РАЗВИТИЕ НА GPGPU СУПЕР-КОМПЮТРИТЕ. ВЪЗМОЖНОСТИ ЗА ПРИЛОЖЕНИЕ НА NVIDIA CUDA ПРИ ПАРАЛЕЛНАТА ОБРАБОТКА НА СИГНАЛИ И ИЗОБРАЖЕНИЯ

ас.д-р Георги Петров, Филип Андонов - НБУ
департамент "Телекомуникации", департамент "Информатика"
e-mail: gpetrov@nbu.bg, fandonov@nbu.bg

Развитието на софтуерните системи и алгоритмите за цифрова обработка и анализ на сигнали и изображения, позволиха такива системи да се интегрират в класически приложения ползвани практически във всяка една човешка дейност, от системите за графична обработка и предпечат на произведения на изкуството, кино филми, аудио записи, 3D анимация и игри, до високо прецизните медицински компютърни томографи и магнитни резонансни скенери, както и системите за сигнална обработка, компресия на цифрово видео и много други сфери и приложения. Макар наличните на пазара компютърни конфигурации днес да надвишават хилядократно по производителност и бързодействие своите предшественици от преди 10 години, тези системи са оптимизирани за работа с така наречените настолни приложения. Този тип приложения са оптимизирани за офис работа, основно текстови и графични редактори, гледане на цифрово видео и слушане на музика, уеб приложения и др. Класическите процесорни архитектури (Intel, AMD x86, x64 модели) предлагани днес на пазара в най-разнообразни конфигурации 2, 4 и 8 ядрени процесорни блокове, които позволяват многократно повишаване на бързодействието на потребителския и системен софтуер. Тези системи са създадени предимно за символна обработка и последователни изчисления, което ги прави нефункционални в приложения за сигнална и статистическа обработка.

За научни изчисления години наред се разработват клъстерни супер компютърни системи, като: Connection Machine (1980), MasPar (1987), Cray (1972-1995, която се слива със Silicon Graphics през 1996), в които става възможно паралелизирането на отделните изчислителни задачи. Естествено този тип системи са прекалено скъпи и достъп до тях имат учените и изследователите само от големите университети. Цената на една такава система е огромна, което ги прави неприложими за конвенционални нужди. Интересно е да се знае, че от системата CM-1 има продадени едва стотина броя, а от MasPar около 200. Такъв тип компютърни системи са финансово неефективни за решения в съвременните болнични заведения в отделенията за образна диагностика, както и в научните центрове, академии и университети, и принципно немислими за индивидуални изследвания.

Суперкомпютрите по света

Ако сравним супер компютрите по света можем да отчетем че през 2009г. Московският държавен университет изгражда най-мощният супер компютър в цяла Източна Европа с производителност от 350 терафлопа, имащ пиковата мощност от 414 терафлопа. Като вторият по мощност в цял свят е изграден в Китай и притежава 1206 терафлопа, сравнен с "Ягуар", който дефакто е най-мощният през 2009г. създаден от Cray Inc. имащ колосалните 224162 микропроцесора и производителност от 1759 терафлопа, с пиковата производителност 2331 терафлопа.

Година	Цена на 1 GFLOP US$	Технология
1961	1,100,000,000,000 $1,100 на FLOP	17 милиона IBM 1620 $64,000 всеки
1984	$15,000,000	Cray X-MP
1997	$30,000	2x16 Beowulf клъстер с Pentium Pro
2000	$640 $1/MFLOPS	KLAT2
2007	$45 /GFLOPS	Ambric AM2045
2009	$1.39 /GFLOPS	NVIDIA Tesla C1060 930GFLOPS

*T-G-K FLOP (Terra, Giga, Kilo FLoating point Operations Per Second)

Табл. 1 Историческо развитие на суперкомпютрите

От своя страна българския суперкомпютър IBM Blue Gene/P [15], внесен през септември 2008г. е с Linux-базираната платформа и има 8192 микропроцесора, предоставящ невижданите до момента 23 терафлопа. По отношение на използваните операционни системи може да споменем, че до 2002г. доминиращ на пазара остават UNIX базираните системи, като след това до 2009г. делът на UNIX остава едва 15%, споделяни с Windows и BSD, като процентите продължава да падат. Почти всички останали нови системи (85% от пазара) са заети от Linux базирани системи, което е напълно обяснимо поради отворените стандарти и ниска себестойност, възможност за бърза модификация, добра документация и липса на ограничения за използване и промяна, както и скалируемост по отношение на добавяне на нови мощности от процесори и дискови масиви.

Как игрите променят света на суперкомпютрите

През 90те години се наблюдава развитието на разпределените многопроцесорни архитектури, което практически представлява множество компютри свързани и работещи в една високоскоростна етернет мрежа.

Ефективността на този тип системи е висока, но въпреки стократно по-ниските цени от класическите супер компютри, тези системи продължават да имат прекалено висок експлоатационен разход на електроенергия, климатизация и пространство. За осигуряване на максимална достъпност на научните организации до голям изчислителен ресурс широко разпространение получи доброволната програма BOINC (Berkeley Open Infrastructure for Network Computing), с помощта на която даден научен проблем се разпределя на отделни малки задачи и посредством Интернет те се стартират на хиляди потребителски компютри [1]. Някои от програмите които имат най-висока популярност са свързани със симулации на климатични процеси, изследване на земетръсната активност, генетика, нано технологии, симулация на флуиди, търсене на извънземен разум и финансови приложения.

С развитието на персоналните компютри след 1990г. те все по-често стават повече средство за забавление, отколкото работен инструмент. Интересно е да се види как това на пръв поглед странично приложение на компютрите позволява агрегирането на нов пазарен сегмент за специализиран хардуер, основно в областта на тримерната графична обработка. Първата 3D игра излязла през 1981г. наричана 3D Monster Maze е основоположник на нова ера в компютърните забавления (Фиг. 1.а). Естествено, днес този екранен изглед буди само усмивки и умиление.

1.а

1.б

Фиг. 1.а Скриншот от 3D Monster Maze - първата 3D игра за персонален компютър, 1.б скриншот от Wolfenstein 3D.

През 80-те години персоналните компютри не предоставят високо бързодействие нито пък имат удобен потребителски интерфейс и добра графика, което налага изчакване от около 10 години до масовото навлизане на 3D игрите. Играта, която променя света е Wolfenstein 3D, тя излиза на пазара през май 1992г. и заема колосалните за тогава 1.44MB флопи дисково пространство. И макар тази история да няма пряко отношение към научните изследвания, именно игрите създават предпоставка за разработката на нов тип видеокарти. В началото на 90-те години масово ползвани са компютрите базирани на историческия Intel 386 и в последствие Intel 486, както и Motorola 68000. Техните графични карти не позволяват интегрирането на 3D графични възможности, затова този тип приложения са прекалено бавни. Първата

графична карта имаща хардуер за 3D графичен ускорител е Cirrus Logic Laguna 3D, обаче за начало на новата ера може да кажем появата на 3Dfx Voodoo през 1996г., която се счита за първият 3D графичен ускорител имащ невероятен пазарен успех сред геймърите [2]. От този момент светът на персоналните компютри се променя. Най мощните персонални компютри започват да се създават именно и заради геймърите [3]. Пазарният сегмент расте, скоро 3D ускорителите, първоначално предлагани като допълнителна PCI пратка, започват да се вграждат в стандартните видео карти, а класическите софтуерни приложения за графична обработка и 3D симулации, като 3D Studio Max, пуснат с това търговско име за Windows NT, пуснат първоначално за MS-DOS с името Autodesk 3D Studio през 1990г., започват да ползват възможностите на новия хардуер, който е многократно по-евтин от ползваните до момента решения на Silicon Graphics.

