Романюк О. Н.,
Даньковська О. В.,
Вяткін С. І

АНАЛІЗ АРХІТЕКТУР ВІДЕОКАРТ КОМПАНІЇ NVIDIA

Наведено детальний аналіз архітектур відеокарт компанії NVIDIA.

Shows detailed analysis of NVIDIA graphics architectures.

Візуалізація - найважливіший етап формування тривимірного зображення, який полягає в проектуванні сцени на екран комп'ютера з урахуванням освітленості, матеріалів об'єктів сцени, положень джерел світла та точки спостереження. Цей етап вимагає великих обчислювальних витрат, оскільки сцена в пам'яті графічної системи зберігається в просторовому вигляді і для створення плоского зображення потрібно розрахувати інтенсивність освітлення для кожної точки зображення. При великій кількості джерел світла для складних за формою об'єктів такі розрахунки вимагають тривалого часу, тому сучасні відеокарти мають велику кількість обчислювальних ядер.

Мікроархітектура GPU має відмінність від мікроархітектури CPU. Завданням графіки є незалежна паралельна обробка даних, тому GPU має багато потоків. Мікроархітектура спроектована так, щоб виконувати велику кількість завдань.

GPU складається з декількох процесорних ядер, які в термінології NVIDIA називаються Streaming Multiprocessor, а в термінології ATI – SIMD Engine.

Для прикладу розглянемо архітектури відеокарт GeForce 7800 [1] і GeForce 8800, оскільки вони мають принципову різну організацію, характерну для різних поколінь відео карт.

У відеокарті GeForce 7800 реалізовано графічний конвеєр із використанням вершинних і піксельних процесорів. Відеокарта має 24

піксельних процесори PS, по одному текстурному блоці на конвеєр, 8 вершинних процесорів і 16 блоків растрових операцій (ROP) (рис. 1). Піксельні процесори згруповано по 4 для обробки квадів

Процесор PS має два векторні АЛП (2), здатні виконувати 2 різні операції над 4 компонентами та два міні-АЛП (найпростіші скалярні АЛП для виконання простих операцій). Кожний піксельний блок може виконувати інструкції типу MADD (множення/додавання). АЛП 3 використовують для формування оптичних ефектів. АЛП 1 за один такт можна або вибрати одне значення текстури й задіяти другий АЛП 2 для однієї або двох операцій, або використати обидва АЛП, якщо не вибирається текстура.

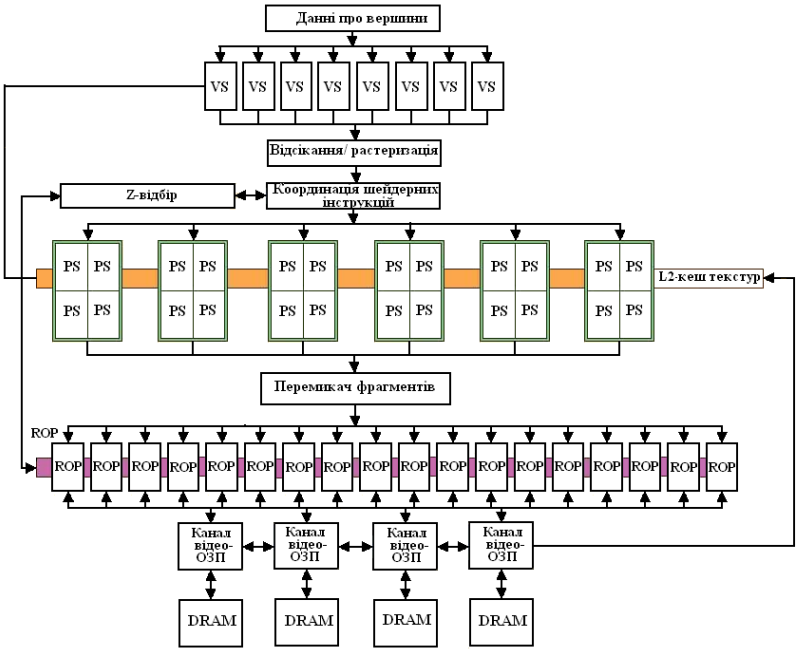

Рисунок 1 – Структура графічного процесора G70

За один такт вершинний процесор (рис. 3) може виконати одну векторну операцію, одну скалярну операцію й здійснити один доступ до текстури. До появи GeForce 8800 усім GPU було властиве одне фундаментальне обмеження – розподіл виконавчих пристроїв для піксельних і вершинних шейдерів. Відповідно, будь-який графічний процесор містив у своєму складі два окремих набори блоків для обробки кожного виду шейдерів. Такий розподіл міг негативно позначатися на

загальній ефективності роботи GPU, оскільки в сценах, насичених піксельними шейдерами, продуктивності наявних піксельних процесорів могло не вистачати, у той час як обчислювальні потужності вершинних процесорів не використовувалися, і навпаки. Проблему дисбалансу вирішила уніфікація шейдерних процесорів, при якій навантаження між ними розподілялися динамічно, залежно від особливостей конкретної сцени. У відеокарті GeForce 8800 вперше використано уніфіковану шейдерну архітектуру рендерингу, потокове оброблення інформації та новий вид шейдера – геометричний.

