УДК 004.92

Романюк О. Н.
Захарчук М. Д.

ЕВОЛЮЦІЯ РОЗВИТКУ АРХІТЕКТУР ВІДЕОКАРТ

Вінницький національний технічний університет

Анотація
У даній статті розглянуто основні етапи процесу еволюції архітектур відеокарт, розглянуто особливості архітектур кожного покоління.
Ключові слова: *відеокарта; архітектура; інновації; графіка; відеопам'ять.*

Abstract
This article presents the basic stages of the process of evolution of architectures of video cards are considered, features of architectures of each generation are also considered.
Key words: *video card; architecture; innovations; memory.*

Вступ

Комп'ютерна графіка динамічно розвивається у напрямку візуалізації тривимірних зображень у режимі реального часу. Її засоби допомагають вирішувати широке коло завдань інтерактивного проектування, автоматизованого навчання, контролю технологічних параметрів, теоретичних і прикладних досліджень. Сьогодні для формування зображень використовують відеокарти [1-14], які стали одним із ключових компонентів обчислювальних систем. Тенденція до подальшого ускладнення графічних сцен, збільшення рівня деталізації поверхонь для коректної апроксимації об'єктів реального світу, використання більш складних моделей освітлення та зафарбовування вимагає збільшення продуктивності графічних процесорів. Різні вимоги до швидкості роботи графічного процесора залежно від виконуваних обчислювальною системою задач стимулюють виробників відеокарт розробляти графічні процесори та відеоадаптери з різними специфікаціями та характеристиками для максимального задоволення потреб предметної галузі.

Мета роботи: проаналізувати особливості архітектур побудови відеокарт.

ВИКЛАД ОСНОВНОГО МАТЕРІАЛУ

Перші архітектури відеокарт

Перші архітектури відеокарт. MDA (Monochrome Display Adapter) [1] – працювала тільки в текстовому режимі та підтримувала п'ять атрибутів тексту: звичайний, яскравий, інверсний, підкреслений і миготливий. Ніякої колірної або графічної інформації вона передавати не могла, і те, якого кольору будуть літери, визначалося моделлю монітора. Архітектура не підтримувала роботу із окремими пікселями та складалася з ядра відеоадаптера, яким служив чіп Motorola MC6845, обсяг пам'яті становив 4 Кбайт. Максимальна роздільна здатність становила 720x350 пікселів.

Першою кольоровою відеокартою стала CGA (Color Graphics Adapter), випущена компанією IBM, яка стала основою для подальших стандартів відеокарт. Вона могла працювати або в текстовому режимі, або в графічному (рис.1). [2] Архітектура CGA, на відмінно від MDA, підтримувала роботу із окремими пікселями та складалася з ядра відеоадаптера, яким служив все той же чіп Motorola MC6845, але обсяг пам'яті був більше в чотири рази , тобто 16 Кбайт. У режимі роботи з кольоровою графікою максимальний дозвіл становило 320x200 пікселів, з монохромним - 640x200 пікселів. Глибина кольорів адаптера становила 4 біта. Це дозволяло використовувати палітру з 16 кольорів.

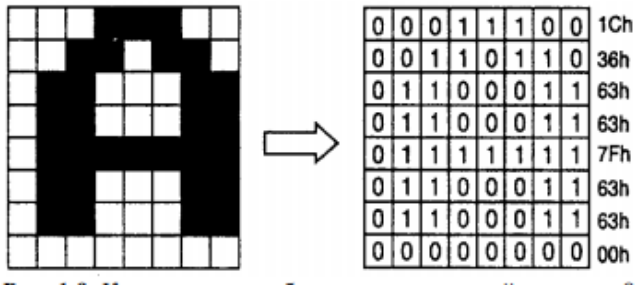

Рис.1 –Графічний режим відеокарти CGA

Логічним продовженням MDA і CGA стало теж рішення IBM під назвою EGA (Enhanced Graphics Adapter), представлене у вересні 1984 року. За своєю суттю новий відеоадаптер[2] став першим у своєму роді рішенням, здатним відтворювати нормальне кольорове зображення. Так само як і CGA, EGA підтримував текстовий і графічний режими, при цьому максимальний дозвіл становив 640x350 пікселів при використанні 16 кольорів з 64 можливих.

