УДК 681.3:004.272

ОРГАНІЗАЦІЯ ПАРАЛЕЛЬНИХ ОБЧИСЛЕНЬ НА ОСНОВІ ТЕХНОЛОГІЇ GPGPU ТА СПОСОБИ ЇХ ОПТИМІЗАЦІЇ.

Пасічник Д.Г.
Науковий керівник – д.т.н., проф. Яровий А.А.

Сучасні відеокарти містять сотні математичних виконавчих блоків, що можуть використовуватися для значного прискорення великої кількості обчислювально-інтенсивних додатків. Нинішні покоління GPU мають досить гнучку архітектуру, що разом із високорівневими мовами програмування та програмно-апаратними архітектурами, повністю розкривають їх можливості і забезпечують їх більшу доступність. Для використання таких можливостей основні виробники відеокарт, – NVIDIA і AMD, розробили і анонсували відповідні платформи під назвою CUDA (Compute Unified Device Architecture) і CTM (Close To Metal, пізніше – AMD StreamComputing).

Метою роботи є дослідження особливостей організації високопродуктивних паралельних обчислень на основі технології GPGPU, аналіз програмної платформи та способів оптимізації паралельних обчислень в CUDA.

Варто зазначити, що для організації обчислень на GPU можна використовувати графічні APIOpenGL і Direct3D, у яких дані до відеокарти передаються у вигляді текстур, а розрахункові програми завантажуються у вигляді шейдерів. Хоча ці стандарти здаються найбільш універсальними, вони дозволяють використовувати один і той же код для відеокарт різних виробників, проте у них є недоліки, оскільки вони менш гнучкі і не зручні у використанні. Крім того, вони не дозволяють використовувати специфічні можливості певних відеокарт, такі як швидка розділена (колективна) пам'ять, присутня в сучасних обчислювальних процесорах. Платформа CUDA ліквідувала деякі з важливих обмежень попередніх моделей GPU і володіє рядом переваг. Платформа забезпечує доступ до розподіленої між потоками пам'яті, яка може бути використана для організації кеша з широкою смугою пропускання, в порівнянні з текстурними вибірками. В CUDA доступна лінійна адресація пам'яті, яка надає можливість запису даних за довільними адресами. Хоча трудомісткість програмування GPU за допомогою CUDA досить велика, вона нижча, порівняно з ранніми GPGPU-рішеннями. У загальних рисах, при оптимізації програми CUDA потрібно домогтися оптимального балансу між розмірністю та кількістю блоків, використовувати ті алгоритми, які добре разпаралелюються, використовувати буферизацію, а також мінімізувати передачу даних між CPU і GPU.

В перспективі практичне застосування результатів даного дослідження дозволить підвищити ефективність паралельних обчислень на основі технології GPGPU та забезпечить оптимальне використання обчислювальних ресурсів.