



Универзитет „Св. Кирил и Методиј“ во Скопје
**ФАКУЛТЕТ ЗА ИНФОРМАТИЧКИ НАУКИ И
КОМПЈУТЕРСКО ИНЖЕНЕРСТВО**

Семинарска по предметот:

Компјутерски компоненти и периферии

На темата:

Споредба на архитектурите на графичките картички од 2010 со 2011 година

Ментор:

Проф. Д-р Коста Митрески

Изработиле:

Дејан Арсиќ, 115005

Бојан Дикиќ, 115020

Содржина

СОДРЖИНА	2
ВОВЕД	3
КРАТОК ИСТОРИЈАТ	4
ОД КАДЕ ПОТРЕБАТА ЗА ГРАФИЧКИ КАРТИЧКИ?	4
АРХИТЕКТУРА НА RADEON HD 5800 (TERASCALE 2)	5
МЕМОРИСКИ КОНТРОЛЕР	6
AA (ANTI-ALIASING) И AF (ANISOTROPIC FILTERING) ПОДОБРУВАЊА	7
DIRECTX 11	8
ВИДЕО	8
АРХИТЕКТУРА НА RADEON HD 7000 (GNC)	9
ПЕРФОРМАНСИ	9
КВАЛИТЕТ НА СЛИКА	10
1. PRT (PARTIALLY RESIDENT TEXTURES)	10
2. ПОДОБРЕНО AF (ANISOTROPIC FILTERING)	11
3. УСОВРШЕНА ТЕСЕЛАЦИЈА НА DIRECTX 11	11
AMD ZERO CORE POWER TECHNOLOGY	11
АРХИТЕКТУРА НА FERMI GF100	12
ПЕРФОРМАНСИ	12
ТЕСЕЛАЦИЈА	13
ФИЗИЧКО ПРОЦЕСИРАЊЕ	13
ВИДОВИ FERMI АРХИТЕКТУРИ	14
АРХИТЕКТУРА НА KEPLER GK110	14
ПЕРФОРМАНСИ	15
1. ДИНАМИЧНА ПАРАЛЕЛНОСТ	15
2. HYPER-Q	15
3. GMU (GRID MANAGEMENT UNIT)	15
4. NVIDIA GPUDIRECT	16
ПОДОБРУВАЊЕ НА ТЕКСТУРАТА	17
L1, L2 И ЕСС КАЈ KEPLER АРХИТЕКТУРАТА	17
ЗА КРАЈ	18
ПРОФЕСИОНАЛНИ НАСПРОТИ ОБИЧНИ ГРАФИЧКИ КАРТИЧКИ	18
ДАЛИ NVIDIA ИЛИ AMD?	19
ЗАКЛУЧОК	20
КОРИСТЕНА ЛИТЕРАТУРА	21

Вовед

Графичка картичка (видео картичка, графички адаптер или графички акцелератор) претставува компонента на компјутерот која е наменета за обработка и за прикажување на визуелни податоци на соодветни излезни уреди (на пр. монитор). Графичката картичка може да се користи и за обработка на неграфички податоци, а во поново време на неа можат да се пренесуваат податоци кои припаѓаат на централните процесори. Современите графички картички вршат бројни функции од полето на компјутерска графика со што доаѓа до помало оптеретување на останатите делови од системот.

Графичките картички постојат во интегриран облик во склоп на матичната плоча, а во поново време сè почесто се јавуваат како посебна компонента. Интегрираните графички картички имаат мал капацитет на меморија и обично користат системска меморија, додека пак поновите модели имаат сопствена меморија која е посебно модифицирана и се користи само за графика. Речиси сите матични плочи имаат опција за исклучување на интегрираната графичка картичка и можност за примање на современа графичка картичка со високи перформанси, преку AGP, PCI и PCI-E магистралите.



Слика 1: GPU произведен од компанијата ATI

Краток историјат

Првите графички картички се конструирани од компаниите Matrox, Creative, S3 и ATI во 1995 година и тие можеле да произведуваат 3D слика. Подоцна во 1997 година 3dfx креирала нов графички чип Voodoo кој можел да произведува и некои 3D ефекти. За кратко време излегол Voodoo 2 со кој се појавиле и појаки чипови како TNT и TNT 2 од nVIDIA. Компанијата Intel започнала да работи на унапредување на начинот на поврзување на графичката картичка со матичната плоча, како резултат на што се појавила AGP магистралата со што се направила разлика помеѓу GPU и CPU. Од 1999 година приматот во производството на графички картички го завзема nVIDIA, која започнува да работи на унапредувањето на 3D алгоритмот и DDR технологијата, со што капацитетот на меморијата на графичките картички десеткратно се зголемил, од 32 на 128Mb.

Од каде потребата за графички картички?

