

АНАЛІЗ ПАРАЛЕЛЬНИХ ОБЧИСЛЕНЬ НА GPU ТА ЇХ ОПТИМІЗАЦІЯ

Богомолів Ю.С., Вознесенський К.Ю.
Науковий керівник – доц., к.т.н. Яровий А.А.

На сьогодні масштабні паралельні обчислення із використанням відоадаптерів є широко відомою та застосовуваною технологією. Саме тому у даній роботі розглядаються особливості уніфікованої архітектури відоадаптерів, зумовлені нею «вузькі місця» та способи оптимізації програм.

За мету даного дослідження поставлено огляд архітектури відоадаптерів та аналіз методів оптимізації програм, що використовують обчислення на відоадаптерах.

Принцип паралелізму реалізовано в сучасних відоадаптерах на багатьох рівнях. У паралельно працюючих над різними задачами (потокami інструкцій) SIMD-блоках виконується паралельна обробка даних однієї задачі. У сучасних відоадаптерах зазвичай використовується від 2 до 8 контролерів пам'яті, що дозволяє збільшити сумарну пропускну спроможність пам'яті.

До найбільш відомих способів оптимізації програм із використанням обчислень на відоадаптерах належать: спосіб маскуванню затримок (при звертанні до глобальної пам'яті та пам'яті центрального процесора); спосіб узгодження потоків в рамках одного SIMD-блоку; спосіб послідовного зчитування даних з пам'яті. Кожний із цих способів має свої переваги та недоліки та може бути використаним лише у певних випадках. Детально розглянуто один із найбільш ефективних методів оптимізації – спосіб маскуванню затримок.

За використання схеми маскуванню затримок відбуваються такі дії: послідовно виконуються запити від всіх потоків, після чого ядро очікує на дані протягом 32 тактів після першого запиту (одноразова затримка); дані надходять із пам'яті безперервно у порядку їх запиту, передача керування між потоками відбувається миттєво, без додаткових затримок. Таким чином, за використання схеми маскуванню затримок максимальний час простою ядра рівний 32 тактам.