Едновременно с това развитие на графичните карти следва да отбележим модификациите на системните шини за свързване на видео картите към дънните платки на персоналните компютри (Табл. 2). Също така и възможностите за ползване на бърза видео памет, която е вградена в платките на самите графични контролери [18]. Тези нововъведения способстват за възможността централния процесор да обменя бързо големи масиви данни с видео платката, като по този начин освобождава системен ресурс за потребителските програми. Друго важно преимущество на този тип архитектура е възможността видео картата да обменя автономно данни с определена част от паметта на компютъра, което дефакто автоматизира процеса на визуализация на 3D структури на екрана, нейното оцветяване с готови текстури зареждане на често използвани в играта обекти, предварително (терен, къщи, дървета), режимите на придвижване и други.

Шина	Битове предавани едновременно	Скорост (MB/s)	Протокол
PCI	32/64	**132/800**	паралелен
AGP 1x	32	264	паралелен
AGP 2x	32	528	паралелен
AGP 4x	32	1000	паралелен
AGP 8x	32	2000	паралелен
PCIe x1	1	250/500	сериен
PCIe x4	1 Ч 4	1000/2000	сериен
PCIe x8	1 Ч 8	2000/4000	сериен
PCIe x16	1 Ч 16	4000/8000	сериен
PCIe x16 2.0	1 Ч 16	8000/16000	сериен
IDE (ATA100)		100 MB/s	
IDE (ATA133)		133 MB/s	
SATA		150MB/s	
Gigabit Ethernet		125 MB/s	
IEEE1394B Firewire]		100 MB/s	

Табл. 2 Еволюция на системните шини за допълнителна компютърна периферия

Основното нововъведение в 3D графичните карти се явява появата на специализирани процесори за обработка на 3D графика, т.н. GPU (Graphics Processing Unit). Като водещи в комерсиалния пазарен сегмент може да се определят компаниите ATI (закупена през 1996г. от AMD) и NVIDIA. Но хардуерът сам по себе си не допринася особено за масовото му ползване в среда MS-DOS. През 1994г. Microsoft представя Windows 95. До този момент производителите на игри считат MS-DOS за по-удобна система при игрите. Това което до този момент отличава DOS, като операционна система е възможността за директен достъп до хардуера, който е скрит в Windows. Друг проблем остава ползването на системните таймери, които не позволяват добро времеделене на операциите и довеждат до съществено влошаване при Windows 3.1 и 95. За решаването на този проблем през 1995г. Microsoft създават DirectX, известна още като Windows Games SDK. Към настоящия момент наличната на пазара версия е DirectX 11. От своя страна функциите на средата дават на програмистите възможност за ползване в реално време на системните ресурси: графична карта, джойстик, клавиатура и мишка, а също така и въвежда т.н. мултимедийни файлови формати и прецизни таймери. Всички тези нововъведения правят възможно лесното приложно ползване на новия хардуер, без това да налага програмистите на игри да пренаписват драйверите за всеки нов хардуер, както това се прави това до момента в MS-DOS, а и разширява броят поддържани видео карти и 3D ускорители имащи драйвери за Windows.

Следва да се отбележи, че в началото на 90-те години Silicon Graphics промотира **IRIS GL** (Integrated Raster Imaging System Graphics Library), която е графична система за работни станции. Потребителските програмни функции поддържат създаване на графичен интерфейс, като прозорци, вход от клавиатура и мишка. Тази среда се появява преди X Window System и е ориентирана предимно към специализиран хардуер. Поради ограниченията на лицензa на IRIS през 1992г. Silicon Graphics създават OpenGL 1.0 (Open Graphics Library) [4], който е крос платформен интерфейс с независима поддръжка на драйверите на видео картите и входно изходните устройства, наличната към момента на написване на тази статия версия е OpenGL 3.2.

Силното развитието на графичния хардуер, както и потребителските функции за 3D графика Direct X и OpenGL, довеждат до идеята за това графичните карти да се използват за общо приложими математически изчисления, основно за обработка на изображения. За това свидетелства появата на вградени възможности във видео картите за обработка на MPEG 2 компресирано видео (ATI - Rage Series, NVIDIA - GeForce). В момента на поява на тези продукти наличните 32 битови процесорни системи работят на честоти едва достигащи 260-300MHz, което ги прави неприложими за компресията и обработката на цифрово видео в реално време. Появата на тази нова възможност в графичните карти и интегрирането и с Direct X и OpenGL значително променя погледа на програмистите занимаващи се с изчислителни проблеми към този тип хардуер. Следва да отбележим, че идеята за ползване на графичните процесори, като основа на математически системи за паралелна обработка датира още от 1978 когато компанията Ikonas Graphics Systems проектира растерен дисплей за оборудване на пилотски кабини по поръчка на

NASA Langley Research Center. Това е последвано от Pixel Machine – 1989 и Pixel-Planes 5 – 1992. Бумът в развитието на графичния хардуер позволява съществено изменение на начина, по който тези системи могат да се ползват за общо приложими математически изчислителни задачи [5]. Тези разработки довеждат до общото развитие на концепцията GP-GPU (General-Purpose Computation Using Graphics Hardware http://www.gpgpu.org/) [6]. Тази концепция позволява бързото решаване на множество сложни последователни алгоритми подлежащи на паралелизация, каквито са обработката и филтрацията на сигналите, компресирането на видео, криптографски анализ, статистически изчисления и други алгоритми обработващи еднотипно големи масиви от данни. Разбира се за нуждите на сигналната обработка водещи производители на чипове, години наред, модифицират и създават т.н. сигнални процесори (DSP Digital Signal Processor), чието внедряване е предимно в специализирани разработки. В комерсиалния сегмент DSP заемат място след средата на 90-те години предимно при направа на цифрови видеокамери, видеоконферентни системи и различни научни области. Тези платформи са трудни за PC програмистите, при тях отсъства универсалност, и дори днес все още остават обект на внедряване във вградени приложения. Програмирането и проектирането на DSP само по себе си не притежава гъвкавостта на програмирането за PC. Възможностите предлагани от съвременните GPU позволява на всеки средно статистически потребител и програмист на PC софтуер да има достъп до високопроизводителни мултипроцесорни евтини разширения. При това цените на подобни платки варират между 60-3000 в зависимост от броят мултипроцесори (16-480) и количеството памет (256MB – 4GB). Тези системи се развиват до такава степен, че днес позволяват постигането на пикова мощност от до 1 тера флоп в една графична видео карта. За да добиете по-ясна представа за развитието на технологията при картите на NVIDIA вижте по-долната сравнителна Табл. 3. Може да се различат две основни продуктови линии Quadro, специализирана за професионални графични приложения и GeForce предназначена основно за крайния потребител и геймърите. Следва да се отбележи, че FASTRA II е базиран изцяло на комерсиална технология от 6 двуядрени GTX295 и 1 едноядрена GTX275.

модел	**код на модела**	**шина**	**обем памет MB**	**трансфер на данните** (GB/s)	**GFLOPs**
GeForce G210M	GT218	PCIe 2.0 x16	512	12.8	72
GeForce GTS 250M	GT215	PCIe 2.0 x16	1024	51.2	360
GeForce GTX 280M	G92b	PCIe 2.0 x16	1024	60.8	562
GeForce GTX 285M	G92b	PCIe 2.0 x16	1024	64	648

Табл. 3.а Сравнение на любителски и геймърски видеокарти на NVIDIA поддържащи CUDA

модел	**код на модела**	**шина**	**памет MB**	**трансфер на данните GB/s**	**брой изобразени пиксели** (M/s)
Quadro	NV10GL	AGP 4x	64	2.66	480
Quadro4 380XGL	NV18GL	AGP 8x	128	8.2	1100
Quadro FX 1400	NV41	PCIe x16	128	19.2	4200
Quadro FX 4500X2	G70	PCIe x16	1024	33.6	11280
Quadro FX 5600[2]	G80	PCIe 2.0 x16	1536	76.8	38400
Quadro FX 5800	G100GL-U	PCIe 2.0 x16	4096	102	52000

Табл. 3.6 Сравнение на професионални видеокарти на NVIDIA поддържащи CUDA

Освен поддържащи CUDA графични видео карти, съществуват и карти, съдържащи само GPU процесори и по-големи обеми RAM. Тези карти са подходящи за вграждане, както в стандартни PC така и в шкафове. По-долу е дадено сравнение на TESLA продуктовата гама на NVIDIA предназначена за създаване на персонални изчислителни системи (Табл. 4). Следва да се обърне внимание, че системите монтирани в шкафове може да бъдат скалирани до стотици терафлопа. Обърнете внимание, че това са пикови мощности и са достижими само за някои алгоритми при условие, че данните са заредени в локалната памет на картите.