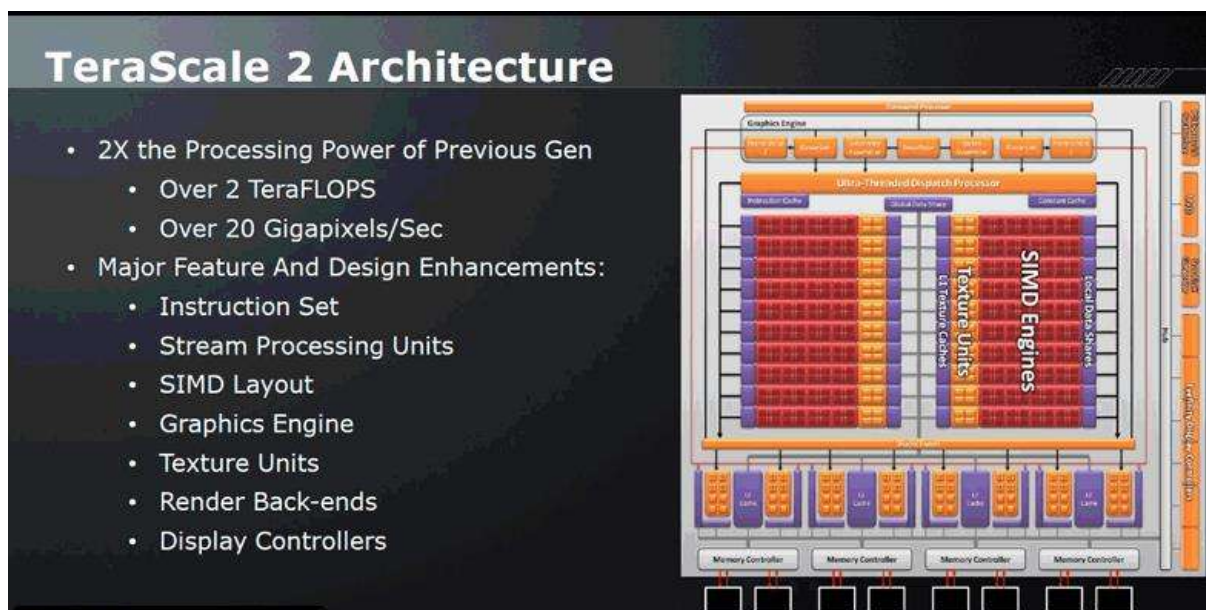
Човекот отсекогаш имал потреба за визуелно претставување на некои информации, денеска со развојот на технологијата речиси сè може да се претстави визуелно, т.е. компјутерски. Затоа графичките картички доаѓаат како голема помош на процесорот, во обработка на графички податоци (вектори, триангулација, сенки, отсјај...), со еден збор тие значително ја намалуваат работата на процесорите. Да земеме на пример 3D објекти кои сами по себе содржат голем број на вектори, пиксели, бои и друго. Доколку сите овие информации ги предадеме на главниот процесор би се соочиле со големи пристапи до меморија што би довело до латенција.



Слика 2: Kepler GK110 (од лево) и Radeon HD 7000 (од десно)

Архитектура на Radeon HD 5800 (TeraScale 2)

Новата серија на графички картички Radeon HD 5800 е многу повеќе од надоградување на добрите делови. TeraScale 2 архитектурата на серијата Radeon HD 5800 е еволуционерна надградба на серијата Radeon HD 4800. Нема некои поголеми промени во постоечката архитектура. Революционерните промени се присутни во претходната генерација TeraScale. TeraScale 2 е проширување на архитектурата, подобрување на перформансите и додавање на нови особини.



Слика 3: TeraScale 2 архитектура

Веројатно, особините и поддршката на DirectX 11 би биле revolucionерни доколку AMD ги поддржеше важните особини DirectX 11 како процесот на теселација (се користи во компјутерската графика со цел да се претстават сите објекти како мозаик од триаголници). Серијата Radeon HD 5870 има 2,7 терафлопи на компјутерска моќ во единечна прецизност и 544 гигафлопи во двојна прецизност. ATI ја додал можноста да изведува Co-issue MUL, зависен ADD во единечен clock циклус со дополнително SAD усовршување. Секој thread процесор содржи четири streaming процесорски единици и една специјална streaming процесорска единица, branch единица и регистри.

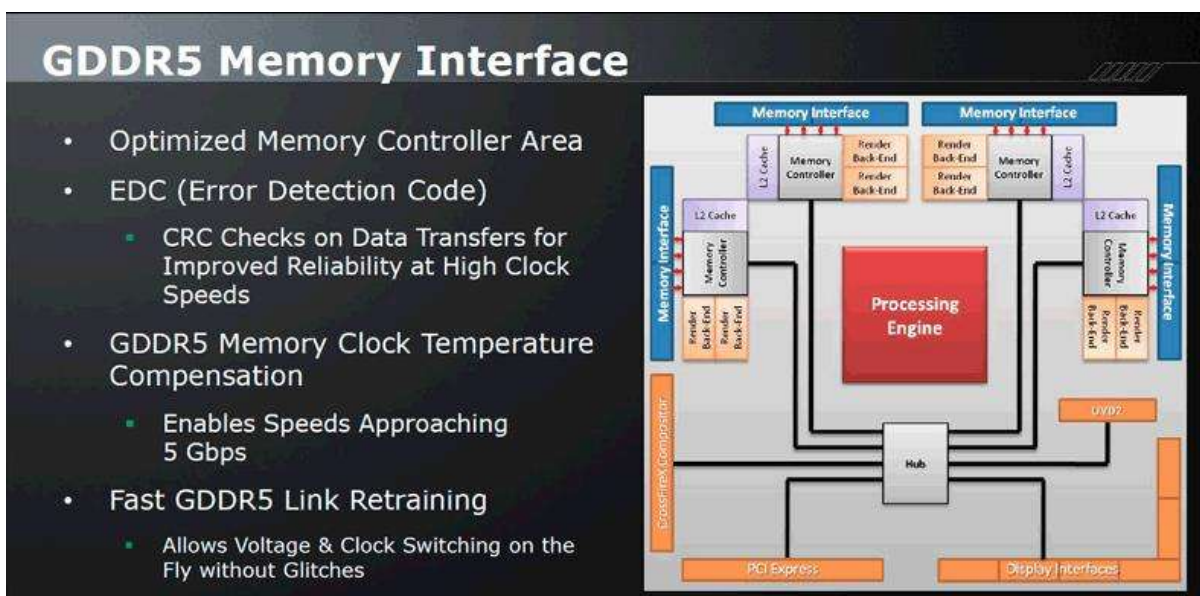
Заедно со надоградбата на thread процесорите ATI ги усовршил и единиците за текстура и кешот. Има 80 единици за текстура и HD 5870 може да изведува 68 милијарди билинеарни филтрирани тексели (основна единица за текстура во компјутерската графика) во секунда и дури до 272 милијарди 32-битни fetches (земање на инструкции) во секунда. Опсегот на L1 кеш е

подобрен и L2 кешот е удвоен на 128KB. Има уште неколку нови DirectX 11 текстурни особини, како што е 32-битен и 64-битен HDR (High-dynamic-range imaging) block compression modes.