На архітектурному рівні EGA був схожий зі своїми попередниками: він також використовував видеоконтроллер Motorola MC6845, оснащувався збільшеним об'ємом пам'яті, рівним 256 Кбайт. Для передачі даних застосовувалася шина ISA. Вся пам'ять поділялась на 4 сегменти. Графічний процесор [2] був спроможний заповнювати сегменти паралельно, що значно підвищило швидкість заповнення кадру. Також, адаптер додатково оснащувався 16 Кбайт пам'яті для розширення графічних функцій BIOS і мав проміжний буфер.

У подальшому фірма IBM розробила VGA [1] (Video Graphics Array) з розширенням EGA. Це фактичний стандарт відеоадаптера, який уособлював увесь розвиток відеокарт 80-х років. Особливістю VGA стало розташування основних підсистем на одній мікросхемі, що робило відеокарту більш компактною. Архітектура VGA [2] створила справжню революцію у розвитку відеокарт. Вона складалася із графічного контролера, який забезпечував обмін даними між ЦП і відеопам'яттю ; була вперше включена спеціальна мікросхема - RAMDAC (Random Access Memory Digital-to-Analog Converter - цифро-аналоговий перетворювач - ЦАП - даних, що зберігаються в ОЗУ) ; секвенсора, який перетворював дані відеокарти у потік бітів, який подавався на контролер, де відбувалося перетворення цих бітів на кольори; синхронізатора, який керував часовими параметрами та забезпечував високу швидкодію; контролера ЕПТ, що генерував сигнали синхронізації для дисплея. Відповідно кольорів стало більше і було створено нові графічні режими, так звані «X-режими» на 256 кольорів, із збільшеною роздільною здатністю.

Отже, у 80-х роках відеокарти тільки почали зароджуватися. Розробники розвивали їхні архітектури та збільшували можливості відеокарт. Почали використовуватися нові технології (ЦАП, контролер) та збільшилась функціональність відеокарт, а саме використання адаптерів для генерування кольорових зображень. Загальний вигляд архітектури відеокарт кінця 80-х років наведено на рисунку 2.

Відеокарти 90-х років

Графічний користувальницький інтерфейс, що з'явився в багатьох операційних системах, стимулював новий етап розвитку відеоадаптерів. З'являється поняття «графічний прискорювач» (graphics accelerator). [1] Це відеоадаптери, які виробляють виконання деяких графічних функцій на апаратному рівні. До числа цих функцій належать, переміщення великих блоків зображення з однієї ділянки екрана в інший, заливання ділянок зображення, малювання ліній, дуг, шрифтів, підтримка апаратного курсору і т. п. Прямим поштовхом до розвитку настільки спеціалізованого пристрою стало те, що графічний користувальницький інтерфейс безсумнівно зручний, але його використання вимагає від центрального процесора чималих обчислювальних ресурсів, і сучасний графічний прискорювач якраз і покликаний зняти з нього частку обчислень з остаточного висновку зображення на екран.