конфигурация	код на модела	архитектура	брой процесори	памет MB	Трансфер (GB/s)	GFLOPS
GPU Processor	C870	G80	128	1536	77	519
Deskside Supercomputer	D870	G80	256	2x1536	154	1037
GPU Server	S870	G80	512	4x1536	307	2074
GPU Processor	C1060	G200	240	4096	102	936
GPU Server	S1070	G200	960	4x4096	410	4320
GPU Server	S1075	G200	960	4x4096	410	4320

Табл. 4 Сравнение на модели графични карти на NVIDIA и броят мултипроцесорни и памет в тях.

Продуктовата линия на TESLA е подходяща за изграждане на дестоп супер компютри и вграждане в шкафове. Типично приложение е ползването на картата C1060 за изграждане на до 3.8 терафлопа в единично PC. Някои от

моделите предлагани (Фиг. 2) на пазара, са поддържани от Dell и ASUS, като препоръчваният процесор е Intel® Xeon®.

Фиг. 2 Типове продукти TESLA, за вграждане в стандартно PC – C1060 и за монтаж в шкафове Tesla S1070 GPU сървър.

Следва да споменем, че гейм конзолите от типа Xbox, Xbox 360, Deramcast, Wii, PS2 и PS3 позволяват създаването на персонални супер компютри [7]. По-долу е дадена сравнителна таблица за броят продадени геймърски конзоли, които не са PC базирани, както и изчислителната им мощност. Тези конзоли се програмират по-трудно и не са така удобни като продуктите GeForce и TESLA, при тях липсват много от възможностите предоставяни от класическите PC, тъй като те са предназначени за крайни потребители на игри, а не програмисти.

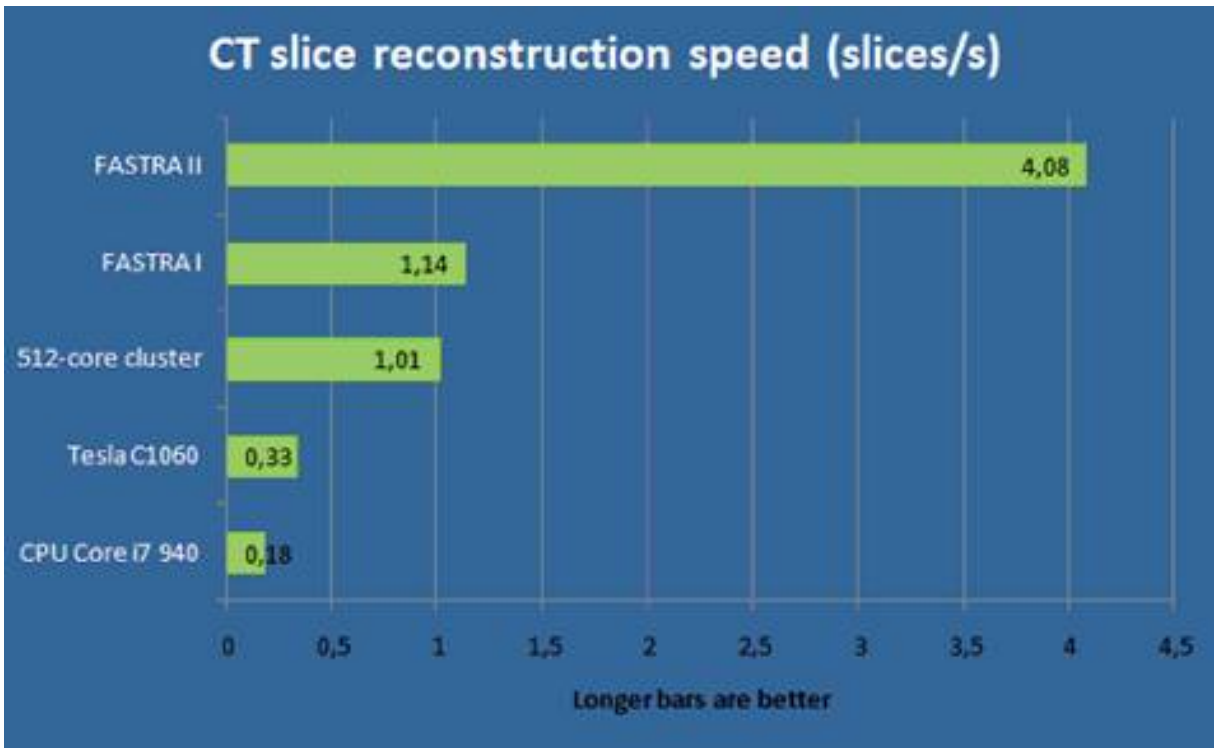
Конзола	Скорост на процесора в гигафлопа	Общо продадени терафлопа	Брой продадени конзоли (милиони)
Xbox	5.8	139 200	24
Xbox 360	115.2	2 188 800	19
Dreamcast	1.4	8 400	6
Wii	2.9	75 400	26
PS2	6.3	768 800	124
PS3	218	283 400	13
Общо	349.6	3 464 000	212

Табл. 5 Сравнение на изчислителната мощност на комерсиални гейм конзоли – видео игри към началото на 2009г. (Изт. [7])

Как CUDA променя ситуацията

След като добихте представа за световните тенденции и класация за супер компютрите през 2009г. и развитието на GPU технологията в професионалните видеокарти се замислете за това дали е възможно да притежавате 12 терафлопа в стандартно PC? Разглеждайки различните модели разпределени изчислителни системи следва да се спрем на един уникален

проект наричан FASTRA I и II създадени във Vision Labs към университета в Антверпен, Белгия. Втората версия на системата е резултат от взаимодействието между студенти, Tones.be и ASUS, за създаването на най-мощнша десктоп супер компютър в света притежаващ 13 GPU и имащ пикова производителност от 11 терафлопа на цена от 6000 евро.



Фиг. 3 Сравнение на скоростта на реконструкция на MRI изображения при ползването на различни технологии и CUDA (Източник: http://fastra2.ua.ac.be/)

Системата е използвана при реализацията на софтуер за реконструкция на MRI (изображения от ядрено магнитен резонанс). Типично този тип приложения изискват много време за реконструкция на изображенията, което е в порядъка на 3 до 12 часа. Тъй като болшинството болници не разполагат с изчислителен ресурс от повече от 1-2 терафлопса изобщо, процесът на диагностика отнема значително време. Обаче използването на CUDA във FASTRA II показва как тази технология може да се ползва за масово въвеждане дори при ниски бюджети практически навсякъде. По-долу са дадени сравнителни данни от производителността на FASTRA II, 512 core cluster, Tesla C1060 и i7940 (Фиг. 3).