Како што претходно е споменато, HD 5870 содржи целосна DirectX 11 теселација што денеска е позната како шеста генерација на ATI технологијата. Геометриското сенчање е побрзо и подобрен е квалитетот на OpenGL со поддржување на 12-битна суб-пикселна прецизност. OpenGL и OpenCL се целосно поддржани и усовршени. Исто така ROP (Render Output Unit) е дуплиран што го има неразделивиот ефект на елиминирање или пеглање на назабеноста кај графички прикази во 32-битна или 64-битна боја. Има и други подобрувања меѓу кои ATI најде начин да ги зголеми перформансите на CFAA (Computer Fraud and Abuse Act). CFAA ги користи stream процесорите за пресметка и не попречува во работата на меморискиот систем.

Мемориски контролер

Меморискиот контролер на Radeon HD 5800 серијата, исто така, добил некои оптимизации. ATI го вклучи и GDDR5 EDC (Error Detection Code) сличен на ECC (Error-correcting code) кој се наоѓа во DRAM (Dynamic random-access memory). Ова овозможува GDDR5 модулите да постигнат повисоки фреквенции кога ги редуцираат грешките. Ова може да помогне графичките картички да станат постабилни и посилни како целина.



Слика 4: GDDR5 мемориски интерфејс

Интересен резултат на овозможувањето на EDC е промената на последиците од overclocking. Во минатото, кога е преоптоварена меморијата на графичката картичка, се зголемува фреквенцијата сè додека картичката не

почне да исфрла артефакти (непосакувана промена на податоците) и потоа ќе прекине со работа. Но денеска кога е преоптоварена мемориската фреквенција, перформансите ќе се зголемат сè додека не ја постигне највисоката точка и EDC се вклучува за да ги поправи грешките во меморијата кои се должат на висока фреквенција. Ова значи дека EDC ќе ги поправи овие проблеми и перформансите ќе се намалат со зголемување на фреквенцијата и затоа EDC мора да работи повеќе за да ги поправи грешките. Со ова нема да се појавуваат артефакти и нема да има прекини во работата при преоптоварена меморија. Само ќе се забележи намалување на перформансите. Затоа целта е да се преоптовари меморијата додека перформансите не се намалат и потоа да се врати назад сè до моментот на намалување на перформансите и тоа е максималната преоптовареност на меморијата.

AA (Anti-Aliasing) u AF (Anisotropic filtering) подобрувања

AMD не ги само подобрија перформансите, туку направија и надградба на AA и подобрувања во квалитетот на AF. Стара особина која е вратена кај Radeon HD 5800 е Supersample AA. Целосен Supersample AA може да биде избран во Catalyst Control Center. Може да се избере или Multisample AA или Supersample AA. Ова бара голема вклученост на хардверот, но овозможува да биде присутен во некое ниво во повеќето игри со пониска резолуција. ATI го подобрил филтрирањето на текстурата – новиот AF алгоритам ја елиминира аголната зависност од претходната генерација.



Слика 5: Филтрирање на текстура

DirectX 11

Radeon HD 5800 серијата целосно го поддржува DirectX 11 во секој поглед. Оваа GPU (Graphics processing unit) беше изработена со оваа особина и AMD сакаше да излезе на пазарот заедно со Windows 7. Radeon HD 5800 серијата ја поддржува и DirectX теселацијата во целост преку Hull и Domain Shaders. Едно од подобрувањата во DirectX е во multi-threading што ја овозможува апликацијата – DirectX runtime и DirectX driver да работат во одвоени нитки. Најважен аспект на DirectX 11 и DirectCompute 11 во DirectX 11 е што може да се користат во игри. Подолу е дадена листа на игри кои ќе го поддржат DirectX 11 во некоја или друга форма.

Studio	Game	Release
EA Phenomic	BattleForge	September '09
GSC Gameworld	S.T.A.L.K.E.R.: Call of Pripyat	Q4 '09
Codemasters	DiRT 2	Q4 '09
Turbine	Lord of the Rings Online	Q1 '10
Turbine	Dungeons and Dragons Online: Eberron Unlimited	
Rebellion	Aliens vs. Predator	Q1 '10
Kylin	Genghis Khan	
EA DICE	Frostbite 2 Engine	2010
Trinigy	Vision Engine	

Слика 6: Листа на игри кои први го поддржале DirectX 11

Видео

Исто така во Radeon HD 5800 серијата има и подобрувања во видеото. UVD 2.0 (Unified Video Decoder) е вратено и подобро за да овозможи два 1080p HD видео прикази одеднаш. HDMI (High-Definition Multimedia Interface) особините се исто така подобри, поддржувајќи HDMI 1.3a. Има и звучни подобрувања поддржувајќи Dolby TrueHD и DTS-HD Master Audio со целосна поддршка за Blu-ray аудио формати и до 8 канали за 192kHz/24-битен звук. Radeon HD 4800 серијата е одлична во декодирање на видеа, а Radeon HD 5800 е само малку подобрена.

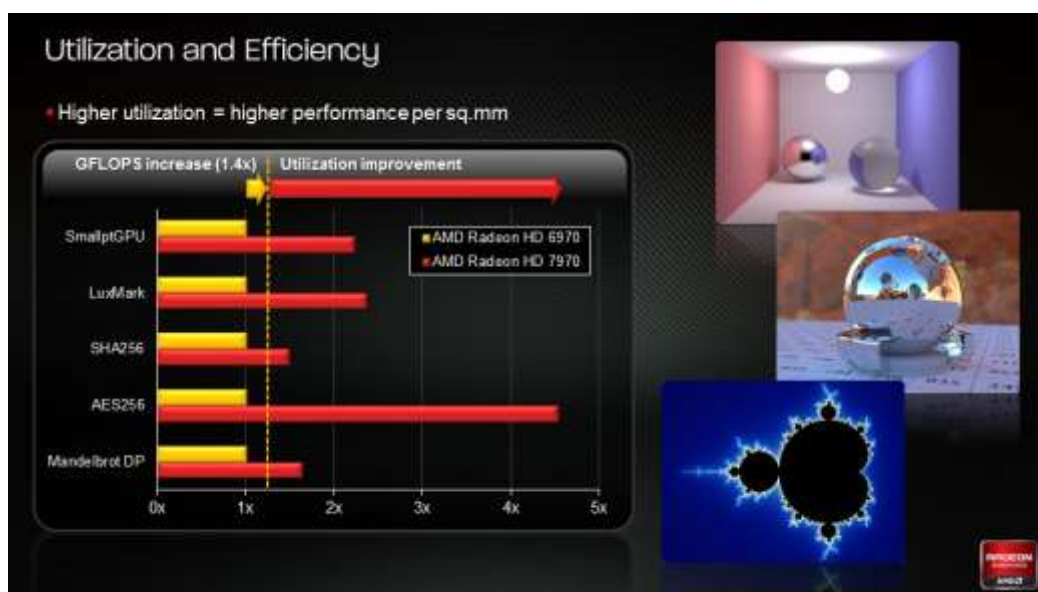
Архитектура на Radeon HD 7000 (GNC)