Чіп (рис. 4) складається з 8 універсальних процесорів, які включають 128 ALU і 32 TMU. Гранулярність виконання складає 8 блоків, кожний з яких може виконувати функції вершинного, піксельного, або геометричного шейдера над блоком із 32 пікселів. Його називають шейдерним процесором. Кожний такий процесор має кеш першого рівня L1, у якому зберігаються текстури й дані, які можуть бути використані шейдерним процесором. Блоки ROP визначать факт видимості, запис у буфер кадру й мультисемплінг. Вони згруповані з контролерами пам'яті, чергами запису та кешем другого рівня L2.

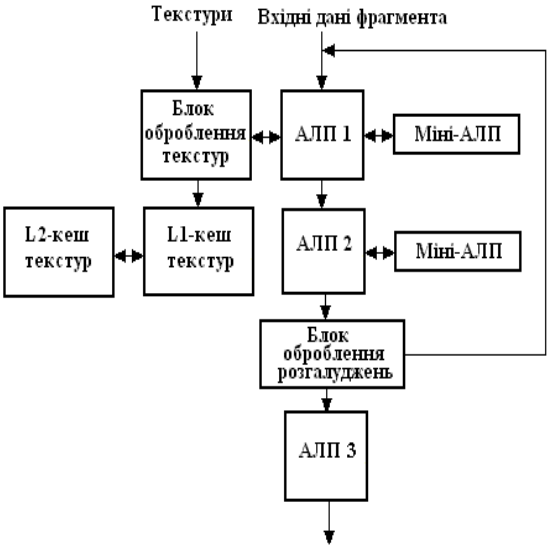

Рисунок 2 – Структура піксельного процесора

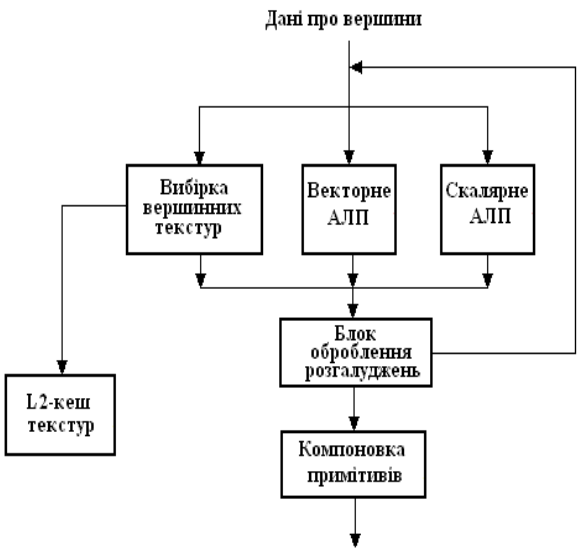

Рисунок 3 – Структура вершинного процесора

Потокові процесори SP є уніфікованими скалярними процесорами із плаваючою комою, що обробляють не тільки графічні, але й інші дані. Об'єднання SP у кластери дозволяє ефективно використовувати апаратні ресурси відеокарти. Кожний потоковий процесор на основі механізмів керування здатний динамічно перепризначуватися для виконання конвеєрних графічних або інших операцій. Thread Processor керує завантаженням потокових процесорів.

Крім шейдерних блоків і ROP у GeForce 8800 є набір керувальних блоків: Input Assembler приймає вихідні дані з пам'яті системи або локальної пам'яті; Setup/Raster/ZCull – блок, що виконує встановлення, растеризацію трикутника на блоки по 32 піксели; блоки, що запускають на виконання програми даних різних форматів: вершинні (Vertex Thread Issue), геометричні (Geometry Thread Issue) і піксельні (Pixel Thread Issue).