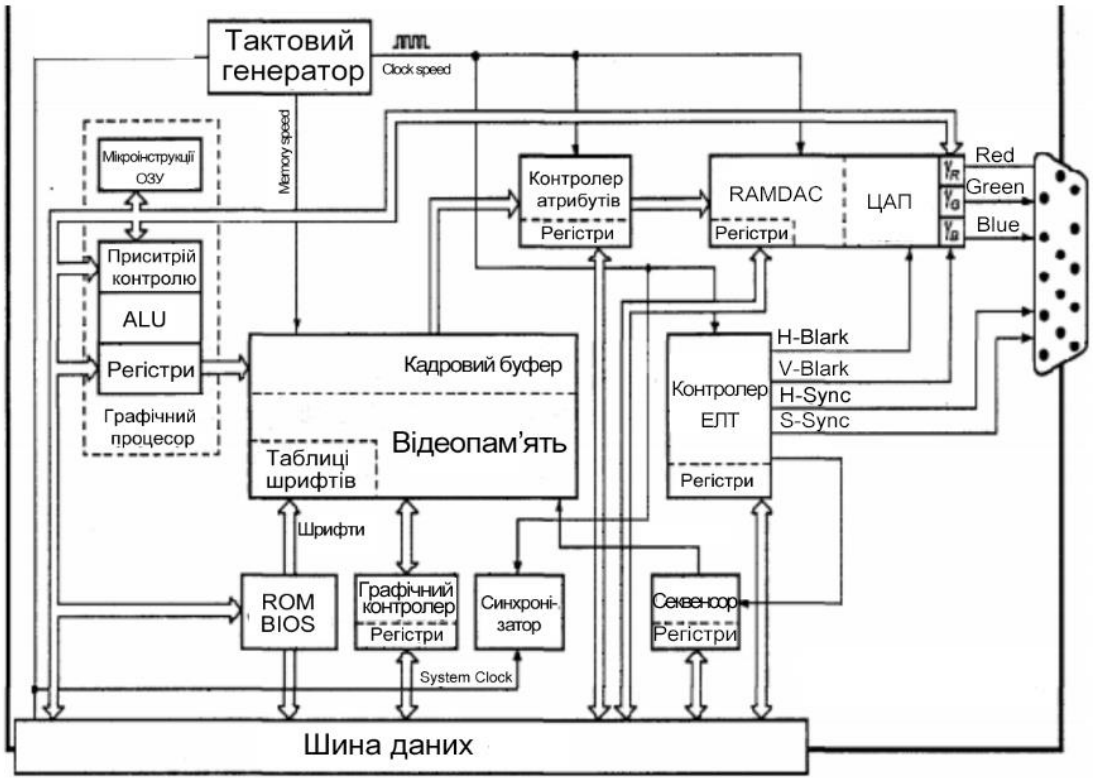

Рис.2 – Зображення архітектури відеокарт 80-х років

У 1991 року з'явилося поняття SVGA [2] (Super VGA) - розширення VGA з додаванням більш високих режимів і додаткового сервісу, наприклад можливості поставити довільну частоту кадрів. Число одночасно відображаються збільшується до 24 біта, з'являються додаткові текстові режими. З сервісних функцій з'являється підтримка розширення BIOS стандарту VESA. З його допомогою програмісти могли визначати специфічні відповідності та використовувати їх в подальшому. При цьому для роботи з будь-яким SVGA-пристроєм використовувався єдиний драйвер.

Відмінною рисою рішень SVGA [2] став вбудований акселератор, поява якого була пов'язана з необхідністю якісної обробки графічної складової нових ОС. Але цього все ще не було достатньо, адже активно розвивалася тривимірна графіка, яка потребувала значно більшого функціоналу від апаратного забезпечення(GPU). Процес утворення зображення показано на рисунку 3.

Перший вдалий 3D-акселератор для [4] масового ринку, Voodoo Graphics (1996 рік), був двочіповим рішенням. Один чіп, TexelFX, являв собою один простий текстурний блок, що завантажував чотири текселі і виконував білінійну інтерполяцію між ними за один такт. Інший чіп, PixelFX, був простим блоком растеризації (ROP), що виводить один піксель за такт. Також з'явився програмний інтерфейс Glide, який був створений на скорочені й бібліотеці OpenGL і використовувався у Voodoo1 та забезпечував роботу лише з [13] трьохвимірною графікою. Процес утворення 3D зображення показано на рисунку 3.

Рис.3 – 3D конвеєр першого покоління

У Voodoo 2 і 3 був [4] додано другий текстурний блок, що дозволило застосовувати кілька більш складних ефектів, накладаючи до двох текстур на піксель за такт або виконувати трилінійну

фільтрацію. Також було додано 2D ядро для обчислення двовимірних операцій. Процес формування 3D зображення показано на рисунку 4. Для інших операцій конвеєру використовувався центральний процесор. В ролі інтерфейсу виступали Direct3D та OpenGL.

Рис.4 – 3D конвеєр другого покоління

Nvidia Riva TNT – [2] відповідь компанії Nvidia на Voodoo. Покращена архітектура дозволяла здійснювати у два рази більше операцій та обчислень, було покращена якість зображення та швидкість обрахунку та створення трьохвимірних сцен, впроваджено нову технологію фільтрації текстур. Новий чіп, який має кодову назву NV4, був виготовлений по 350-нм технологічному процесу, містив 7 мільйонів транзисторів, а його частота склала 90 МГц. Як чіп [8] пам'яті використовувалися модулі SDRAM, їх сумарний обсяг становив 16 Мбайт. Частота пам'яті дорівнювала 110 МГц, а ширина шини пам'яті - 128 біт. RIVA TNT підтримувала більше функцій: наприклад, 32-бітний колір і текстури з дозволом 1024x1024 точок. Також відеокарта отримала підтримку технології Twin-Texel, яка дозволяла накладати дві текстури на один піксель за такт в режимі мультитекстурування.