AMD Radeon HD 7000 серијата со високи перформанси е радикално нов пристап кон дизајнот на GPU. Најважно е тоа што таа е прва архитектура во светот со 28nm на GPU што овозможува AMD да собере до 4,3 милијарди транзистори во приближно ист простор што собира до 2,6 милијарди транзистори. Овој тип на GPU е еден од најмоќните и најнапредните графички процесори. Оваа архитектура е опремена со околу 2048 stream процесори, а секој од нив содржи свој скаларен копроцесор. Во комбинација со големиот пораст на GPU програмските јазици како C++ AMP и OpenCL – GCN (Graphics Core Next) архитектурата е вистинската архитектура која што се појавила во вистинско време.

Перформанси

Зголемувањето на бројот на транзисторите во GPU има големо влијание на потенцијалните перформанси на графичката картичка, но транзисторите сами по себе не се доволни. Потребен е навистина одличен дизајн, како GCN архитектурата, за ефикасно искористување на потенцијалните перформанси. Почнувајќи со искористување на GPU, AMD направи големи напори за да се осигура дека GCN архитектурата е способна за ефикасно искористување на хардверските ресурси. GCN архитектурата е дизајнирана за подобро искористување со што се осигурува дека GPU врши оптимално искористување на своите ресурси за максимални перформанси.

DirectX 11 теселацијата е исто така подобра од претходните AMD Radeon серији на графички картички.



Слика 7: Споредба на искористеност и ефикасност помеѓу AMD Radeon HD 7000 серијата и AMD Radeon HD 6000

GCN архитектурата има придобивки од драматично подобрените теселациски перформанси. Игрите со оваа технологија се значително побрзи отколку во претходните генерации на AMD Radeon производите. Секоја графичка картичка е дизајнирана да одземе само одредена количина на енергија од напојувањето, познато како TDP (Thermal Design Profile). GCN архитектурата може да го направи тоа благодарение на AMD PowerTune технологијата. AMD PowerTune технологијата претставува интелигентен систем кој што врши real-time анализа на игрите и апликациите кои го користат GPU. Оваа технологија ги подобрува перформансите на апликацијата со забрзување на clockspeed-от на GPU до 30%. Најдобро од сè е тоа што оваа технологија е комплетно автоматска и е специјално дизајнирана да ги подобри гејмерските перформанси.

Квалитет на слика


Перформансите не се доволно добри за денешните гејмери. Квалитетот на сликата или јасноста и прецизноста на текстурата и ефектите, е исто така важен дел од дизајнирањето на графичките картички. GCN архитектурата е опремена со три клучни технологии кои драстично го зголемуваат нивото на квалитет.

1. PRT (Partially Resident Textures)

Во многу игри може да се забележи дека често се повторуваат текстури, особено сценографијата во позадина (планини, дрвја и сл.). Тоа е затоа што ако се зголеми големината или бројот на текстурите, играта може да има негативно влијание на перформансите на GPU. PRT е радикална нова технологија која го решава овој проблем.

Partially Resident Textures (PRT)

- Local graphics memory behaves like a hardware-managed cache
 - Texture data can be streamed in on demand
- Improved memory efficiency and image quality with very large, detailed textures
 - Hardware accelerated mapping & filtering for Virtual Texturing or "MegaTexture"
 - Texture sizes up to 32 TB (16k x 16k x 8k x 128-bit)
 - Expected to feature in next-gen game engines
 - High value to apps working with very large data sets



Слика 8: PRT

PRT може да искористи огромен број на текстурни датотеки, дури до 32TB, со минимално влијание врз перформансите. Ова се постигнува со протекување на мали делови од тие масивни текстури, давајќи виртуелно бесконечно снабдување со уникатна текстура. PRT им овозможува на идните игри да користат текстури со ултра висока резолуција со истите перформанси како денешните.

2. Подобро AF (Anisotropic Filtering)

Достапна на секој модерен GPU, AF е технологија која му помага на GPU во правење на убедливи текстури во игрите. Секој GPU се разликува во начинот на кој AF се извршува. GCN архитектурата е специјално оптимизирана за да произведе супериорни резултати кога AF е овозможено. Подобрено AF во Radeon 7000 серијата овозможува да се добијат поостри и подобри текстури.

3. Усовершена теселација на DirectX 11

Како што DirectX 11 станува поразвиен, креаторите на игри го пробиваат факторот на реалистичност со користење на повисок степен на специјални цртачки ефекти. Еден таков ефект е теселацијата, која може динамично да создаде дополнителен детаљ на сцената. Како и AF и теселацијата не е нова во GPU, но начинот на кој теселацијата е изведена може да има силно влијание начинот на претставување на графиката. Заради ова, GCN архитектурата е оптимизирана до 4x од перформансите на Radeon HD 6000 серијата.

AMD ZeroCore Power Technology

AMD ZeroCore Power Technology го засилува лидерството на AMD во ефикасноста на моќноста на преносливите компјутери со цел да им овозможи на GPU на десктопот да се исклучат кога мониторот е исклучен, познато како долга idle состојба. Ова е одлично кога ќе се оддалечите од компјутерот за да се јавите, да гледате телевизија или да отидете до продавница. Исто така AMD ZeroCore Power овозможува дополнителни GPU во AMD CrossFire конфигурацијата да се исклучат кога не се користат. Оваа AMD ексклузивна технологија обезбедува некористени GPU да се ефикасни колку што е можно повеќе. Дури и најстраствениот гејмер(играч) со AMD CrossFire multi-gpu конфигурација има корист од AMD ZeroCore Power. Некористени GPU се исклучуваат со цел да заштедат моќност.