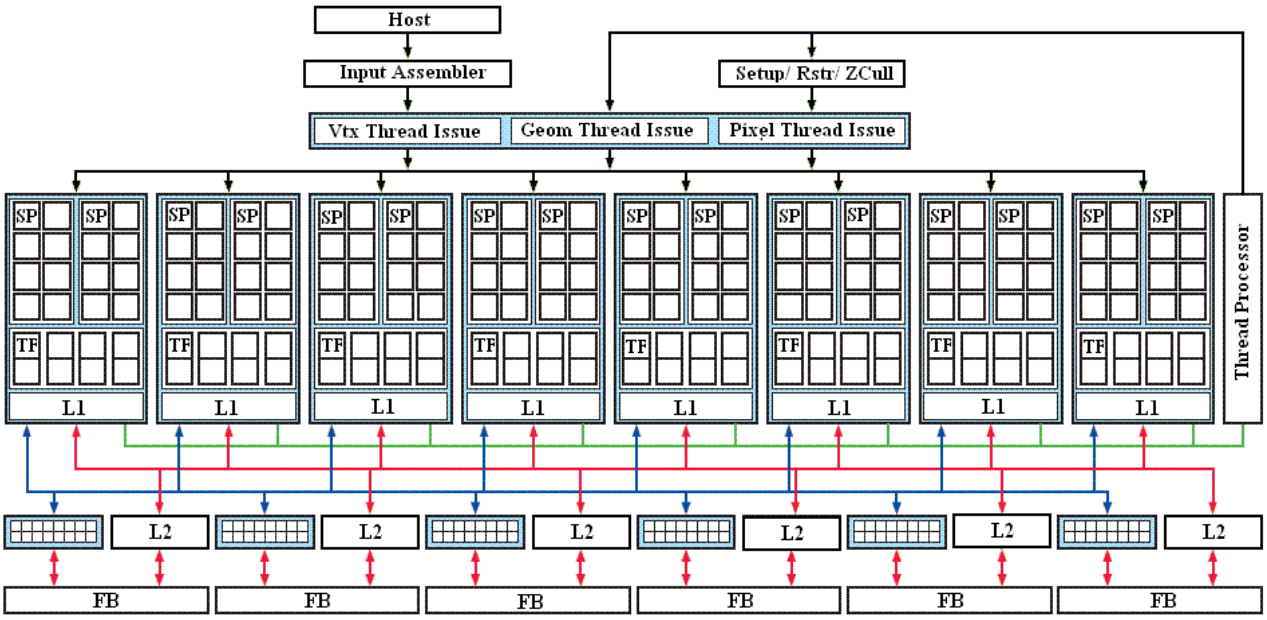

Рисунок 4 – Структура графічної відеокарти GeForce 8800

АЛП у потокових процесорах фірм Nvidia і AMD нерівнозначні. На рисунку 5 зображено структуру шейдерного процесора відеокарти ATI Radeon HD 3800 фірми AMD.

