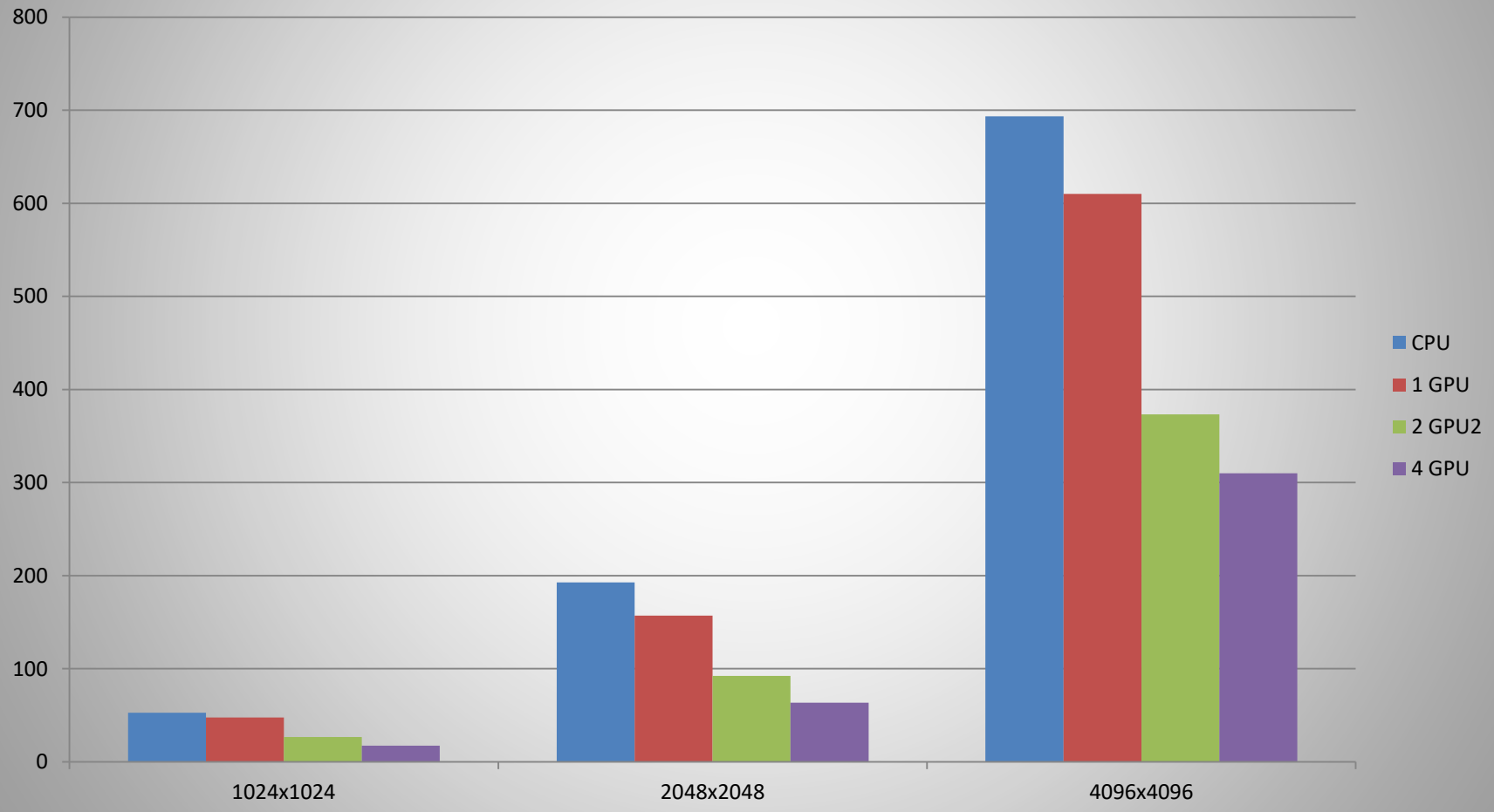
Схема взаємодії NVIDIA CUDA та OpenMP при реалізації ПІ перетворення



Графічне представлення часу виконання роботи програми



Діаграма часу виконання роботи програми



Відносний приріст швидкодії відносно послідовної реалізації

	Час,с		
Розмір зображення (масиву)	1 GPU	2 GPU	4 GPU
128x128	0,77	0,871	0,96
1024x1024	1,109	1,98	3,035
2048x2048	1,226	2,086	3,038
4096x4096	1,136	1,858	2,238

Відносний приріст швидкодії відносно реалізації на основі одного GPU

	Час,с	
Розмір зображення (масиву)	2 GPU	4 GPU
128x128	1,12	1.236
1024x1024	1,78	2,735
2048x2048	1,7	2,47
4096x4096	1,63	1,96