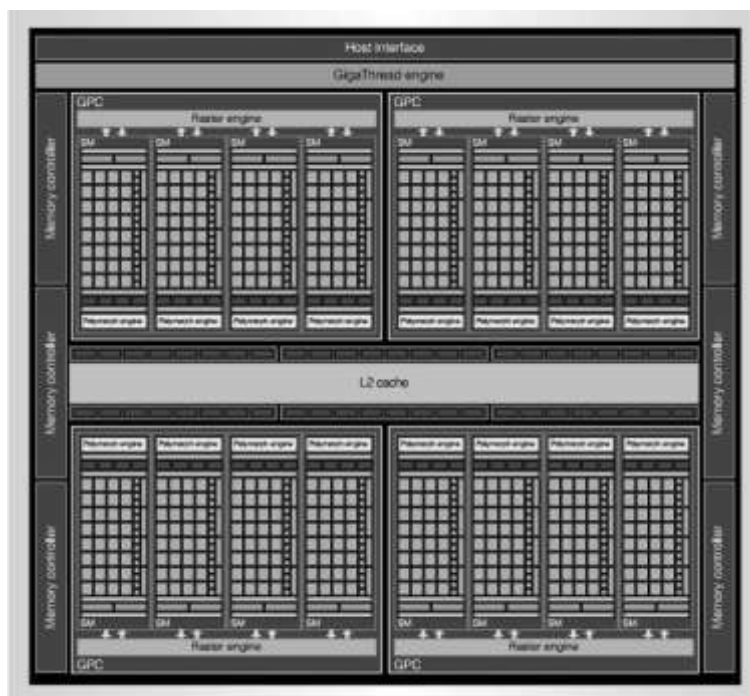
Архитектура на Fermi GF100

Fermi GF100 архитектурата обезбедува неколку нови способности во однос на GT200 или Tesla архитектурата. Оваа архитектура нуди до 512 CUDA јадра и специјални функции за компјутери со високи перформанси. Клучните карактеристики на Fermi GF100 архитектурата вклучуваат 6 мемориски контролери, 4 GPU, GigaThread технологија и L2 кеш за запишување или читање. Секој GPU содржи 4 streaming мулти процесори, а исто така 16 PolyMorph уреди – по еден за секој процесор.

Перформанси

Fermi GF100 архитектурата има 32 CUDA јадра, а секое јадро содржи 4 streaming мулти процесори – вкупно 512 јадра. Секој streaming процесор содржи 48Kb L1 кеш меморија. Секое CUDA јадро има floating-point и целобројна логика на извршување и инструкциите ги извршува паралелно. Секој streaming процесор е поврзан со PolyMorph уредот кој што има посебна фиксна функција и е специјално програмиран за графика и теселација.

Fermi GF100 архитектурата има неколку кеш хиерархии кои ги чуваат податоците при нивна обработка. Податоците на L1 кешот може да се користат за операции кои се извршуваат со помош на стек ориентирана машина или за да се подобри ефикасноста за зачувување и читање на операции. L1 кешот е поддржан од L2 кешот кој служи за читање и запишување.



Слика 9: Преглед на GF100 архитектурата

Меморискиот систем вклучува повеќенаменски мемориски контролери. Бидејќи L2 кешот е унифициран и сите клиенти го користат како читачко-запишувачки кеш, барањата се распределени помеѓу повеќенаменските уреди, како што е PolyMorph и текстурните уреди. За да ефикасно ја поддржи теселацијата податоците на PolyMorph уредот остануваат на чип во кешот, додека често текстурните мапи се доста големи и затоа мора да бидат земени и донесени во кешот. Ова кешот го прави со заменување на старите податоци. За графика се користат блокирачки формати за да “шетаат” низ меморијата со цел да се задоволат конфликтните барања.

Теселација

Теселацијата е нова, значајна особина во GF100 изработен од Microsoft DirectX 11. За да се разбере архитектурата, потребно е да се разберат чекорите на процесирање. Најпрво, вертекс податоците (податочна структура која опишува точка во 2D или 3D простор) се процесираат со тоа што се проектираат на екранот. Следно, patch-овите (софтверски елементи дизајнирани да решаваат проблеми или да ги усовршуваат компјутерските програми или пак да ги подржуваат податоците) се процесираат со цел да ги спојат заедно вертекс податоците. Вертекс податоците и patch-овите може да бидат земени од меморијата. Факторот на теселација го контролира геометрискиот детаљ и кодира колку нова геометрија да создаде. Теселаторот зема фактори на теселација и на излез дава триаголници и линии. Може да се случи значајна ефикасна експанзија на податоци

Физичко процесирање

Физичкото процесирање на GPU овозможува подобрена визуелна реалност. Во GF100 е овозможено двојно подобрување на физичкото процесирање. Со користење на физичкото моделирање, аниматорите и дизајнерите на игри можат да го одредат типот на ефектот, но real-time процесирањето овозможува повеќе реални елементи. Во оваа архитектура се додадени особини во GPU за да се забрза PhysX, меѓу кои и побрзата атомичност (операција која што не може да се раздели на помали делови) и редуцираност, паралелна гарантирана синхронизација и сумирани пресметки, како и L1 и L2 кеш хиерархиите.

Видови Fermi архитектури

Со менување на Fermi архитектурата во дистрибуиран растеризациски систем постигната е едноставна поделба. Главни промени се модифицираните streaming процесори. Ако GF100 е предодреден за потрошувачка графика и CUDA со високи перформанси, тогаш GF104 е предодреден примарно кон потрошувачот. Затоа на GF104 не му е потребна ECC и му треба помала двојна прецизност при операции со floating-point. GF104 користи наизменична архитектура на streaming процесори. Може да се создадат други чипови со комбинирање на различен број на единици затоа што Fermi архитектурата овозможува поделба на фамилии на GPU. GT200 архитектурата со нереплицираниот растеризациски уред бараше повеќе време за поделба на помали системи. Меморискиот систем е исто така поделен - GF100 има 6 мемориски контролери, додека пак GF104 има 4.