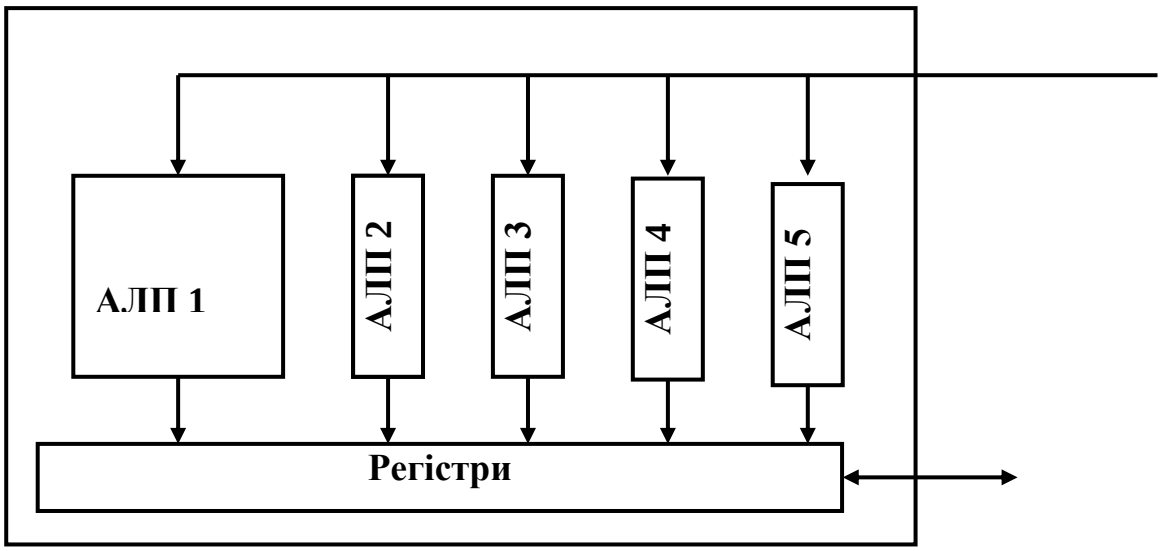

Рисунок 5 – Структура шейдерного процесора відеокарти ATI Radeon HD 3800

Один із п’яти АЛП може виконувати такі спеціальні функції, як синус, логарифм, експонента й т.д. Інші чотири АЛП виконують операції додавання-множення. Кожний із суперскалярних процесорів додатково має блок розгалужень, що підвищує ефективність роботи на шейдерах із великим числом переходів. Суперскалярна архітектура досягає найбільшої ефективності тоді, коли всі АЛП зайняті виконанням незалежних операцій, а домогтися цього досить складно, тому що в 3D-додатках багато операцій залежить від результатів виконання попередніх. Саме тому графічним ядрам ATI Radeon HD для досягнення найкращих результатів потрібна ретельна оптимізація драйверів під конкретний додаток.

Найновішими архітектурами відеокарт фірми NVIDIA є Fermi, Kepler та Maxwell.

Архітектура CUDA з кодовою назвою «Fermi»[3] (рис. 6) – це одна з найновіших архітектур. Більше трьох мільярдів транзисторів і 512 ядер CUDA дозволяють архітектурі Fermi забезпечувати суперобчислення і високу продуктивність.

NVIDIA GF100 (GT300) – 40-нм графічний процесор (GPU), розроблений корпорацією NVIDIA, перший представник лінійки GeForce 400. До нововведень чіпа відносяться дію за схемою Multiple Instructions

Multiple Data, підтримка ECC, перехід на 64-розрядні регістри відеопам'яті, підтримка технологій DirectCompute, OpenCL, що дозволяють проводити обчислення на GPU, тому NVIDIA Fermi можна віднести до розряду General-Purpose Graphics Processing Unit. Чіп NVIDIA GF100 має 512 суперскалярной шейдерними процесорами (або ядрами CUDA, як називає їх NVIDIA) і 3 мільярдами транзисторів.

За оцінками NVIDIA чіп показує 400% приріст продуктивності в обчисленнях з подвійною точністю в порівнянні з попереднім поколінням продукції компанії.

