УДК 004.925.5

АНАЛІЗ АРХІТЕКТУРИ TURING ДЛЯ ПОБУДОВИ ВІДЕОКАРТ

Романюк О.Н., д.т.н., проф., E-mail: rom8591@gmail.com
Чан А.Л.В.., студентка
Панфілова Ю.О., студентка

Для формування реалістичних графічних зображень широко використовують відеокарти, зокрема, компанії NVIDIA, яка модернізувала серію відеокарт на базі Turing, що збільшило її продуктивність і зробило цю архітектуру значно ближчою до Volta, ніж до Pascal [1]. У статті проаналізовано особливості модернізації та нововведень.

Метод трасування променів (англ. ray tracing) у комп'ютерній графіці є методом формування зображення тривимірних об'єктів чи сцени за допомогою відстеження ходу променя світла крізь точку екрану і симуляції взаємодії цього променя з уявними об'єктами, що підлягають відображенню. Цей метод дозволяє створювати надзвичайно реалістичні зображення.

NVIDIA Turing — одні із найпотужніших і перші в світі графічні процесори для відслідковування променів, які використовують компоненти, що здатні обраховувати 10 млрд. променів за секунду [2]. Порівняно з Pascal, це є 25-кратним вдосконаленням характеристик щодо трасування променів. У серію відеокарт на базі Turing входять три моделі: Quadro RTX 5000, 6000 та 8000.

Однією з головних переваг нової архітектури Turing фірма Nvidia називає технологію трасування променів в реальному часі, яку раніше не застосовували в графічних прискорювачах споживчого класу. Дана технологія, як метод візуалізації, не є новою і про неї було відомо ще раніше [1]. Тим не менш цей напрямок і сьогодні залишається актуальним. Головною проблемою технології трасування променів в реальному часі є відсутність необхідних обчислювальних ресурсів. Із впровадженням вдосконаленої архітектури Turing дану проблему вирішили шляхом поєднання трасування променів із методами растеризації, таким чином створивши гібридний конвеєр. Як результат, була отримана можливість отримувати якість графіки в реальному часі близьку до повноцінного трасування променів [3].

Водночас для Turing у архітектури Volta було запозичено можливість одночасного виконання арифметичних інструкцій FP32 (основне навантаження під час обробки шейдерів) і операцій блоку INT32, що дозволяє графічним процесорам даної архітектури паралельно виконувати операції з плаваючою і фіксованою комою [4], а також забезпечує адресацію та вибірку даних, порівняння тощо. Новизною ж тут є підтримка більш широкого діапазону точності, що дозволяє в кілька разів прискорити певні робочі навантаження, які не вимагають високої точності. Окрім режиму половинної точності обчислень з плаваючою комою FP16, було впроваджено підтримку цілочисельних інструкцій INT8 і INT4, що, відповідно, в 2 і 4 рази швидше, ніж у FP16[3]. Використовуючи великі й точні типи даних тільки за необхідності, їх можна

запакувати разом аби зменшити час виконання роботи та збільшити її об'єм. (Саме це стало причиною збільшення продуктивності ядер Turing, ніж Pascal, на одній і тій же тактовій частоті.) Зменшення точності нейронної системи при використанні INT4 дозволило багатократно пришвидшити обчислення, що є надзвичайно важливим, особливо в процесах визначення логічних висновків при реалізації штучного інтелекту. Так, архітектура Turing оснащена тензорними ядрами, які можуть забезпечити продуктивність обчислень штучного інтелекту вище 100 Терафлопс [5].

В основі нової архітектури лежить чіп Turing GPU з 18,6 мільярдами транзисторів — це пояснює, чому розміри кристалу графічного процесора становлять 754 мм2 [2]. З усієї лінійки компанії він першим отримав підтримку пам'яті GDDR6 (найбільш швидкісної в світі пам'яті, що забезпечує велику кількість кадрів в секунду при відтворенні зображення), з 256- та 384-розрядними шинами [6].

Архітектура NVIDIA Turing використовує кластерну структуру виконавчих модулів. Кожен кластер (GPC — Graphics Processing Cluster) має 8 (у TU104) та 12 (у TU102 і TU106) SM-модулів. У свою чергу SM-модуль складається із CUDA-ядер кількістю 64 одиниць, 8 Tensor-ядер для завдань машинного навчання з продуктивністю до 500 трлн тензорних операцій за секунду, та одного RT-ядра для розрахунку трасування променів (RayTracing) [4].

Основним нововведенням в архітектурі Turing є ще більша порівняно з архітектурою Volta апаратна орієнтованість на трасування променів, яка є реалізована в нових RT-ядрах для трасування. Ці процесорні блоки прискорюють перевірку перетину променів, трикутників і маніпуляцій з ієрархіями обмежувальних об'ємів (Bounding Volume Hierarchies, BHV), що є широковживаною структурою даних для зберігання об'єктів при трасуванні променів. RT-ядра прискорюють розрахунки руху світла та звуку в 3D-середовищі.

Важливі зміни відбулися на рівні мультипроцесорних блоків SM, що мають стандартну структуру в усіх варіантах GPU Turing. Всі обчислювальні блоки всередині SM згруповано в чотири масиви обробки даних із логікою управління, що включає регістри, 1 планувальник на кожні 16 ядер (удвічі більше, ніж у Pascal) і один порт диспетчера на кожні 16 ядер (стільки ж, як і у Pascal). При цьому в одному SM наявні 64 потокових процесори. Така конструкція дозволяє усунути задачу об'єднання інструкцій у пари. Враховуючи, що в Turing вдвічі більше планувальників, для роботи ядер CUDA їм достатньо просто відправляти по одній інструкції за такт. При цьому за наявності групи з 32 потоків 16 ядер CUDA достатньо всього два такти для виконання [1].