Table 2. GF100 versus GF104 scale of units.								
GPU	GPC	CUDA cores	Frame buffer pins	ECC	Total L2	Total L1	Tex	Double precision Gflops/sec
GF100	4	512	384	Yes	768 Kbytes	256 Kbytes	16 units	768
GF104	2	384	256	No	512 Kbytes	128 Kbytes	16 units	96

Слика 10: Разлика помеѓу GF100 и GF104

Архитектура на Kepler GK110

Како што расте побарувачката за паралелна компјутеризација со високи перформанси во многу области од науката, медицината, инженерството и финансиите, nVIDIA продолжува со иновации и со задоволување на потребите со извонредно моќни GPU компјутерски архитектури. Постоечките GPU на Fermi архитектурата имаат редефинирани и забрзани HPC (High Performance Computing) можности во области како што се сеизмичко процесирање, биохемиски симулации, климатско моделирање, сигнално процесирање, компјутерски финансии, компјутерско инженерство, компјутерска флуидна динамика и анализа на податоци. Новата архитектура на nVIDIA, Kepler GK110 GPU го подига паралелната компјутеризација и ќе помогне во решавање на најтешките светски компјутерски проблеми. Со нудење на повисока процесирачка моќност отколку претходната GPU генерација и со нудење на нови методи за оптимизација, Kepler GK110 го поедноставува создавањето на паралелни програми и уште повеќе ќе ги револуционизира високите перформанси на компјутерите.

Перформанси

Со опфаќање на 7,1 милијарди транзистори Kepler GK110 не е само најбрзата, туку и најкомплеско градената архитектура на микропроцесор. Со додавање на нови иновативни особини кои се фокусираат на компјутерските перформанси, Kepler GK110 е дизајнирана да биде центар на паралелното процесирање за Tesla и HPC пазарот. Kepler GK110 овозможува капацитет со повеќе од 1 терафлопи на двојна прецизност. Заедно со подобрените перформанси, Kepler архитектурата направи голем чекор напред во ефикасноста на моќноста овозможувајќи до 3x од перформансите на Fermi.

Следниве нови особини во Kepler GK110 овозможуваат подобрена GPU утилизација, поедноставен дизајн на паралелни програми и помош во користењето на GPU во спектарот на компјутерски средини почнувајќи од персонални компјутери до суперкомпјутери:

1. Динамична паралелност – овозможува GPU да создава нова работа за себе, да синхронизира резултати и да го контролира распоредот на таа работа преку забрзани хардверски патишта и сето тоа без вклучување на CPU. Со овозможување на флексибилноста на адаптација на количината и формата на паралелноста во извршување на програмот, програмерите можат да откријат повеќе видови на паралелна работа. Ова овозможува помалку структурирани, но покомплексни задачи да течат поефективно и полесно овозможувајќи поголеми порции на една апликација целосно да течат во GPU. Исто така полесно може да се создадат програми и CPU е ослободен од други задачи.

2. Hyper-Q – овозможува повеќенаменски CPU јадра да ја започнат работата на еден GPU истовремено и со ова драматично се намалува GPU утилизацијата и значително се намалуваат CPU idle времињата. Hyper-Q го зголемува вкупниот број на конекции помеѓу компјутерот и GK110 GPU со овозможување на 32 истовремени, хардвер-контролирани конекции за разлика од единечната конекција која е присутна кај Fermi. Hyper-Q е флексибилно решение кое овозможува одделни конекции од повеќенаменски CUDA јадра, повеќенаменски MPI (Message Passing Interface) процеси.

3. GMU (Grid Management Unit) – за да се овозможи динамична паралелност потребна е понапредна, пофлексибилна менаџерска мрежа и експедитивен, контролен систем. Новиот GMU на GK110 ги распределува мрежите кои што треба да бидат извршени на GPU. GMU може да ја паузира брзината на новите мрежи и тие ќе бидат ставени во ред на чекање, сè додека не бидат подготвени да се извршат. GMU овозможува работите и на CPU и на GPU да се соодветно менаџирани и забрзани.

4. **nVIDIA GPUDirect** – ги овозможува GPU во еден компјутер или GPU во различни сервери лоцирани низ мрежата директно да разменуваат податоци без потреба да се оди во меморијата на CPU. RDMA (Remote Direct Memory Access) особината во GPUDirect овозможува уреди од трета страна како што се SSD, NIC и IB адаптери директно да имаат пристап на меморијата од повеќенаменски GPU во истиот систем и со тоа значајно се намалува прикриеноста на MPI пораките до/од GPU меморијата. Исто така го намалува опсегот на барања на системската меморија и ги ослободува GPU DMA уредите за користење во други CUDA процеси. Kepler GK110 исто така ги поддржува останатите особини на GPUDirect вклучувајќи ги Peer-to-Peer и GPUDirect за видео.

	FERMI GF100	FERMI GF104	KEPLER GK104	KEPLER GK110
Compute Capability	2.0	2.1	3.0	3.5
Threads / Warp	32	32	32	32
Max Warps / Multiprocessor	48	48	64	64
Max Threads / Multiprocessor	1536	1536	2048	2048
Max Thread Blocks / Multiprocessor	8	8	16	16
32-bit Registers / Multiprocessor	32768	32768	65536	65536
Max Registers / Thread	63	63	63	255
Max Threads / Thread Block	1024	1024	1024	1024
Shared Memory Size Configurations (bytes)	16K	16K	16K	16K
	48K	48K	32K	32K
			48K	48K
Max X Grid Dimension	2 ¹⁶ -1	2 ¹⁶ -1	2 ³² -1	2 ³² -1
Hyper-Q	No	No	No	Yes
Dynamic Parallelism	No	No	No	Yes

Слика 11: Споредба помеѓу Fermi и Kepler архитектурите

Главна цел на дизајнот на Kepler архитектурата е подобрување на енергетската ефикасност. Инженерите се потрудија да ги подберат сите перформанси во однос на Fermi архитектурата. 28nm технологија одигра важна улога во намалувањето на потрошувачката на енергија, но беа потребни многу модификации на GPU за да се редуцира потрошувачката на енергија, притоа да се одржат одличните перформанси. Секоја хардверска единица во Kepler архитектурата е дизајнирана да обезбеди извонредни перформанси во однос на потрошувачката на енергија. Најдобар пример за ова е дизајнот на новиот streaming мулти процесор кој вклучува значително повеќе алгоритми за пресметување со двојна прецизност.