Отже, 1990-і породили велику кількість компаній, що виробляють дискретні відеокарти. Основною інновацією архітектур відеокарт 90-х стала можливість виконання Twin-Texel обчислень на GPU. Крім того, важливим нововведенням стала поява мультитекстурування, що дало можливість накладання в реальному часі карт висот, карт освітлення та інших. Також стало можливим використання 2 ядер (3D та 2D), які дозволили перенести основну масу обчислень графіки на відеокарту. Загальний вигляд архітектури відеокарт кінця 80-х років наведено на рисунку 5.

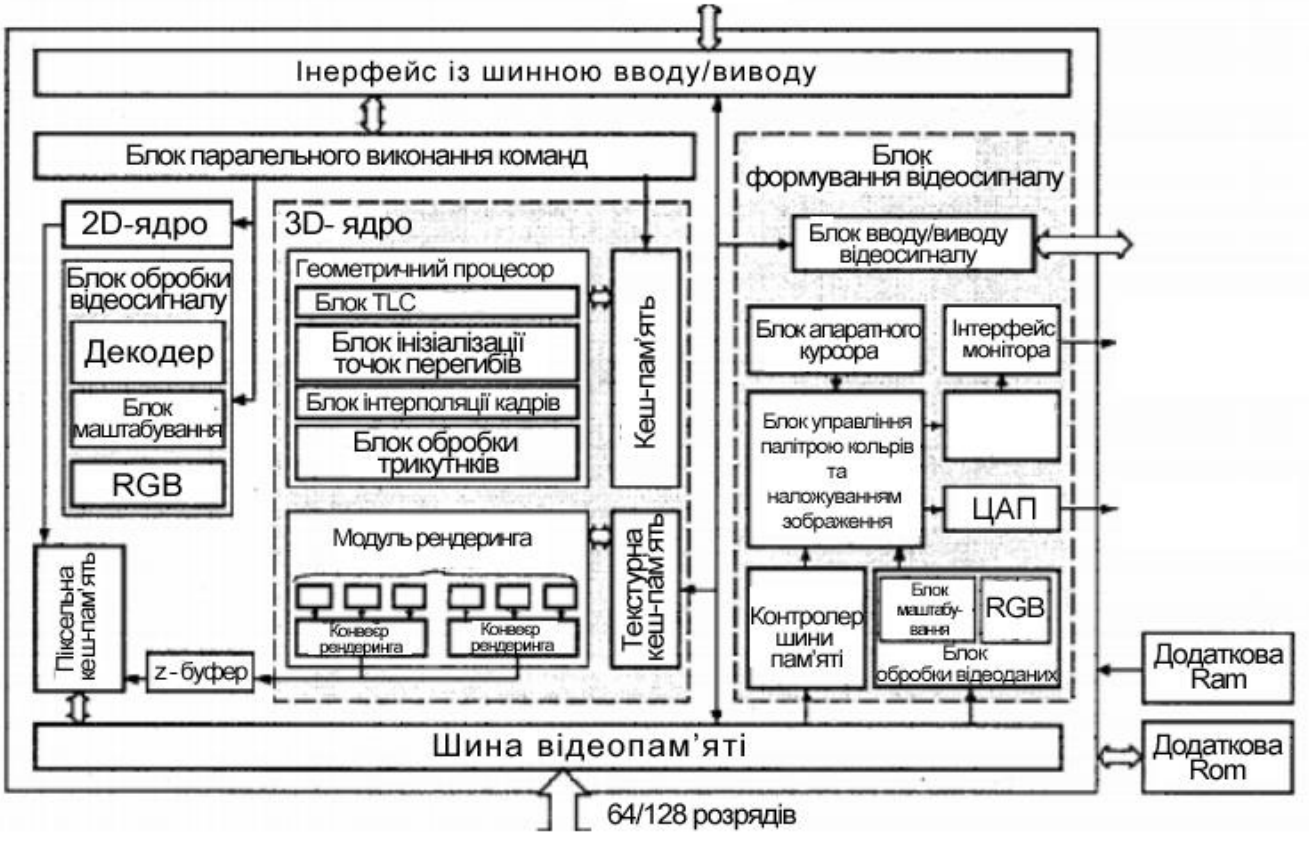

Рис.5 – Архітектура відеокарт 90-х

Відеокарти 2 тисячоліття

NVIDIA GeForce 7800 [3] – справжня революція у архітектурі відеокарт, а саме перехід до уніфікованих потокових вершинних та шейдерних процесорів.

