УДК 004.42
А. Воронка
(Тернопільський національний технічний університет імені Івана Пулюя, Україна)

МОДЕЛЬ ПАМ'ЯТІ ТЕХНОЛОГІЇ CUDA

UDC 004.42
A. Voronka

CUDA TECHNOLOGY MEMORY MODEL

Можливість вільного доступу до пам'яті з побайтовою адресацією, є надзвичайно важливим моментом. Самі потоки CUDA можуть адресуватися до даних з різних просторів пам'яті в один і той самий час. Будь-який потік має приватну локальну пам'ять. Кожен блок потоку володіє загальною пам'яттю, котра видима всім потокам блоку і з тим самим часом життя, як і блок. У всіх потоків є доступ до однієї й тієї ж глобальної пам'яті. Є також два додаткові простори для зчитування, котрі доступні для всіх потоків, а саме постійної та текстурної пам'яті. Пам'ять текстур також пропонує ряд варіантів звертань та фільтрацію даних для окремих форматів.

Глобальна пам'ять. Є у пам'яті пристрою, а вона – через 32-, 64- або 128-байтові транзакції. Вони повинні бути вирівняні: тільки 32-, 64- або 128-байтові сегменти пам'яті пристрою, які вирівняні за тим, чия власне перша адреса кратна їх розміру, можуть бути прочитані чи записані в пам'ять угоди. Скільки транзакцій та пропускної здатності зрештою потрібно, залежить від обчислювальної здатності пристрою. Compute Capability 2.x, 3.x, 5.x та 6.x дають більш детальну інформацію про те, як обробляються глобальні звернення до пам'яті для різних обчислювальних можливостей.

Локальна пам'ять. Доступ до неї відбувається лише для деяких автоматичних змінних. Автоматичними змінними, які компілятор може розмістити в такій пам'яті, є: масиви, для котрих він не може визначити, що їх індексують зі сталими одиницями, великі структури чи масиви, які споживатимуть занадто багато простору для реєстрації, будь-яка змінна, якщо власне ядро бере більше регістрів, чим доступно (це також відомо як розлиття регістрів).

Загальна пам'ять. Так як вона вмонтована в чіп, тому володіє вищою пропускною здатністю та значно меншою затримкою, ніж локальна чи глобальна пам'ять. Для одержання високої пропускної здатності пам'ять поділяється на однакові за розміром модулі однакового розміру (банки), до яких можна одночасно звертатися. Таким чином, будь-який запит на читання або запис в пам'ять, що складається з n адрес, які потрапляють у n різних банків пам'яті, може обслуговуватися одночасно, що дає загальну пропускну здатність, яка в n разів перевищує пропускну здатність одного модуля.

Постійна пам'ять. Є в пам'яті пристрою та кешується у постійному кеші, вказаному в Compute Capability 2.x. Потім запит розбивається на стільки окремих запитів, що у вихідному запиті є різні адреси пам'яті, що зменшує пропускну здатність на коефіцієнт, що дорівнює кількості окремих запитів. Потім одержані запити обслуговуються за пропускної здатності постійного кешу у випадку попадання в кеш або за пропускної здатності власне пам'яті пристрою в іншому разі.

Текстурна пам'ять. Пам'ять текстури та поверхні є у пам'яті пристрої та кешуються в кеші текстури, тому при їх зчитуванні стоїть одна пам'ять, що зчитується з пам'яті пристрою тільки при пропуску кеша, інакше це просто коштує одного читання з кешу текстур. Кеш текстури оптимізовано для 2D просторової локальності, тому потоки одного і того ж детектора, які читають текстурні або поверхневі адреси, що знаходяться близько один до одного в 2D, досягнуть найкращої продуктивності. Крім того, він призначений для потокового завантаження із постійною затримкою. Кеш- кеш зменшує потребу в пропускній здатності, але не покриває затримку.