Результат роботи програмного продукту

```
Размер матрицы хвостовых элементов для изображения Frame00001.bmp составляет 643
9984 12800 17664 4000 7140 672 6720 31068 804 6492 7812 8316 7044 2864 2380 688 5668 11228 31
040 1136 288 276 752 4076 896 4516 60 1140 512 592 524 1312 212 668 132 460 1152 68 160 656 3
96 2432 3844 268 2292 1020 48 3328 80 1664 3060 1792 1696 176 924 1136 3728 1248 2276 3020 15

Качество изображения Frame00001.bmp - очень высокое

Размер матрицы хвостовых элементов для изображения Frame00002.bmp составляет 651
1097728 31232 762 262 36 192 1510 600 3194 3816 1688 710 922 3982 32 938 776 494 5032 2 798 5
0 1450 370 288 1882 584 106 972 352 660 130 96 212 554 1622 258 844 460 136 166 1184 524 436
90 266 68 1148 924 214 878 1070 188 3962 266 922 1062 200 1036 1702 398 282 2564 348 488 96 5

Качество изображения Frame00002.bmp - очень высокое

Размер матрицы хвостовых элементов для изображения Frame00003.bmp составляет 657
1064960 30976 1524 2000 72 288 576 4896 3600 224 2512 5504 240 1072 5408 2832 1296 2592 5408
04 252 828 880 244 1352 1208 172 668 20 1420 2304 2860 268 276 1624 2140 168 1052 520 504 40
1260 120 1792 472 1040 1392 180 200 656 324 440 264 640 1288 1272 724 344 876 3880 660 272 37

Качество изображения Frame00003.bmp - очень высокое

Размер матрицы хвостовых элементов для изображения Frame00004.bmp составляет 648
1097728 31232 1024 500 144 384 940 2240 340 588 692 64 204 1652 3020 1908 1868 2292 140 4596
8 476 128 368 484 476 1248 272 84 1488 128 620 1300 208 624 592 332 20 496 812 212 60 240 240
336 700 1172 480 236 1840 496 116 28 452 2076 2016 36 1548 16 704 244 60 240 3796 1500 512 43

Качество изображения Frame00004.bmp - очень высокое

Размер матрицы хвостовых элементов для изображения Frame00005.bmp составляет 625
1114112 31232 1024 1244 324 288 1440 192 1056 48 48 2928 1936 4192 2768 528 2192 112 496 1008
288 1004 420 144 76 512 292 16 572 1124 224 448 800 400 80 1892 104 208 120 64 240 860 1148 1
16 764 1140 160 492 264 788 480 928 356 1068 448 668 1784 208 1344 20 576 368 1284 844 320 64

Качество изображения Frame00005.bmp - среднее

Размер матрицы хвостовых элементов для изображения Frame00006.bmp составляет 649
1081344 30464 2048 1000 1016 192 144 768 3520 80 1720 720 552 6104 4056 568 1416 4944 984 119
512 24 248 608 1376 280 552 680 288 360 408 1536 80 1144 664 232 136 1544 752 416 96 1976 496
08 1016 848 1496 328 224 96 192 2416 552 720 1656 5248 32 1296 248 712 352 296 184 320 920 11
```

Економічна частина

- Загальні витрати на виконання та впровадження результатів виконаної наукової роботи становить 15723,62 грн.
- Очікуванні комерційні ефекти від реалізації результатів розробки:
  - абсолютна ефективність вкладених інвестицій дорівнює 49862 грн.;
  - відносна щорічна ефективність вкладених коштів становить 38,9%;
  - термін окупності вкладених інвестицій становить 2,57 років.

Висновки

- Згідно з отриманими результатами тестування, при розмірності зображень 1024x1024 пікселів за рахунок застосування 4 графічних карт досягається приріст швидкодії на 203% в порівнянні з послідовною реалізацією на CPU та на 173% в порівнянні з реалізацією на основі одного GPU.
- При обробленні зображень розмірності 2048x2048 пікселів з застосуванням 4 графічних карт досягається приріст швидкодії на 203% в порівнянні з послідовною реалізацією на CPU та на 173% в порівнянні з реалізацією на основі одного GPU.
- При розмірності зображень 4096x4096 пікселів застосування 4 графічних карт дозволяє досягається приріст швидкодії на 123% в порівнянні з послідовною реалізацією на CPU та на 96% в порівнянні з реалізацією на основі одного GPU.
- Такі результати свідчать про досягнення мети магістерської кваліфікаційної роботи, яка мала наступний вигляд: підвищення швидкодії процесу паралельно-ієрархічної обробки плямоподібних зображень на основі гетерогенного програмно-апаратного забезпечення.
- Отримані результати підтверджують доцільність застосування технологій гетерогенних обчислень при паралельно-ієрархічній обробці зображень та вказують на перспективність застосування комбінованого паралелізму, який досягається за рахунок взаємодії різних програмних та апаратних засобів.

Зв'язок роботи з науковими програмами, планами, темами

- Магістерська кваліфікаційна робота виконана в межах НДР GP/F61/083 "Методологія побудови високопродуктивних інтелектуалізованих паралельно-ієрархічних систем на основі сучасних мережевих обчислювальних комплексів з гетерогенною архітектурою".

Апробація результатів

- Результати роботи апробовані на 5 наукових конференціях, зокрема на VII Міжнародній науково-технічній конференції «Фотоніка – ODS» 2015, м. Вінниця; X Міжнародній науково-практичній конференції «ІОН-2016», м. Вінниця; VI Міжнародній конференції студентів і молодих науковців «МІТ-2016», м. Одеса; XLIV та XLV НТК професорсько-викладацького складу, співробітників та студентів ВНТУ, м. Вінниця, 2015-2016р.р.

Публікації

- На тему магістерської кваліфікаційної роботи опубліковано 9 друкованих робіт, у тому числі 4 статті у наукових журналах, що входять до переліку фахових видань України (входять до наукометричних баз даних РІНЦ, Google Scholar). Окрім того, отримано свідоцтво про реєстрацію авторського права на твір (комп’ютерну програму)
- За результатами роботи подано та прийнято до друку статтю у наукове видання, що входить до наукометричної бази даних “Scopus”.
- Окрім того, основні результати досліджень впроваджено на ТОВ "Практик-С", м. Вінниця.

Дякую за увагу