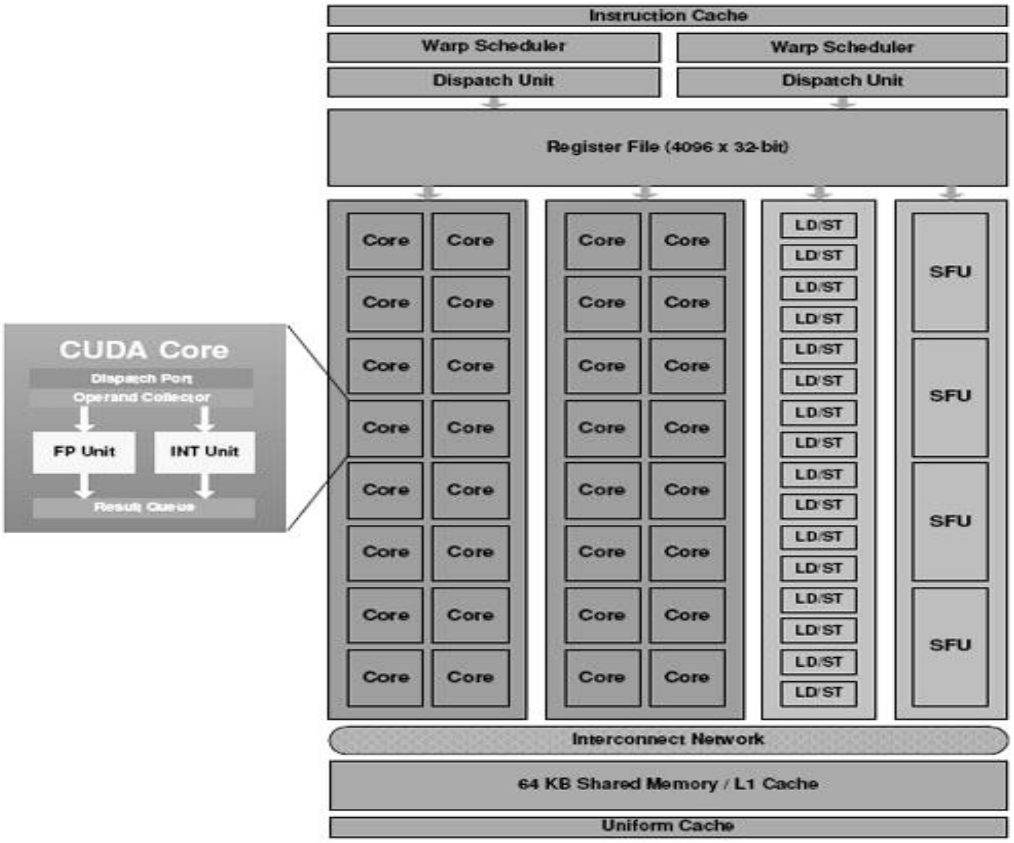

Рисунок 6 – Архітектура Fermi

Архітектура Fermi [3] забезпечує більш широке застосування гетерогенних обчислень на GPU і CPU, підтримуючи повний спектр обчислювальних додатків. Вбудовані можливості роботи з C ++ і сумісність з середовищем розробки Visual Studio роблять паралельне програмування з Fermi ще простіше і дозволяють досягати неймовірно

високої швидкості роботи, включаючи значне прискорення трасування променя, обробки фізики, аналізу кінцевих елементів, високоточних наукових обчислень, роботи з розрідженими матрицями в лінійній алгебрі, сортування та пошукових алгоритмів.

Досить великий проміжок часу компанія NVIDIA не вносила зміни у графічну архітектуру своїх топових рішеннях. Модель GeForce GTX 680, випущена навесні 2012 року, стала першою відеокартою з архітектурою Kepler, потім з'явилися і більш потужні рішення GTX 780 (Ti) на базі потужніших GPU тієї ж архітектури.

Фахівці NVIDIA виділяють 3 ключових нововведення Kepler [4] порівнянно з Fermi: SMX (є і в GeForce, і в Tesla), Hyper-Q і Dynamic Parallelism (тільки в Tesla GK110).

SMX (Streaming multiprocessor) - це новий обчислювальний модуль, що прийшов на зміну SM (Fermi). Оскільки енергоспоживання вже давно є головною проблнмою постачальників процесорів і одним з головних обмежень для збільшення продуктивності, то при проектуванні Kepler інженери компанії орієнтувалися на максимізацію співвідношення Продуктивність / Споживана потужність.

І, дійсно, стверджується, що в метриці «Продуктивність / Ватт» Kepler виграє у Fermi в 3 рази. Кількість ядер CUDA на SMX становить 192 (було 32 на Fermi SM). Так що тепер топові відеокарти Kepler обладнані 8 SMX модулями з 1536 ядрами замість 16 SM з 512 ядрами для Fermi, що дає зростання абсолютної продуктивності також в 3 рази.

Під Hyper-Q розуміється можливість одночасного виконання декількох (до 32) завдань на GPU, запущених, наприклад, з різних CPU-процесів. Для Fermi користувач теж міг отримати доступ до однієї відеокарти з різних процесів і запустити кілька завдань одночасно.

Однак через те, що була тільки одна апаратна чергу для завдань, їх виконання відбувалося завжди послідовно. Наприклад, якщо завдання

завантажувала ресурси відеокарти на 20%, то решта 80% не використовувалися, хоча в черзі чекали інші завдання.

У Kepler з технологією Hyper-Q ситуація змінилася - тепер є підтримка 32 апаратних черг завдань, так що вони можуть бути запущені з сьогоденням паралелізмом. Якщо один з них використовує ресурси відеокарти не повністю, то драйвер запускає на виконання завдання з іншої апаратної черги, що корисно для великої кількості невеликих завдань.

Головним винаходом в Kepler, найбільш цікавим для програмістів і розробників алгоритмів, є Dynamic Parallelism - можливість створювати обчислювальні потоки (threads) всередині вже створених потоків без передачі управління назад у CPU.

Важливість цього нововведення стане зрозумілою, якщо згадати про деревоподібну структуру величезної кількості алгоритмів обчислювальної та дискретної математики. Для Fermi було необхідно завершувати потоки, повертати управління на CPU, створювати нові і т.д., або істотно видозмінювати сам алгоритм. Це не тільки додавало накладні витрати і знижувало підсумкову ефективність коду, але й збільшувало час розробки.

У цілому Kepler [5] цілком можна назвати повністю переробленої архітектурою, яка продовжує тенденції оптимізації ефективного виконання складних обчислювальних задач, а також має дуже швидку обробку геометрії та тесселяции.

Як і у випадку з Fermi, новий GPU має у своєму складі кілька блоків GPC (кластери графічної обробки - Graphics Processing Clusters), які є незалежними пристроями в складі відеочипа, здатними працювати самі як окремі пристрої, так як в їх складі

Є всі необхідні власні ресурси: растерізатор, геометричні двигуни і текстурні модулі. Тобто, більшість функціоналу виконується всередині блоків GPC. Блок-схема GK104 виглядає так.