У відеокарті GeForce 7800 реалізовано графічний конвеєр із [3] використанням вершинних і пікселних (шейдерних) процесорів. Відеокарта має 24 пікселних процесори PS, по одному текстурному блоці на конвеєр, 8 вершинних процесорів і 16 блоків растрових операцій (ROP) (рис. 6). Пікселні процесори згруповано по 4 для обробки квадів. Процесор PS має два векторні АЛП, здатні виконувати 2 різні операції над 4 компонентами та два міні-АЛП (найпростіші скалярні АЛП для виконання простих операцій). Кожний пікселний 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. За один такт вершинний процесор може виконати одну векторну операцію, одну скалярну операцію й здійснити один доступ до текстури.

У відеокарті GeForce 8800 вперше [3] використано уніфіковану шейдерну архітектуру рендерингу, потокове оброблення інформації та новий вид шейдера – геометричний. Чіп (рис. 7) 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 [9] ROP визначать факт видимості, запис у буфер кадру й мультисемплінг. Вони згруповані з контролерами пам'яті, чергами запису та кешем другого рівня L2. Потокові процесори [8] SP є уніфікованими скалярними процесорами із плаваючою комою, що обробляють не тільки графічні, але й інші дані [5]. 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 [3] 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).

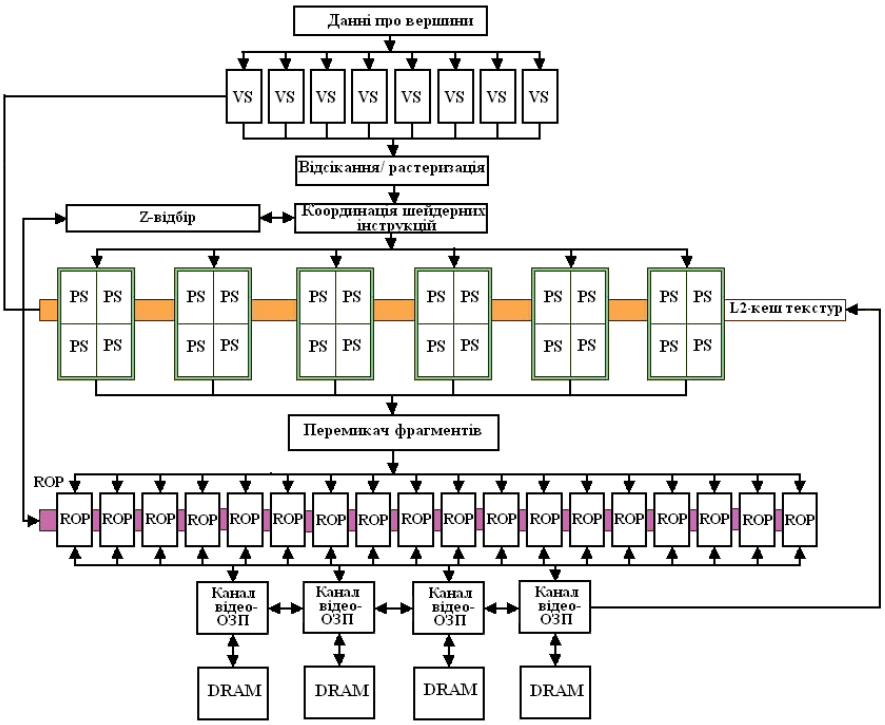

Рис.6 – Архітектура ранніх відеокарт другого тисячоліття

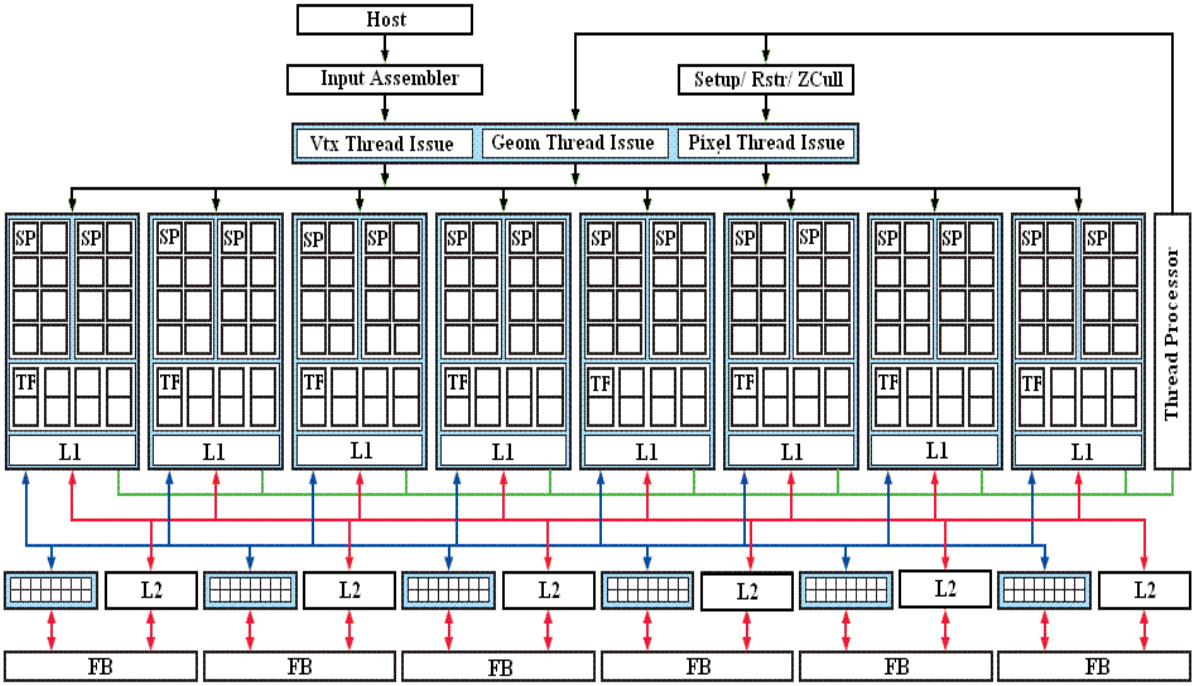

Рис.7 – Архітектура GeForce 8800

Отже, характерними особливостями архітектур відеокарт 2000-х років є перехід на вершинні та шейдерні процесори, з подальшим утворенням геометричних процесорів, збільшення кількості та швидкості обчислень та покращення якості картинки на виході.

Перехід до сучасних архітектур відеокарт