Подобрување на текстурата

Специјално наменетиот GPU за текстура е вреден извор на компјутерски програми кои имаат потреба за обработка и филтрирање на податоци со слики. Текстурата која е додадена во Kepler архитектурата е значително зголемена споредена со онаа на Fermi – секоја единица на streaming процесорот содржи 16 единици за филтрирање на текстурата што е 4 пати повеќе од Fermi архитектурата.

Kepler го менува начинот на функционирање на текстурата. Во Fermi генерацијата, за да GPU повика една текстура, треба да биде доделен слот во поврзувачка табела со фиксна големина пред да започне лансирањето на мрежата. Бројот на слотови во таа табела ограничува колку уникатни текстури еден програм може да прочита. На крај, програмот во Fermi беше ограничен да има пристап само на 128 истовремени текстури. Со користење на bindless текстури во Kepler, дополнителниот чекор на користење на слотовите не е потребен – текстурата е зачувана во меморијата и хардверот ги зема овие зачувани објекти по потреба, со што поврзувачките табели се вон употреба. Ова ефективно ги елиминира сите ограничувања на бројот на уникатни текстури кои може да бидат повикани од компјутерскиот програм.

L1, L2 и ECC кај Kepler архитектурата

Мемориската хиерархија кај Kepler архитектурата е организирана слично како кај Fermi. Kepler архитектурата поддржува унифицирани патеки за внесување и чување податоци со помош на L1 кеш кој го има на секој streaming процесор. Исто така оваа архитектура овозможува употреба на дополнителни нови кеш мемории за читање на податоци. Овозможена е и дополнителна флексибилност во распределбата помеѓу внатрешната меморија и L1 кешот со што се дозволува 32KB/32KB поделба помеѓу нив. Зголемен е и опсегот на внатрешната меморија од 64b на 256b. Како дополнување на L1 кешот, воведени се и 48KB кеш меморија за податоци за читање.

Особините на GPU од 1536KB на KeplerGK110 со L2 кеш меморија го удвојуваат количеството на L2 во Fermi архитектурата. L2 кешот е примарната точка на унификација на податоци помеѓу SMX единиците, сервисирајќи ги сите барања во однос на меморија и текстура со што се обезбедува ефикасно споделување на податоци со голема брзина низ GPU. L2 кешот во Kepler нуди повеќе од 2 пати од опсегот достапен во Fermi. Како и кај Fermi и кај Kepler архитектурата регистерските фајлови, внатрешната меморија, L1 и L2 кешовите се заштитени со Single-Error Correct Double-Error Detect (SECDED) ECC код.

ECC ги проверува земените битови од DRAM што резултира со разлика во перформансите помеѓу ECC-возможна и ECC-невозможна операција, посебно на апликации со осетлив опсег на меморија. KeplerGK110 имплементира неколку оптимизации врз ECC проверката на земените битови базирана на искуството на Fermi.

За крај

Со лансирањето на Fermi архитектурата, nVIDIA влезе во нова ера со компјутери со високи перформанси базирани на хибридни компјутерски модели каде што CPU и GPU работат заедно во решавање на компјутерската, интензивна оптеретеност. Сега, со лансирањето на Kepler GK110, nVIDIA уште повеќе ги извиши границите на високите перформанси на компјутерските системи.

Kepler архитектурата е дизајнирана со цел да ги подобри компјутерските перформанси, како и да ги зголеми капацитетите на компјутерите и сето ова да се постигне со извонредна енергетска ефикасност. Во оваа архитектура има вметнато многу нови иновации, како што се streaming процесорите, динамичка паралелност и Hyper-Q технологијата кои направија хибридните компјутерски модели драматично да се забрзаат, полесни да се за програмирање и на нив да може да се применуваат поширок опсег на компјутерски апликации. Kepler архитектурата ќе се користи во бројни системи, почнувајќи од работни станици, па сè до суперкомпјутери за решавање и на најсложените предизвици.

Професионални наспроти обични графички картички

Да се избере вистинската професионална графичка картичка не е лесно. Од хардверска гледна точка разликата помеѓу професионална и обична графичка картичка не е голема. Денеска повеќето професионални картички имаат ист хардвер со обичните картички, иако чиповите се рачно изработени од најквалитетните делови. Исто така поседуваат повеќе RAM меморија, отколку обичните. Но, најголемата разлика помеѓу професионалната и обичната графичка картичка е во нивната поддршка на софтвер. Хардверот на обичните картички е повеќе насочен кон fill rate (број на пиксели што графичката картичка ги носи и ги запишува во видео меморија за време од една секунда) и сенчење, додека на професионалните е насочен кон 3D операции како што се геометриски трансформации и вертекс матрици.

Професионалните картички се исто така значително оптимизирани, тестирани и потврдени за употреба со CAD и DCC апликации. Ова не само што ги зголемува перформансите, туку нуди и одлична стабилност и предвидливост

споредено со обичните графички. Генералниот консензус дека апликациите како 3ds Max, Maya, Softimage, AutoCAD, SolidWorks ќе работат и на обичните картички, но не со такви перформанси како кај професионалните и често може да се појават дефекти и аномалии. Овие проблеми се помалку чести кај професионалните картички и кога ќе бидат откриени, брзо се адресираат до производителите.

Дали nVIDIA ili AMD?