Хардуерни особености на CUDA

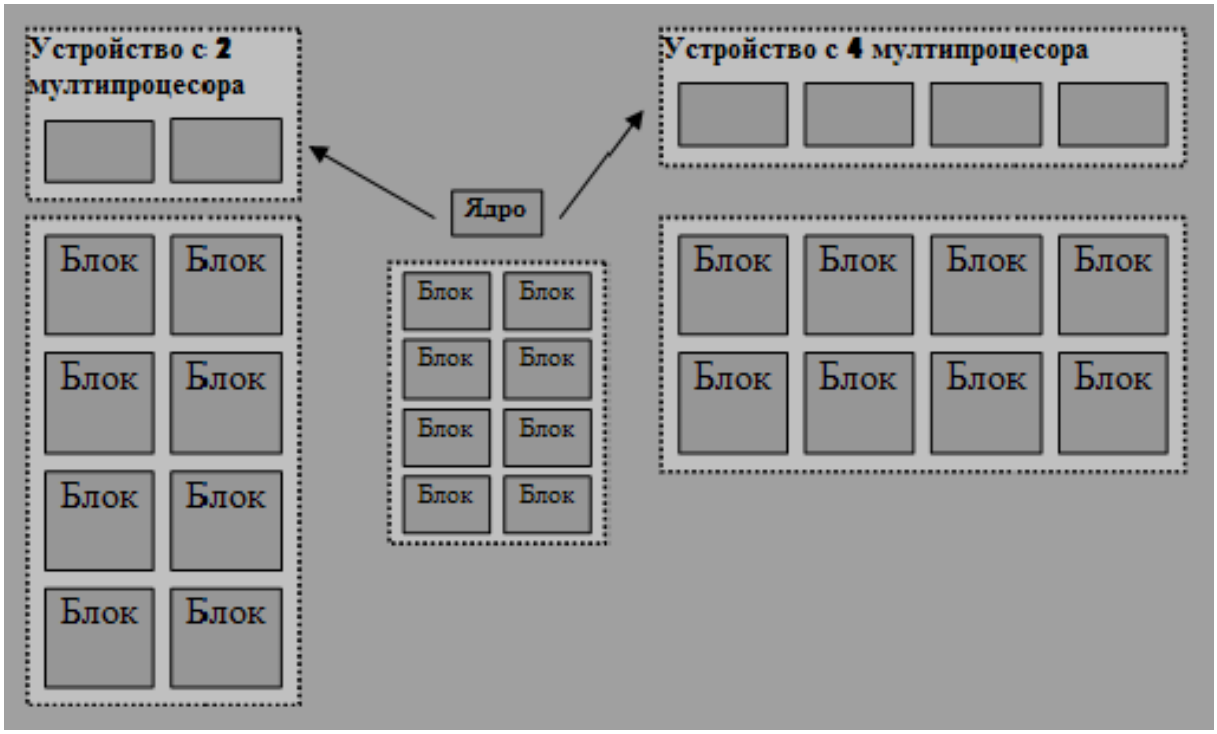
Технологията CUDA се счита за основоположник на "супер компютинг демократизацията" [9]. Чрез CUDA съвместимите продукти всеки програмист и/или учен има достъп до огромен изчислителен ресурс в рамките на стандартните бюджети за закупуване на съвременно PC, като цените на

професионалните продукти започват от 250$ до 1300$ за PC платформи. Приложения и алгоритми чиято, проверка изискваше десетки дни, днес могат да бъдат изпълнени за часове или дори минути в зависимост от ползвания хардуер [8]. Бързодействието на програмите, писани за CUDA съвместими устройства зависи основно от ползвания хардуер, като могат да работят на всички поддържащи технологията устройства. Принципно архитектурата на GPU се различава от тази на CPU (Фиг. 4). При GPU се заделя сравнително малка част от кеша, наричан споделена памет (__shared__) в която може да се извършват обработки от множество АЛУ едновременно, друга особеност е че се ползва мнококанална видео DRAM (__global__), която е в пъти по-бърза от стандартната RAM ползвана в съвременните персонални компютри. Характерно за CUDA е ползването на SIMD архитектура, като при това в данните записани в отделни области обща памет могат да бъдат обработвани едновременно от множество АЛУ. При стандартните процесори AMD и Intel голяма част от транзисторите в чипа се ползват за изграждане на локални кеш памети и по-сложна контролна логика на контролера на процесора. При CUDA повечето АЛУ (аритметично логически устройства) означава, че повече еднотипни изчисления ще бъдат извършени върху даден масив от данни. Разбира се различни техники за оптимизацията на този процес могат да се използват и за тях ще стане дума малко по-късно.

Масово използваните днес процесорни архитектури са Харвардска и Фон Нойман. Първият модел, разпределя отделни блокове памет за съхранение на инструкциите и данните, а вторият тип позволяват на паметта да съдържа едновременно програмни инструкции и данни (Фиг. 4). Освен това процесорите могат да се различават по начина по който обработват инструкциите и данните, като съществуват основно четири модела: **SIMD** (Single Instruction, Multiple Data), MISD (Multiple Instruction, Single Data), SISD (Single Instruction, Single Data) и MIMD (Multiple Instruction stream, Multiple Data stream). Първият тип предполага използването на множество АЛУ работещи едновременно над съседни области памет с еднакви инструкции, каквито са съвременните DSP процесори и CUDA включително.

Фиг. 4 Архитеркурни особености на GPU и CPU, (изт. NVIDIA)

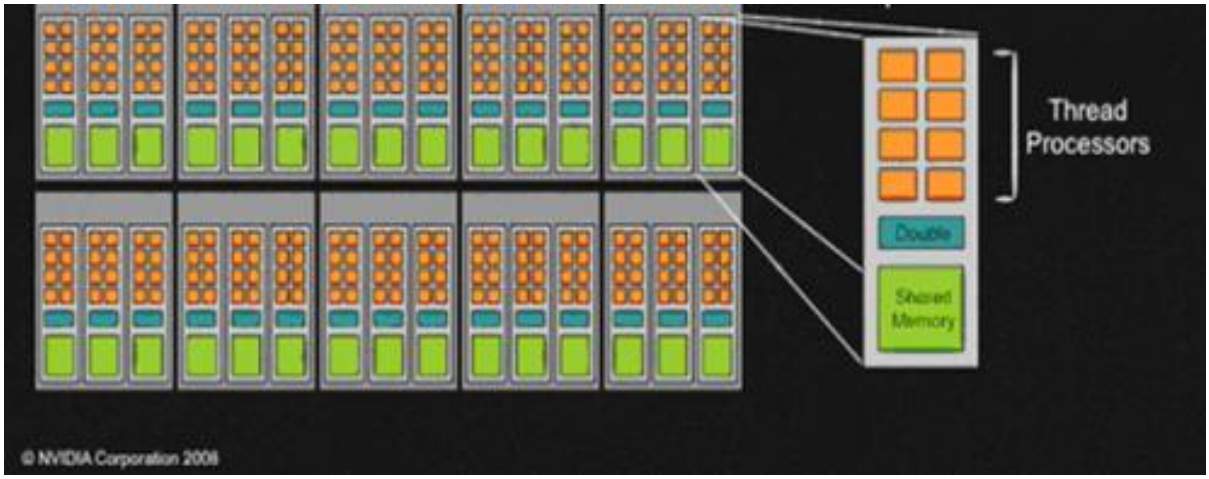
MISD позволява обработката на една област памет едновременно от множество АЛУ. SISD се ползва при микроконтролери и микропроцесори от нисък клас, като е налично едно АЛУ можещо да обработва една данна едновременно. MIMD позволява мащабна асинхронна обработка на данни, като всеки процесор обработва данни намиращи се в различни области на паметта, каквито са много ядрените и многопроцесорни системи. Спецификата на CUDA я доближава много до съвременните DSP процесори, като при това запазва гъвкавостта и лесната промяна на програмната конфигурация характерна за съвременните персонални компютри. Драйверите на CUDA разпределят автоматично отделните изчисления над пълния набор налични процесори в системата, което дава възможност една част групирани в блокове задачи да бъде осъществима на карта с по-малък брой мултипроцесори или такава с по-голям брой (Фиг. 5).



Фиг. 5 Изпълнение на една и съща изчислителна задача от едно ядро с 8 блока на две различни архитектури имащи 2 и 4 мултипроцесора ще отнеме различно време, но резултатът ще бъде един и същ.