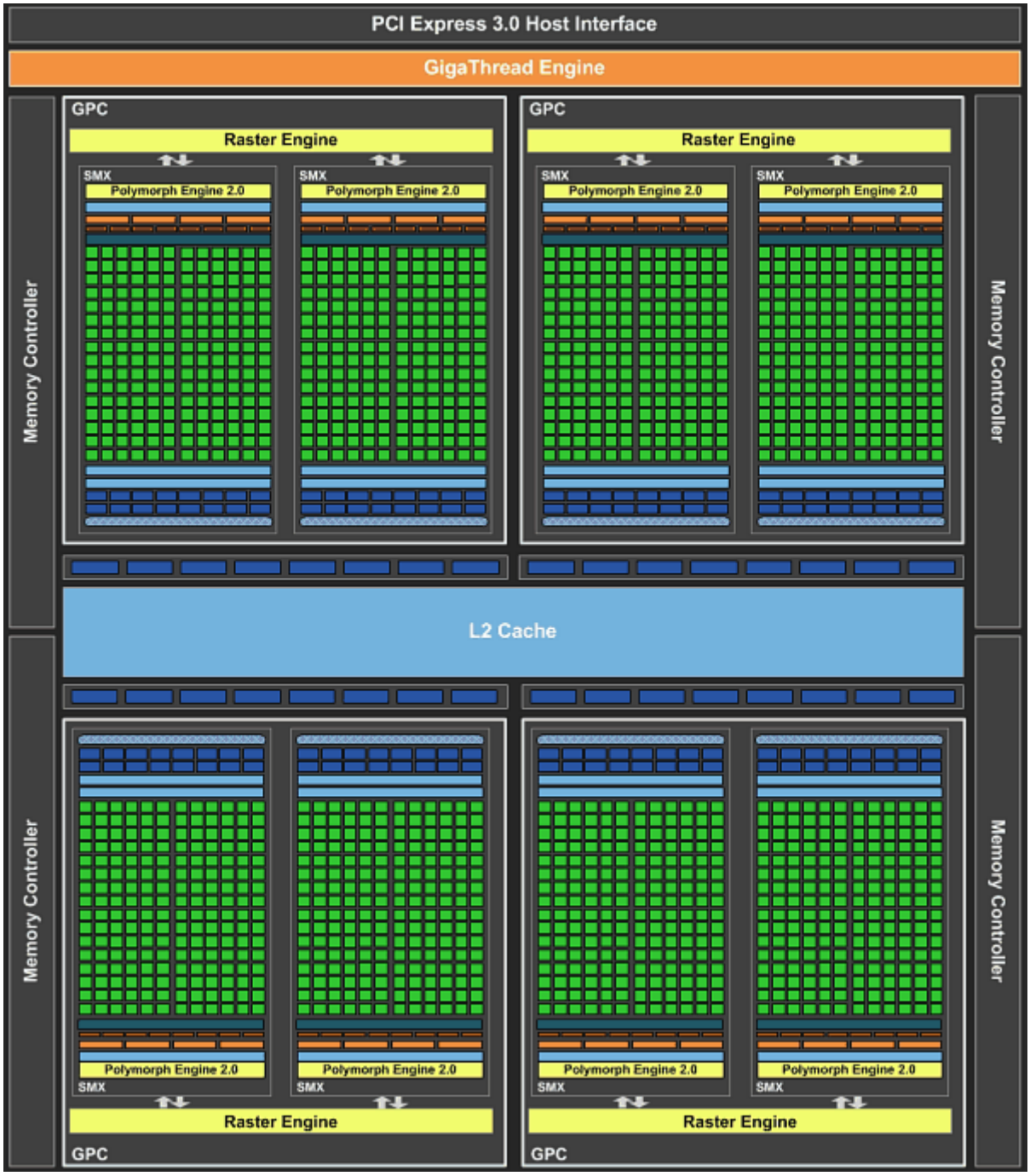

Рисунок 7. Архітекура GK104

Новий GPU має чотири блоки GPC, як і попередній топовий чіп GF100 / GF110, але на відміну від них, кожен з цих блоків містить по два потокових мультипроцессора, що відрізняються від того, що мали місце у всіх попередніх чіпів NVIDIA. Нове рішення використовує наступне покоління потокових мультипроцессоров (Streaming Multiprocessor), які тепер називаються SMX (рис.8), на відміну від SM в попередніх чіпах.