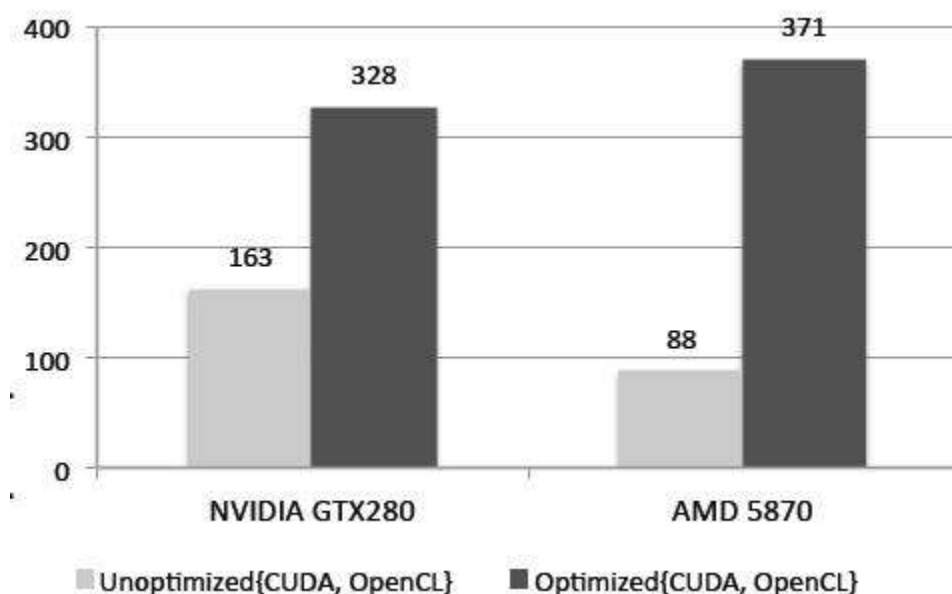
Може да се заклучи дека nVIDIA и AMD делат многу слични концепти при дизајнирањето, но сепак се разликуваат во многу аспекти, почнувајќи од процесорските јадра па се до меморискиот систем. Затоа треба да се спроведе студија со цел да се испитаат карактеристиките во нивните архитектури со помош на неколку апликации. И nVIDIA и AMD имаат различни предности и даваат предност на различни апликации за подобри перформанси и енергетска ефикасност.

Релативните перформанси на картичките доста варираат од еден тест на друг со што тешко е да се прогласи “победник” помеѓу nVIDIA и AMD. Но секако за победник може да ја прогласиме последната употребена архитектура на nVIDIA. Нивната картичка направи вистински бум, опремена е со многу повеќе меморија отколку меморијата на компјутерите пред неколку години. Ова резултираше со апсолутна доминација во сите тестови. Но, овие огромни перформанси доаѓаат со огромна цена и за многумина цената од 4000 долари е прескапа. Во тестирањата и архитектурата на AMD не заостанува многу. Најдобрата графичка картичка во професионалниот графички сектор е онаа од последната генерација на AMD.

Сè на сè работите добро стојат и за nVIDIA и за AMD во професионалниот графички сектор. Ниту еден од нив не е доминантен, но одлуката за купување на вистинската графичка картичка ќе зависи од тоа кои апликации ќе се користат и од нивната цена. nVIDIA картичките се со подобри перформансии во повеќе апликации, отколку од оние на AMD, но се поскапи. AMD картичките се поефтини, но тие го поддржуваат само OpenCL со што ги ограничуваат перформансите на GPU.

Заклучок

Предностите на GPU се во перформансите, нивната цена и нивната енергетска ефикасност и затоа се наоѓаат високо во светот на компјутеризацијата. Исто така тие сè повеќе и повеќе се користат за десктоп апликации. GPU може да обезбедат оптимизација на програмите со цел да се обезбеди ефикасна употреба на хардверот. Оптимизирачките кодови за nVIDIA се добро проучени и има доста студии за нив, но оптимизирачки кодови за други GPU платформи, како што се оние на AMD многу ретко се спомнуваат.



Слика 12: Споредба на ефикасноста на неоптимизирана и оптимизирана архитектура на nVIDIA и AMD

Споредена е оптимизацијата помеѓу двете различни архитектури. Ова е направено со помош на апликација чија оптимизација е детално проучена на nVIDIA платформа, како и на постариот модел од AMD. Со овој процес откриено е дека оптимизацијата од nVIDIA има подобрувања во перформансите, додека онаа на AMD нема некои значајни подобрувања. Затоа некои од оптимизациите кои предизвикуваат пад на перформансите кај nVIDIA, кај AMD предизвикуваат забрзувања и обратно. За крај да заклучиме дека добро познати оптимизации од една архитектура не секогаш се применливи во друга. Така малку познатите оптимизации за OpenCL кај AMD даваат добри резултати и се далеку подобри од оние добро познатите.

Користена литература

- [1]. http://www.hardocp.com/article/2009/09/22/amds_ati_radeon_hd_5870_video_card_review/5#.UWRld70z91E
- [2]. <http://www.amd.com/us/products/technologies/gcn/Pages/gcn-architecture.aspx>
- [3]. http://www.nvidia.co.uk/object/what_is_cuda_new_uk.html
- [4]. <http://blog.cuvilib.com/2012/03/28/nvidia-cuda-kepler-vs-fermi-architecture/>
- [5]. <http://www.nvidia.co.uk/object/nvidia-kepler-uk.html>
- [6]. developer.download.nvidia.com/CUDA/training/GTC_Express_David_Luebke_June2011.pdf
- [7]. http://www.it.uu.se/katalog/davbl791/gpu_architecture.pdf
- [8]. <http://yourstory.in/2010/11/nvidia-today-announced-the-latest-consumer-gpu-based-on-its-fermi-architecture-worlds-fastest-dx11-gpu/>
- [9]. <http://www.cgchannel.com/2011/10/review-professional-gpus-nvidia-vs-amd-2011/>
- [10]. <http://www.nvidia.com/content/PDF/kepler/NVIDIA-Kepler-GK110-Architecture-Whitepaper.pdf>
- [11]. S. Ryoo, C. I. Rodrigues, S. S. Baghsorkhi, S. S. Stone, D. B. Kirk, and W. W. Hwu, "Optimization Principles and Application Performance Evaluation of a Multithreaded GPU Using CUDA,"
- [12]. AMD, "AMD Stream Computing OpenCL Programming Guide."