В най-общия случай за да извършим някакви изчисления паралелно се стартират n на брой процеса, като данните и съответно инструкциите се разпределят в т.н. рамки, които съдържат блокове, а всеки блок съдържа отделни процеси. Това позволява на драйвера да ползва различни хардуерни архитектури, зареждайки в тях повече или по-малко на брой блокове за обработка. Това става прозрачно за програмиста и позволява ползването на хардуер с различни възможности без да се налага повторно компилиране на софтуера. Естествено ако даден блок съдържа повече процеси отколкото процесори са налични в един мултипроцесор те могат да бъдат групирани. При тези случаи, а това са над 95% от реалните примери, се налага ползването на

допълнителни механизми за синхронизация на процесите преди да се прехвърли управлението върху друга група процеси от блока. Следва да се каже, че ползването на мултипроцесорите за обработка на елементи намиращи се в глобалната памет е многократно по-бавно, това налага ползването на споделена и константна памет в зависимост от ситуацията.



Фиг. 6 Архитектурата 10x, предоставя 240 процесора за изпълнение на 240 процеса едновременно, обединени в 30 мултипроцесора всеки от които има по 8 процесора и 1 АЛУ за изчисления с двойна прецизност (изт. NVIDIA)

Начинаещите програмисти следва да обърнат внимание на следните основни термини и тяхното значение:

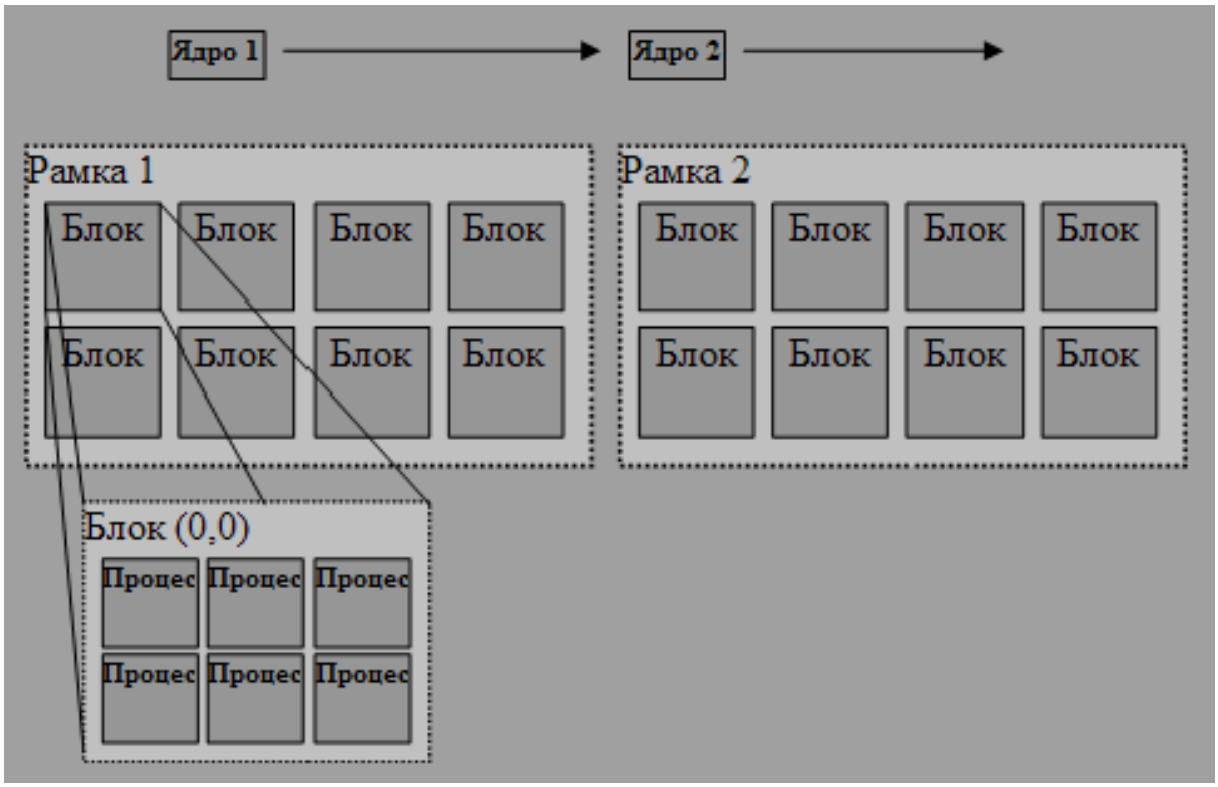
Half-Warp – това е група от 16 процеса чакащи в опашката за последователно изпълнение. Half-warp процесите се изпълняват заедно и са подравнени. Например процесите от 0->15 ще се намират в един и същи под блок, 16->31 в друг и т.н.

Warp – това е група от 32 съседни процеса, те принципно се изпълняват заедно в паралел. Това налага специфични методи за синхронизация, целящи изчакване на завършването на всички процеси на дадена итерация.

Block – блокът е набор от процеси, които правят едно и също нещо над различни елементи от масив с данни или едни и същи споделени данни**.** Поради технически причини минимум 192 процеса са нужни в един блок за оптимизация на закъсненията между тях. Един типичен блок има 256 или 512 процеса и дори 768. Следва да се има предвид, че процесите в един блок се синхронизират по-бързо и така по-лесно обменят данни помежду си чрез споделената памет.

Grid – "решетката" позволява да се създават макро блокове от обособени процесни блокове. Отделните блокове се синхронизират трудно, процесите в един блок също не могат да се синхронизират с процесите принадлежащи на друг блок. Отделни "grid" масиви се генерират за всяка специфична изчислителна задача, обхващащи определени данни, като

при това всеки различен грид може да се създава за нови и вече записани в паметта на картата данни (фиг. 7).



Фиг. 7 Разпределение на отделните процеси в блокове и ползване на различни рамки от процеси за различни изчислителни задачи.

Типове памет

За CUDA програмистите е от съществено значение да правят разлика мужду типовете памет в системата и съответно предимствата и недостатъците от нейното използване:

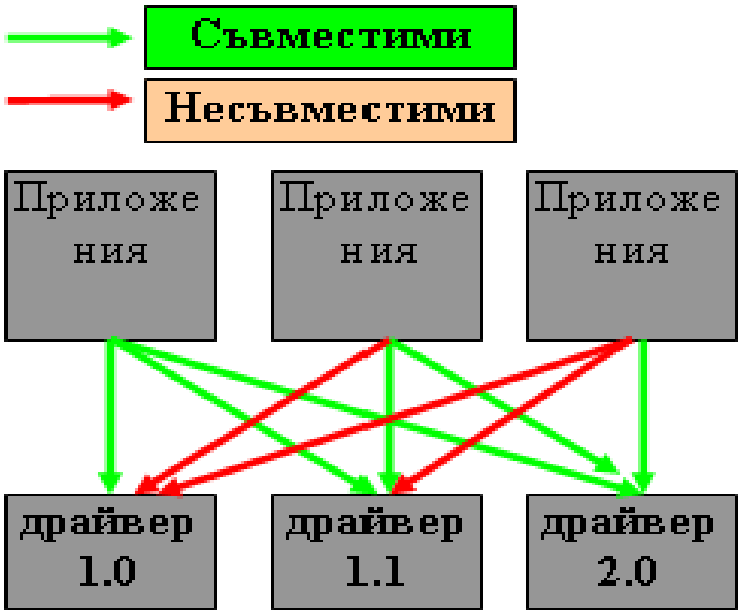
- **Глобална памет** – това е инсталираната физическа памет на графичната карта, чийто обем може да варира между 128 и 4000 MB. Организацията ѝ е линейна, в тази памет, както отделните мултипроцесори, така и хост компютърът могат да пишат и четат данни. Всички процеси могат да пишат и четат данни в глобалната памет, също така процеси стартирани на хост компютъра могат да пишат и четат от и в тази памет.
- **Споделена памет** – всеки един мултипроцесор има малък обем от памет наричан споделен, с размер 16KB. Тази памет се използва от отделните процеси, за бързо четене и запис на данни. Паметта се разпределя на блокове. Като пример ще кажем, че е възможно да имаме няколко блока работещи едновременно върху един и същи мултипроцесор. Разпределението на паметта между отделните блокове трябва да става по следния начин: блоковете = 16KB/ броя блокове. Изпълняваните в

рамките на един блок поцеси могат да комуникират помежду си като записват и четат адреси от тази памет. Тази памет е регистрова, и е около 100 пъти по-бърза от глобалната памет на картата.