Архітектура CUDA з кодовою назвою «Fermi» (рис. 8) – це одна з найкращих архітектур минулого десятиліття. Більше трьох [3] мільярдів транзисторів і 512 ядер CUDA(поєднання вершинних та шейдерних процесорів у одне ядро) дозволяють архітектурі Fermi забезпечувати суперобчислення і високу продуктивність.

NVIDIA GF100 (GT300) [5] – 40-нм графічний процесор (GPU), розроблений корпорацією NVIDIA, перший представник лінійки GeForce 400. До нововведень чіпа відносяться дію за схемою Multiple Instructions. Регістри АЛП, 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 [5] показує 400% приріст продуктивності в обчисленнях з подвійною точністю в порівнянні з попереднім поколінням продукції компанії.

Отже, перехід до сучасних відеокарт став можливий із появою ядер CUDA, які контролювали процеси текстуризації, та потокових мультипроцесорів SM, що дозволили збільшити продуктивність, якість та зменшити енергозатратність.

Сучасні відеокарти

Нові відеокарти NVIDIA на [8] архітектурі Turing не просто привносять чергове підвищення продуктивності, вони несуть в собі ряд технологічних інновацій і є першими ігровими рішеннями, які підтримують трасування променів в реальному часі.

Нові чіпи використовують кластерну структуру, при цьому кожен такий кластер GPC (Graphics Processing Cluster) містить по 8 або 12 потокових мультипроцесорів SM. Крім традиційних ядер CUDA було введено нові 2 типа ядер: RT-ядра для розрахунку трасування променів в реальному часі і тензорні ядра для задач, пов'язаних зі штучним інтелектом. NVIDIA переробила архітектуру

потокових мультипроцесорів (streaming multiprocessors, або SM), і тепер кожен SM складається з 64 ядер CUDA, плюс до них додані 8 тензорних ядер і одне RT-ядро .

В основі нової архітектури лежить 12-нанометровий чіп FFN [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. Також присутній інтерфейс NVLink.

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. Архітектура Turing оснащена тензорними ядрами, які можуть забезпечити вищу продуктивність обчислень штучного інтелекту.