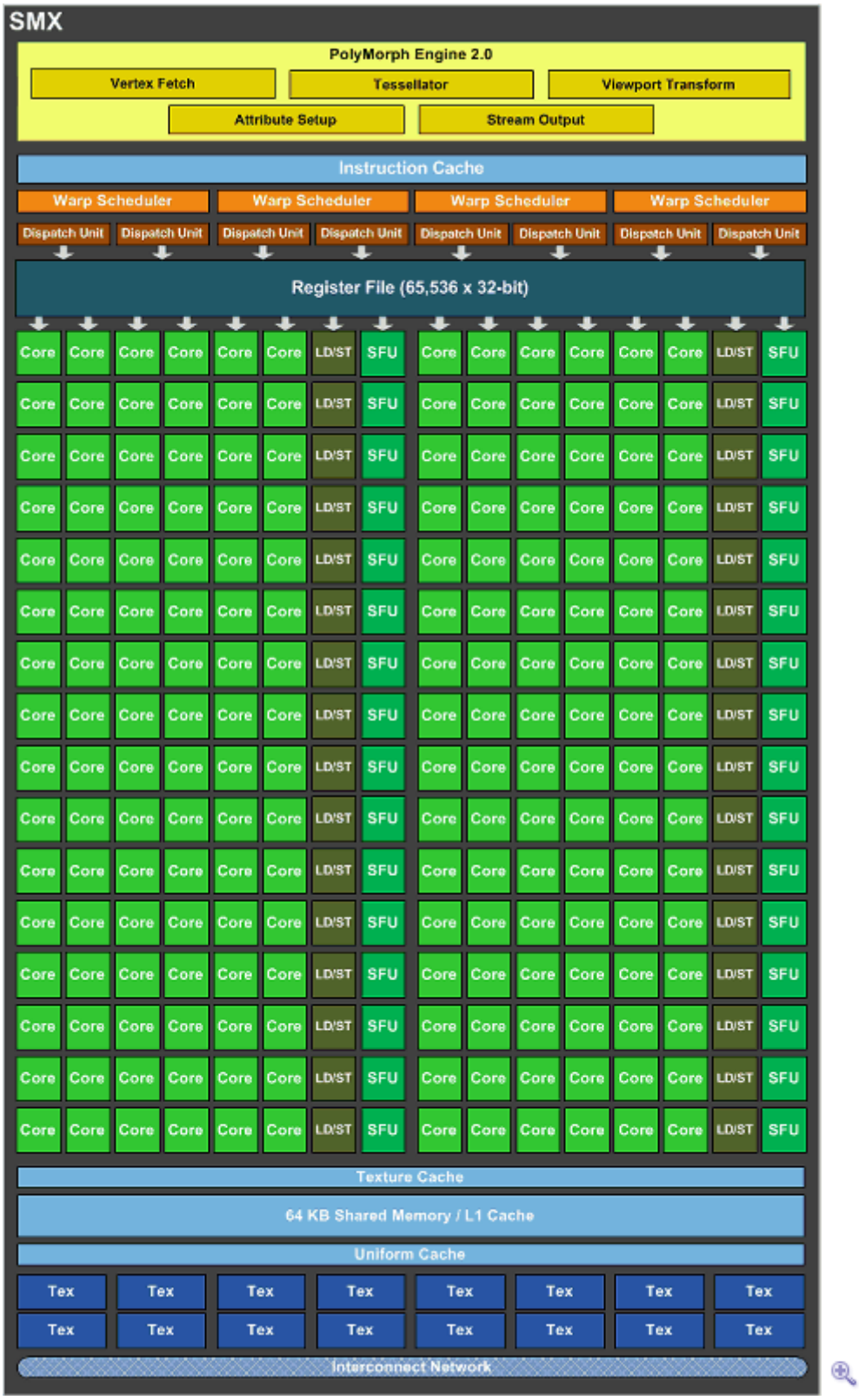

Рисунок 8. Архітектура SMX

Мультіпроцесори - це основна складова частина GPU компанії NVIDIA, і саме вони зазнали найбільше змін до Kepler. Порівнянно з попередніми SM, нові SMX забезпечують більш високу продуктивність,

що видно по кількості функціональних пристроїв у складі SMX, але при цьому споживають значно менше енергії.

Велика частина ключових блоків GPU включена до складу SMX: потокові процесори (CUDA Cores) виконують всі математичні операції над пікселями, вершинами і займаються неграфічних обчисленнями, текстурні модулі (TMU) фільтрують текстурні дані, завантажують і записують їх у відеопам'ять, блоки спеціальних функцій (Special Function Units, SFU) виконують складні операції (обчислення синуса, косинуса, квадратного кореня і т.п.) і інтерполяції графічних атрибутів. Ну а двигун PolyMorph забезпечує вибірку вершин, займається тесселяцію, перетворенням в екранні координати, установкою атрибутів і потоковим виведенням (stream output).