- **Памет за текстури** – GPU притежава и текстурна памет, която може да се ползва при определени обстоятелства, като например за линейна филтрация на данни. Тази памет се кешира и е принципно само за четене.

Развитие на технологията CUDA

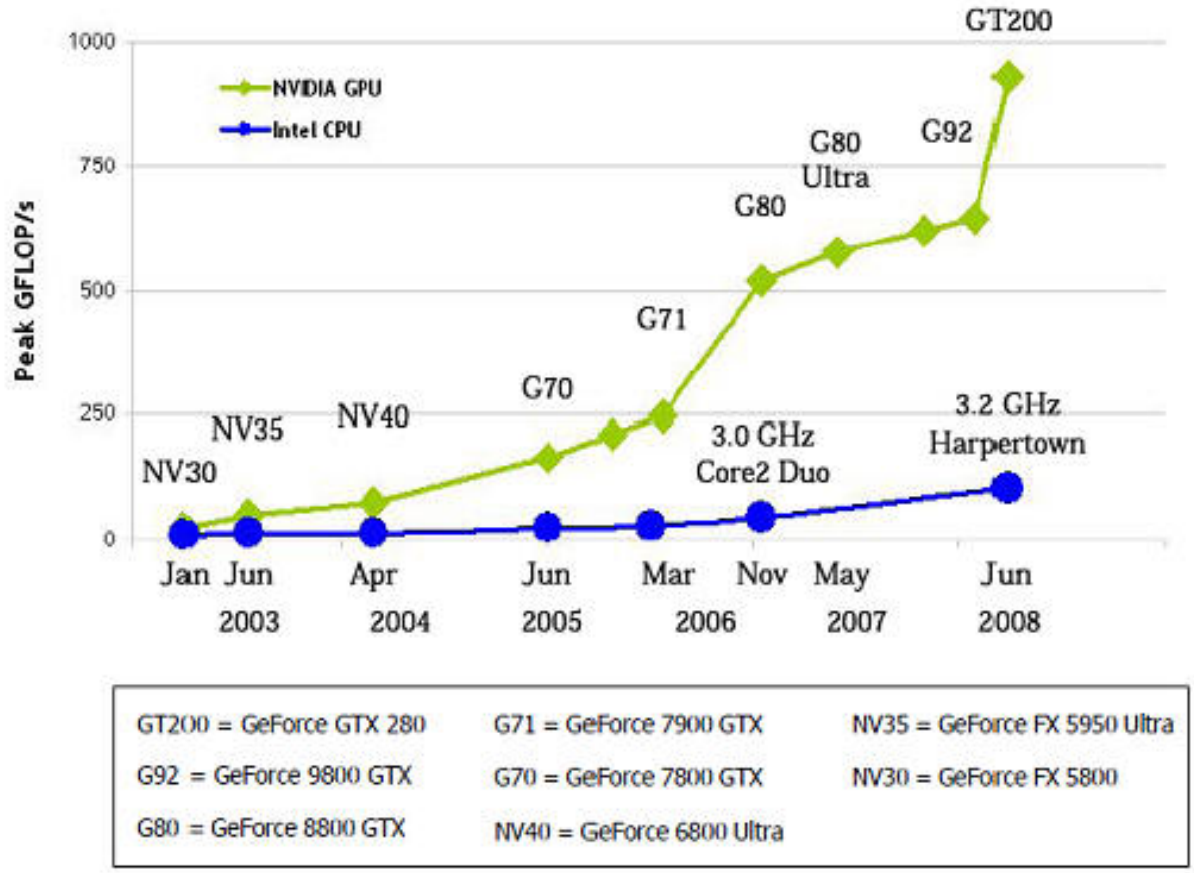
Проектът CUDA излиза за първи път на бял свят заедно с моделът G80 през ноември 2006г., като първата SDK е пусната на пазара през февруари 2007г. Първата версия с номер 1.0 се поддържа до излизането на професионалните клъстерни карти Tesla (юни 2007г.) и е съвместима с процесорите G80. Та е предназначена за пазара на изчислителни системи. Версия 1.1 въвежда функции за ползването на NVIDIA драйверите, като са поддържани карти след GeForce 8 и по-новите версия 169.xx. Това е ключов момент за програмистите, тъй като прави възможно ползването на CUDA върху всички видове видеокарти, а също така и като симулатор върху стандартен компютърен процесор. Следва да се отбележи, че много от преимуществата са достъпни само за 64-битовата версия на Windows. Версия CUDA 2.0 е готова заедно с картите GeForce GTX 200, като бета версията се въвежда през началото на 2008. Следващата версия поддържа: двойна прецизност на изчисленията, като хардуерно това е постижимо само на моделът GT200, а системата има поддръжка за Windows Vista (32- и 64-битова версия), Mac OS X и Linux. Следва да се има предвид, че разработените с всяка следваща версия приложения могат да работят само със същата и по-нови версии на CUDA SDK, като не могат да бъдат поддържани от драйверите на по-старата версия (Фиг. 8).



Фиг. 8 Поддръжка на приложенията е само в права посока в зависимост от версията на CUDA SDK (изт. NVIDIA) http://developer.nvidia.com/object/cuda_2_3_downloads.html

- **Версия 1.0:** пъддържа общ до 512 процеса изпълнявани в блок с размерност на обработвания блок 512x512x64 процуса, като максималният размер на всяка размерност в масива е 65535. Размерът се изчислява на базата 32 процеса, като броят регистри в мултипроцесор е 8192, а обемът на споделената памет е 16 килобайта в 16 банки. Ползва се 64 килобайта памет за съхранение на константи, локалната памет към всеки процес е до 16 килобайта, набор команди към мултипроцесор побиращи се в 8 килобайта, брой активни блокове до 8, брой основи към мултипроцесор 24, брой активни процеси към мултипроцесор 768, размерност на обработвания масив за: 1D 2^{13}, 2D с максимална ширина от 2^{16}, и височина от 2^{15}, 3D должина, ширина и височина до 2^{11}, обем на ядрения код до 2 милиона инструкции.
- **Версия 1.1:** въвежда поддръжката на т.н. 32 битови атомарни процесорни операции, позволяващи елементарни изчисления като събиране умножение, деление и др. над операнд намиращ се в глобалната или споделената памет и обратното записване на резултата в адреса на операнда. Тези операции са хардуерно обезпечени по такъв начин, че да не си пречат при изпълнението на множество едновременни процеси.
- **Версия 1.2:** поддържа атомарни операции над 64 битови стойности и изчисления с двойна прецизност в глобалната памет, като броят на регистри в един мултипроцесор е увеличен на 16384 и максималният брой активни процеси в мултипроцесор е 1024.
- **Версия 1.3:** поддържа аритметика с двойна прецизност на операнди с плаваща запетая.