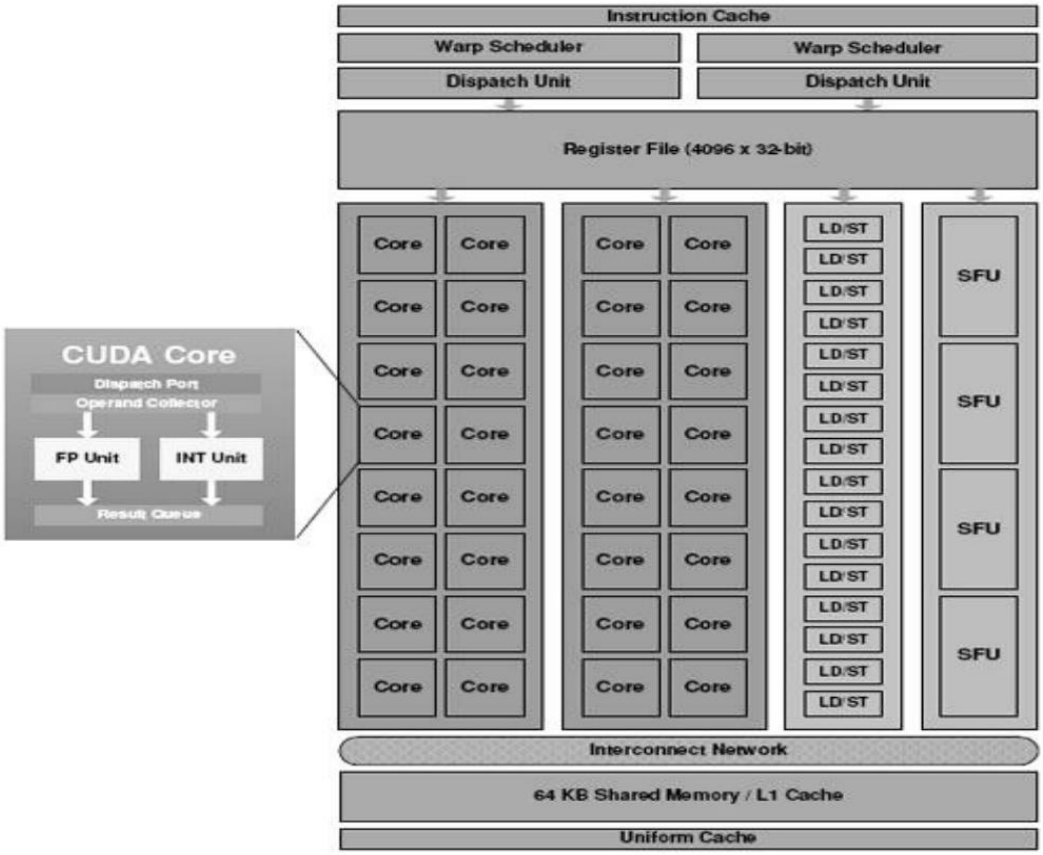

Рис.8 – Архітектура Fermi

Основним нововведенням в архітектурі Turing є 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, що є широковживаною структурою даних для зберігання об'єктів при трасуванні променів. RT-ядра прискорюють розрахунки руху світла та звуку в 3Dсередовищі.

Важливі зміни відбулися на рівні мультипроцесорних блоків SM, що мають стандартну структуру в усіх варіантах GPU Turing та дозволяють обчислювати 16 млн. операцій в секунду. Всі обчислювальні блоки всередині SM згруповано в чотири масиви обробки даних із логікою управління, що включає регістри, 1 планувальник на кожні 16 ядер і один порт диспетчера на кожні 16 ядер. При цьому в одному SM наявні 64 потокових процесори.

Більш досконалою є архітектура [11] Ampare. Завдяки покращеним тензорним та RT ядрам другого покоління, SM мультипроцесорам та збільшеною пропускною здатністю шини даних, ігрові рішення нової архітектури приблизно в півтора рази швидше аналогічних Turing в традиційних завданнях растеризації і до двох разів швидше при трасуванні променів.

Висновок

Проведений аналіз показав, що архітектури відеокарт змінювалися під впливом наукових інновацій. Їхній розвиток вплинув не тільки на архітектуру комп'ютера, а й на розвиток нових індустрій, технологій.