Блоки LSU використовуються для передачі даних з / в кеш і пам'ять, що розділяється, що може негативно позначитися на завданнях GPU обчислень. Втім, зменшення кількості LSU не повинно значно вплинути на продуктивність в графічних застосуваннях.

Найсучаснішими є архітектури Nvidia MAXWELL і Nvidia VOLTA.

Архитектура Nvidia MAXWELL характеризується уніфікованою віртуальною пам’яттю.

Вперше GPU на базі архітектури Maxwell можуть динамічно візуалізувати відбите світло, використовуючи нову технологію VXGI (воксельна глобальна ілюмінація). Сцени будуть виглядати значно більш натурально, так як світло взаємодіє в ігровому середовищі більш реалістично.

Буде реалізований багато кадровий антиаліайзинг (MFAA)

Ігровий процес в графічно насичених іграх означає вибір між високими настройками або високою частотою зміни кадрів з низькими настройками.Графічні процесори GeForce GTX 980 і 970 підтримують ексклюзивну технологію MFAA, яка забезпечує і те, й інше: збільшуючи

продуктивність у порівнянні з відеокартами попереднього покоління, дозволяє грати в високому дозволі з високими FPS.

Графічні процесори GeForce GTX 980 і 970 забезпечують необхідну потужність для рендеринга зображень у дозволі 4K. Вони використовують технологію DSR, яка покладається на просунуті фільтри для масштабування зображення, забезпечуючи геймерам ігровий процес в 4K навіть на 1080p моніторах. Кожна гра автоматично оптимізується за допомогою утиліти GeForce Experience ™ без зменшення продуктивності.

Забезпечуючи високу продуктивність з малими затримками, графічні процесори на базі архітектури Maxwell представляють собою нове покоління графічних рішень для створення захоплюючої і плавною віртуальної реальності.

GTX 980 і 970 є найшвидшими відеокартами, представляючи собою ідеальне рішення для дисплеїв з високою роздільною здатністю 4K і 4K Surround. Висока продуктивність плюс ексклюзивні технології, такі як NVIDIA G-SYNC ™ і захоплення відео в 4K за допомогою NVIDIA ShadowPlay ™ , означають, що отримано найбільш просунутий ігровий процес в 4K.

Архітектура Nvidia VOLTA буде впроваджена в 2016-му році.

У графічних процесорах Volta, які підуть за Maxwell, чіпи пам'яті будуть розташовані в одному корпусі з кристалом GPU (багатокристальна компоновка). Таке взаємне розташування пам'яті і GPU дозволить збільшити пропускну здатність з'єднання між ними до 1 ТБ / с. Для прикладу: щоб передати всі дані з повністю записаного диска Blu-Ray по такому з'єднанню досить 1/50 с. Таким чином, можна констатувати, що архітектури графічних відеокарт стрімко розвиваються у напрямку підвищення продуктивності і зменшення споживаної потужності.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. NVIDIA's GeForce 7800 GTX graphics processor . [Електроний ресурс] Режим доступу:

http://techreport.com/review/8466/nvidia-geforce-7800-gtx-graphics-processor

2. Революция в мире графических процессоров [Електронний ресурс] Режим доступу:http://compress.ru/Article.aspx?id=16963.
3. Програмно-апаратна платформа CUDA, архітектурний ряд Fermi, програмування за допомогою засобів програмної платформи CUDA SDK. Технологія GPGPU [Електронний ресурс] Режим доступу: http://knowledge.allbest.ru/programming/2c0a65635a2ad68a4d53a89521306c36_0.html.
4. Технічні особливості архітектури Kepler [Електронний ресурс] Режим доступу: http://isicad.ru/ru/articles.php?article_num=15312.
5. NVIDIA GeForce GTX 680Новый однопроцессорный лидер 3D-графики. [Електронний ресурс] Режим доступу: http://www.ixbt.com/video3/gk104-part1.shtml.
6. MAXWELL – Архитектура GPU нового поколения. [Електронний ресурс] Режим доступу: http://www.nvidia.com.ua/object/maxwell-gpu-architecture-ru.html.
7. За архитектурой NVIDIA Maxwell последует Volta, объединяющая GPU и
память в одной микросхеме, [Електронний ресурс] Режим доступу http://www.ixbt.com/news/hard/index.shtml?16/63/96.