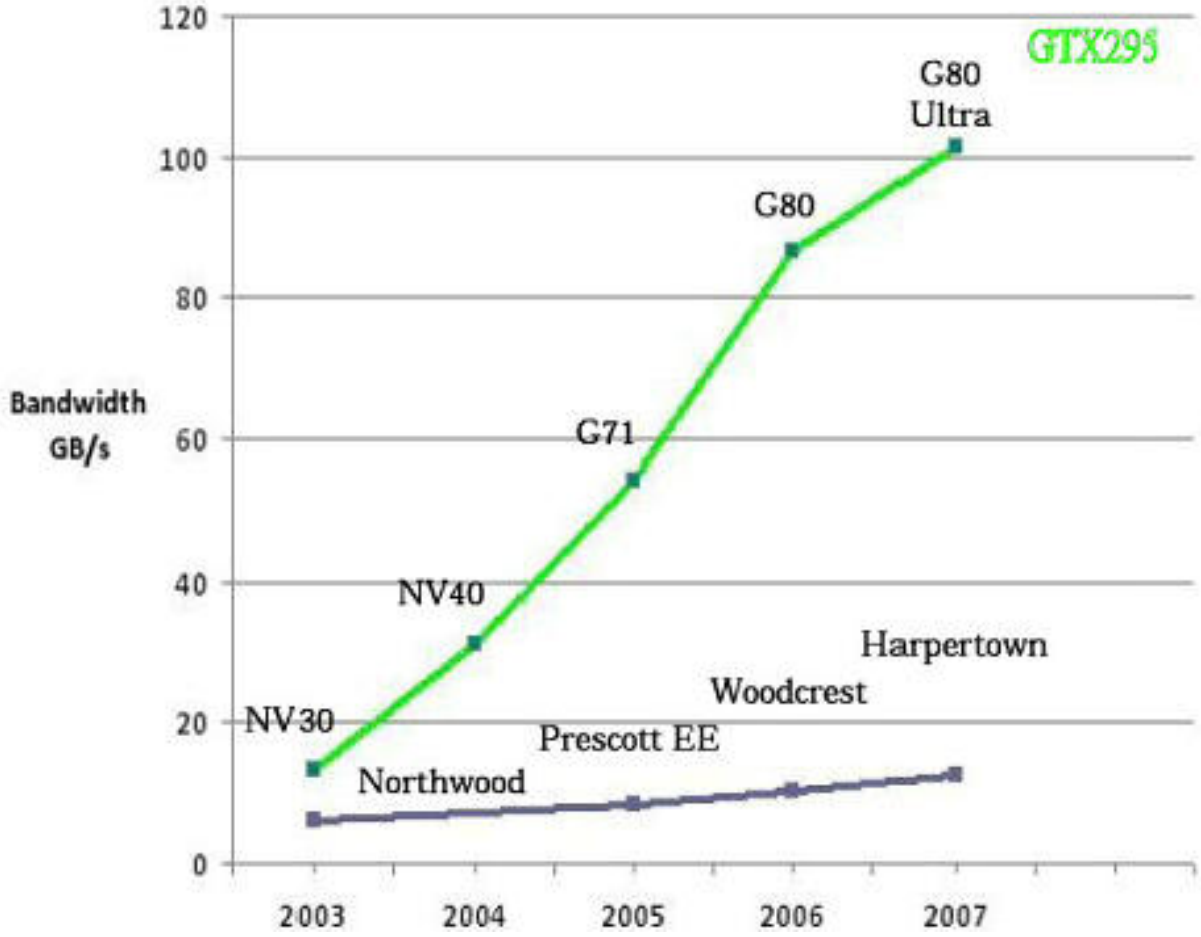
Интересно е да сравним как производителността на GeForce видео платките се изменя във времето. За един кратък период от 6 години технологията увеличава своята производителност с над 800%. Първият съществен скок се осъществява от моделът G70 последван от почти двойно увеличаване на производителността на модела G71. Следващите скокове са при моделите G80 и G200. На фиг. 9 е дадено сравнение на производителността със съвременни Intel процесори излезли от производство в същия интервал от време.



Фиг. 9 Сравнение на производителността между поредни GeForce продукти и процесори Intel (изт. NVIDIA)

Тъй като това са видео карти, нормално е да се очаква, че ползваните в тях технологии за съхранение и обмен на данни ще надвишават десетократно скоростите на трансфер на данните в кеша на процесора и между процесора и външната RAM памет на хост компютрите (Фиг. 10). Едновременно с това следва да се отчете, че NVIDIA забавят излизането на пазара на продукти с вградени атомични операции за работа с глобалната памет и споделените ресурси. Вместо това се представят нови платки имащи увеличена тактова честота и потребление на енергоконсумацията. В действителност за пазара на игри това не е нужно, но именно този клас платки са масово достъпни за

повечето програмисти. За разлика например една карта TESLA има до 4-5 пъти по-висока цена за същата производителност както една GeForce карта, с особената разлика, че поддържа всички видове атомарни операции и има вградени АЛУ за аритметични операции с двойна точност.



Фиг. 10 Сравнение на скоростта на обмен на данни между GeForce продуктите и Intel процесори. Заб. На графиката предоставена от NVIDIA не се вижда следващото поколение графична карта GTS295, която по спецификация следва да има приблизително 220GB/s скорост на обмен на данните. (изт. NVIDIA)

Програмиране с CUDA

За щастие CUDA значително опростява използването на GPU процесорите от обикновените програмисти и учените. Класическият подход на GPGPU архитектурата е, че програмистът ползва някои от достъпните приложни функции на драйверите или директно работейки с драйверите на видео картата може да осъществява даден вид изчисления. Този метод е труден защото изисква като минимум познания за работа с една от двете среди за 3D графика: OpenGL или Direct X, и то за тези версии на видео платките, които се поддържат. CUDA прави нещо съвсем различно - тя определя дадена част от

графичния чип само и единствено за създаването на обособен набор мултипроцесори, имащи достъп само до глобалната памет на картата. Следва да се има предвид факта, че видео картата, ако не се ползват продуктите от серията Tesla, заделят даден обем памет за съхранение и извеждане на данни върху графичния дисплей на компютъра. Тази памет е малко и обикновено е в границите на няколко десетки мегабайта. Останалата част от RAM паметта е достъпна, като глобална памет за GPU изчисления. CUDA поддържа следните типове приложения:

- Работещи с DirectX Compute OpenCL (Open Computing Language) драйвер
- Работещи директно с CUDA драйвера приложения : "C" компилатор за CUDA (nvcc), смесване на код със стандартни "C/C++" приложения
- Интегриране в програми на други езици: "Fortran", "Java", "Python"

И трите метода за разработка на приложения използват специфичната за дадената операционна система ядро и функции за организация на паралелни изчисления върху даден хардуер.

Начинаещите програмисти следва да ползват CUDA Developer SDK

В предлагания от NVIDIA комплект за програмиране се включват примери с програмен код, без които усвояването на новата технология е немислимо. Примерите включват:

- Паралелни сортировки;
- Умножение на матрици;
- Транспониране на матрици;
- Измерване на производителността с таймери;
- Паралелно събиране (scan) на големи и малки масиви;
- Конволюция на изображения;
- Уейвлет трансформация;
- Изобразяване на графика с OpenGL и Direct3D;
- Използване на CUDA BLAS FFT библиотека;
- CPU-GPU C и C++ миксиране на код;
- Монте-Карло алгоритми;
- Паралелен генератор на случайни числа;
- Паралелно изчисление на хистограми;
- Отстраняване на шум в изображения;
- Реализация на филтър на Собел за оконтуряване на изображения.

Експериментални резултати

Опитът, който споделяме, е свързан с изпълването на CUDA за нуждите на цифровата сигнална обработка на изображения в две направления:

статистически анализ и фурие трансформация и филтрация. Наличната система е **GPU 250 GTX** имаща **128 CUDA GPU** ядра, с графичен процесор работещ на 738 MHz, и основен процесор работещ на 1836 Mhz. Важно условие при създаването на ефективни приложения е да се има предвид изчислителните възможности на хардуера, които до момента са следните версии: 1.0, 1.1, 1.2 и 1.3.