Таким чином, пришвидшення роботи відбувається не за рахунок ускладнення конструкції, а за рахунок її оптимізації. В Turing використовуються ті ж самі ресурси, але більш збалансовано. Схема SM чіпу попереднього покоління GP102 виглядає більш складною і насиченою, хоча в TU102 72 мультипроцесори SM, а в GP102 їх не більше 30. В результаті

флагманський чіп з архітектурою Turing має на 21% більше ядер CUDA і текстурних блоків.

У сучасних додатках обчислення цілих чисел займають близько 36% з усіх обчислень при виконанні. Виконання операцій двох типів в один потік значно прискорить загальні обчислення. Оновлена уніфікована структура кешу L1 дозволяє конвеєру TPC працювати з ним ефективніше. При збереженні загального об'єму кешу L1 на рівні 96 КБ загальна пропускна здатність може вирости до двох разів (до 100 Гбайт/с). Водночас у всіх процесорах збільшений об'єм загального кешу L2 (в GPU TU102 він становить 6 МБ, тоді як у старого GP102 лише 3 МБ) [7].

Архітектура відеокарт Turing реалізована на трьох базових графічних процесорах: TU102, TU104 та TU106.

Графічний процесор TU102 — це мікросхема площею 754 мм^2. Порівняно з найбільшим десктопним графічним процесором GP102 на архітектурі Pascal, цей процесор має на 60% більшу площу та містить на 55% більше транзисторів (18,6 млрд). Повноцінний процесор TU102 складається з шести обчислювальних кластерів, у кожен з яких входить модуль растеризації і шість блоків обробки текстур, одного модуля поліморфів і двох потокових мультипроцесорів SM. Загалом отримуємо 72 мультипроцесори SM, 4608 ядер CUDA, 576 тензорних ядер, 72 ядра RT, 288 текстурних блоки та 36 модулів обробки поліморфів [8].

Графічний процесор TU104 — мікросхема площею 545 мм2, що містить 13,6 млрд транзисторів, 6 графічних кластерів, кожен з яких має по 4 TPC. То ж загалом маємо: 48 мультипроцесорів SM, 3072 ядер CUDA, 384 тензорних ядер, 48 ядер RT, 192 текстурних блоків та 24 модулів обробки поліморфів [8].

Графічний процесор TU106 — остання відеокарта покоління Turing, складається з 10,8 млрд транзисторів, має площу 445 мм2. Має 3 кластери, в кожному з яких по шість TPC. Внутрішня конструкція TPC така ж, як і в інших чіпах на базі Turing. В результате отримуємо: 2304 ядер CUDA, 288 тензорних ядер, 36 ядер RT та 144 текстурних блоків [8].

Сьогодні в 3D-графіці більш поширеним є метод растеризації, який полягає у постійному перерахунку тривимірного зображення, поданого найпростішими геометричними фігурами, в пікселі. Причому візуальні ефекти реалізуються за допомогою шейдерів. На сьогоднішній день технології растеризації досягли досить високого рівня реалістичності, але всі ефекти, відображення і формування тіней моделюються при цьому штучно, в той час, як при трасуванні променів зображення формується точно, шляхом розрахунку траєкторій променів, відбитих від поверхонь полігонів.

Проаналізовано архітектуру Turing відеокарт. Визначено особливості архітектури. Розглянуто побудову базових процесорів на даній архітектурі, комбінованої чіпової пам'яті, управляння потоками, оптимізованого набору інструкцій. Наведено технічні характеристики графічної архітектури NVIDIA Turing.

Перелік посилань

1. «3D-ускорители Nvidia Quadro RTX, основанные на архитектуре Turing» KEDDR, 14 серпня 2018 р., [Електронний ресурс]. Режим доступу: https://keddr.com/2018/08/predstavlenyi-3d-uskoriteli-nvidia-quadro-rtx-osnovannyie-na-arhitekture-turing/
2. «NVIDIA. Раскрывая тайны архитектуры GPU Turing следующего поколения: удвоенный Ray Tracing, GDDR6 и многое другое» habr, 13 вересня 2018 р., [Електронний ресурс]. Режим доступу: https://habr.com/company/ua-hosting/blog/422773/
3. К. Ходаковский, «Что принесёт на рынок новая архитектура NVIDIA Turing?» 3DNEWS, 14 серпня 2018 р., [Електронний ресурс]. Режим доступу: https://3dnews.ru/973967
4. «Изучаем блок-схему графических процессоров NVIDIA Turing» OCclub, 12 вересня, 2018 р., [Електронний ресурс]. Режим доступу: https://occlub.ru/news/hardware/31179-izuchaem-blok-shemu-graficheskih-processorov-nvidia-turing
5. «NVIDIA TURING ИЗОБРЕТАЯ ГРАФИКУ ЗАНОВО» NVIDIA, [Електронний ресурс]. Режим доступу: https://www.nvidia.com/ru-ru/geforce/turing/
6. «Nvidia Turing: представлена архитектура видеокарт нового поколения» HI-Tech, 14 серпня 2018 р., [Електронний ресурс]. Режим доступу: https://hi-tech.mail.ru/news/nvidia-turing/
7. «Архитектура Turing и особенности новых видеокарт GeForce RTX» OverClockers, 27 вересня 2018 р., [Електронний ресурс]. Режим доступу: https://www.overclockers.ua/video/nvidia-turing-geforce-rtx/all/
8. «Архитектура Nvidia Turing: трассировка лучей и многое другое» Tom`s HARDWARE, 2 жовтня 2018 р., [Електронний ресурс]. Режим доступу: http://www.thg.ru/graphic/nvidia-turing-architektura/nvidia-turing-architektura-01.html