Список літератури

1. Відеоадаптери. Архітектура комп'ютерів [Електронний ресурс] – Режим доступу до ресурсу: http://archcom.ptngu.com/newtema_24.html
2. Еволюція дискретних відеокарт [Електронний ресурс] – Режим доступу до ресурсу: https://www.ferra.ru/review/computers/graphic-apapter-videocards-evolution-part-4.htm
3. Романюк О. Н. Аналіз архітектур відеокарт компанії NVIDIA/ О. Н. Романюк, Даньковська, С. І. Вяткін // Збірник матеріалів Міжнародної науково-практичної Інтернет-конференції "Електронні ресурси: створення, використання, доступ", Вінниця, грудень 2014 р. –2014. – С. 3-15.
4. Романюк О. Н. Аналіз тенденцій розвитку відеокарт / О. Н. Романюк, О. О. Дудник // Оптико-електронні інформаційно-енергетичні технології. - 2017. - № 2. - С. 114-119.
5. Революція в світі графічних процесорів [Електронний ресурс] – Режим доступу до ресурсу: https://compress.ru/article.aspx?id=169 https://compress.ru/article.aspx?id=16963.
6. Романюк О. Н. Класифікація графічних відеоадаптерів / О. Н. Романюк, Р. Ю. Довгалюк, С. В. Олійник // Наукові праці Донецького національного технічного університету. Сер. : Інформатика, кібернетика та обчислювальна техніка. - 2011. - Вип. 14. - С. 211-215.
7. Романюк О. Н. Високопродуктивні методи та засоби зафарбовування тривимірних графічних об'єктів. Монографія. // О. Н. Романюк – Вінниця : УНІВЕСУМ-Вінниця, 2006.
8. Романюк О. Н. Аналіз висопродуктивних відеокарт / О. Н. Романюк, С. О. Романюк, О. В. Поліщук // «Predni vedeske novinky-2013» : materialy IX mezinarodni vedecko-prakticka conference, 27 srpna-05 zari 2013 roku. – Praha, 2013. – С. 11-13
9. Романюк О. Н. Аналіз архітектур графічних відеокарт / О. Н. Романюк, Д. Т. Обідник, О. В. Поліщук, П. О. Величко // П'ята міжнародна науково-технічна конференція "Моделювання та комп'ютерна графіка", Донецьк, 24-27 вересня 2013 р. – Донецьк : ДонНТУ, 2013. – С. 132–138.
10. Вяткин С. И. Function-based GPU architecture /С. И. Вяткин, С. А. Романюк, С. В. Павлов, А. А. Дудник // Вимірювальна та обчислювальна техніка в технологічних процесах. - 2015. - № 1. - С. 139-144.
11. Романюк О. Н., Кательніков Д. І., Денисюк А. В., Захарчук М. Д. Аналіз архітектури AMPERE побудови відеокарт. The 6th International scientific and practical conference "Priority directions of science and technology development" (February 20-22, 2021), Kyiv, Ukraine. 2021. p 264-269.
12. Романюк, О. Н. Аналіз архітектури VOLTA відеокарт / О. Н. Романюк, Ю. О. Панфілова, А. Л. В. Чан // Інформаційні технології і автоматизація – 2018 : зб. доп. XI Міжнар. наук.-практ. конф., Одеса, 4–5 жовт. 2018 р. / Одес. нац. акад. харч. технологій ; ред. кол.: С. В. Котлик, В. А. Хобін. – Одеса, 2018. – Ч. II. – С. 13–15.
13. Вяткін С. І. Еволюція конвеєра рендерингу в відеокартах / С. І. Вяткін, О. Н. Романюк, О. О. Дудник. // Електронні інформаційні ресурси : створення, використання, доступ : збірник матеріалів Міжнародної науково-практичної Інтернет-конференції, м. Вінниця, жовтень 2016 р. – Вінниця, 2016. – С. 486-489..
14. Романюк О. Н. Застосування відеокарт для неграфічних обчислень / О. Н. Романюк, Р. Ю. Довгалюк, С. І. Вяткін, Д. Л. Благодир // Одинадцята Міжнародна науково-технічна конференція «ВОТТП-2011», 5-8 червня 2012 р. – Хмельницький, 2012. – С. 23-24.

Романюк Олександр Никифорович – доктор технічних наук, професор, завідувач кафедри програмного забезпечення, Вінницький національний технічний університет, м. Вінниця.

Захарчук Максим Дмитрович – студент групи 2ПІ-20б, Вінницький національний технічний університет, м. Вінниця.

Romanyuk N. Oleksandr - doctor of technical sciences, professor, head of the Software Department, Vinnytsia National Technical University, Vinnytsia.

Maksym D. Zakharchuk – student of 2SE-20b group, Vinnytsia National Technical University, Vinnytsia.