Изчисление на глобална статистика

На сайта на NVIDIA са дадени примери за ползването на CUDA при реализацията на паралелни алгоритми за определяне глобална статистика на данни, в това число: средно значение, стандартно отклонение, минимум, максимум и средно квадратична грешка. Желаещите могат да се запознаят с [10], реализацията върху CUDA платформа налага особен вид оптимизации, за да стане възможно ползването на цялата налична пикова мощност на системата, както и скоростите за обмен на данни. Акцентът на готовите примери се поставя над производителността, а не толкова на произволната им реализация. Следва да се отбележи, че прилагането на класически подходи при подобен вид изчисления не води до постигане на по-висока производителност, тъй като графичните процесори са предвидени за работа с пиксели, проблемите следва да се сведат до интерпретация на данните, като изображения от пиксели.

Изчисление на хистограми

Ускоряването на изчислението на 2D хистограмите е над 26 пъти в сравнение с изчислението им върху класически Intel процесор. Тъй като процесът изисква зареждане и четене на данни от паметта на картата, общото ускорение е между 4-8 пъти [12]. Ефективността значително се повишава ако се буферират за обработка по-големи масиви данни (над 500MB) в паметта на видео картата. Това е особено полезно, когато става въпрос за изчисление на стотици хиляди хистограми, като например е случай с обработка на изображения от видео филми. Времето за обработка на една хистограма може да бъде до 4 секунди, при хистограма с 8 битови елементи и обем 128MB, докато същото изчисление на CUDA със 128 процесора и 1GB/RAM GeForce GTS 250 тази скорост се снижава до 0.91-0.95 секунди. Така обработката на едночасов 2MB видео поток ще отнеме 27 минути в изчислителната си част, а при обработка на стандартен процесор ще отнеме над 120 минути. Към това време следва да добавим и времето за отваряне и разкомпресиране на видео кадрите, които априори се прави на хост платформата. Тъй като CUDA работи асинхронно от хост платформата може да се постигне допълнително спестяване на време, като се стартират независимо процеси за четене на данни от дисковите масиви и тяхното обработване. Времето съществено може да се снижи в частта касаещи обработката, ако се ползват карти с повече паралелни

процесори каквито са: GeForce GTX 295. По-високи ускорения могат да се постигнат при тримерните хистограми, където веднъж зареден блокът от 100-200 поредни кадъра се обработва едновременно, и се пести време за зареждането на кадри във видео паметта на картата и връщане на резултати за довършителни изчисления. Съществени ускорения се постигат при време-пространствената обработка на изображения, особено при предварително сегментирани визуални сцени, който процес е неосъществим дори при много поточно програмиране върху много ядрени платформи [17].

Умножение на матрици

Много математически операции могат да бъдат сведени до операции с матрици. Умножението и транспонирането на правоъгълни матрици с размери 10x10 не е проблем за конвенционалните процесори, независимо от това ефективна сигнална обработка може да бъде постигната при ползване на филтърни ядра с размери над 25x25 елемента. Този процес може значително да бъде подобрен при ползване на CUDA. По-долу са дадени експериментални резултати от умножението на квадратни матрици на CUDA GTS250 и Intel Core Duo 2.4GHz с 1MB L2 кеш. Експериментите са правени с квадратни матрици, но резултатите са еднакво приложими и за правоъгълни матрици. Следва да се отбележи, че умножението на матрици изисква индексиране на глобалната памет на отделни блокове, и обработката им локално, като общия резултат се записва след умножението на всички блокове. Само така става възможно да се използва споделената памет към мултипроцесорите, която работи много по-бързо и позволява паралелизация на умножението и скалиране върху системи с различен брой мултипроцесори.

Размер на матрицата (float) 16 x размера	**Време за умножение на CUDA (128GPU) SHARED**	**Време за умножение на CUDA (128GPU) с обща памет**	**Време за умножение на CPU (Core II Duo) в куда програма в режим DEBUG**	**Време за умножение на CPU (Core Duo) на C**
50x50	70 ms	340 ms	12sec	2.5sec
100x100	170 ms	2750 ms	140sec	37sec
150x150	340 ms	8900 ms	440sec	120sec
200x200	800 ms	15484 ms	3350sec	670sec

Табл. 6 Експериментални данни от умножението на матрици върху GTS250 и Intel Core Duo 2.4GHz E2220.

Фурие трансформация

Фурие трансформацията е особено удобен инструмент при реализацията на системи за анализ на сигнали и изображения в тяхната честотна област. Чрез обработка на сигналите в честотното пространство много по-лесно и бързо се реализират алгоритми за линейна филтрация на 1D, 2D и 3D сигнали. Друго преимущество на CUDA е съвместимостта с MEX C компилатора за Matlab. Използването на примитиви написани за умножение на матрици и Фурие трансформация и Фурие конволюция за Matlab и LabVIEW позволява многократно повишаване бързодействието на изчисленията правени с тази програма. Постигнатото ускорение е от порядъка на 14 пъти (като от 230 секунди времето за изчисление на 2D спектър се свежда до 17 секунди използвайки CUDA MEX файл). Към настоящия момент за Matlab се поддържа версия CUDA 1.0 и компилаторите на Microsoft Visual Studio 2005 C/C++ или Visual Studio 8.

Заключение

Надявам се направеното тук общо представяне на спомогналите за развитието на GPGPU супер-компютрите движещи сили, както и заложените в GPU архитектурата концепции, да могат да дадат обща представа за възможностите на този тип системи. От извършените експерименти бе установено, че ползването на непрофесионални видео карти поддържащи технологията CUDA могат значително да подобрят бързодйствието на множество често ползвани алгоритми в повечето софтуерни приложения за сигнална обработка. Можем да очакваме, че CUDA продуктите ще продължат да навлизат все по-масово на пазара за потребителски софтуер, в това число финансов, медицински, статистически и други. Като подкрепа на ползата им за научни приложения може да се каже, че в Япония бе построен най-мощният CUDA клъстер, който се нарежда на 29 място сред всички световни супер компютри през 2009г. Ниските цени на професионалните версии TESLA позволяват ползването на големи изчислителни ресурси от самостоятелни учени и изследователи, като при това те могат да провеждат независими изследвания, без да се налага да чакат машинно време за достъп до големи учрежденски супер-компютри.

Библиографска справка

[1] http://boinc.berkeley.edu/

[2] The 3D Chipset Wars: A Chronicle of the Past, Present, and Future, Nicholas Paufler

[3] From Voodoo to GeForce: The Awesome History of 3D Graphics, 05.2009, Paul Lilly

[4] The OpenGLR Graphics System: A Specification (Version 3.2 (Core Profile) - December 7, 2009), Mark Segal, Kurt Akeley

[5] General calculations using graphics hardware,with application to interactive caustics, Chris Trendall and A. James Stewart, iMAGIS–GRAVIR/IMAG and University of Toronto
[6] Physically-based visual simulation on graphics hardware, SBN:1-58113-580-7 Authors:Mark J. Harris, Greg Coombe, Thorsten Scheuermann, Anselmo Lastra
[7] How many FLOPS are in game consoles?, J.Peddie, May 2008
[8] Rob Farber, "CUDA, Supercomputing for the Masses", 2008
[9] David Luebke, "The Democratization of Parallel Computing", High Performance Computing with CUDA, SUPERCOMPUTING November 2007
[10] NVIDIA CUDA Compute Unified Device Architecture, Version 1.0
[11] http://gpgpu.org/
[12] N.Satish, M.Harris, M.Garland, "Designing Efficient Sorting Algorithms for Manycore CPUs"
[13] V. Podlozhnyuk, "Histogram calculation in CUDA", 2007
[14] http://fastra2.ua.ac.be/?page_id=53
[15] "Родният суперкомпютър: беше ли в топ-100? ", www.technews.bg
[16] "What is CUDA? An Introduction", http://supercomputingblog.com/
[17] А.Арденска, "Паралелна обработка на информация в системи за обработка на изображения", 2009
[18] What is PCI Express? A Layman's guide to high speed PCI-E technology
by Lee Penrod