Ville Suvanto

GPGPU - PROSESSORIN KORVAAMINEN NÄYTÖNOHJAIMELLA YLEISHYÖDYLLISISSÄ OHJELMISSA

Tietotekniikan koulutusohjelma
Tietotekniikan suuntautumisvaihtoehto
2011

GPGPU - PROSESSORIN KORVAAMINEN NÄYTÖNOHJAIMELLA YLEISHYÖDYLLISISSÄ OHJELMISSA

Suvanto, Ville
Satakunnan ammattikorkeakoulu
Tietotekniikan koulutusohjelma
Yritys: io-media
Valvoja: Ekholm, Ari
Huhtikuu 2011
Sivumäärä: 43
Liitteitä: 1

Asiasanat: GPGPU, näytönohjain, grafiikkapiiri, prosessori

Opinnäytetyön tarkoituksena on tutustua GPGPU-tekniikkaan, jolla näytönohjaimia voidaan hyödyntää yleishyödyllisessä laskennassa ja näin siirtää laskentaa pois pelkästään prosessorin harteilta. Tekniikka on osoittanut potentiaalinsa viimeisen parin vuoden aikana, kun markkinoille on julkaistu näytönohjaimia, joiden grafiikkapiirit rakentuvat lukuisista pienistä suorittimista. Nämä pienet suorittimet ovat vahvoilla varsinkin rinnakkais- ja vektorilaskennassa.

Työn alussa tutkitaan näytönohjainten historiaa, nykyistä markkinatilannetta eri valmistajien ja toimitusmäärien valossa, nykynäytönohjaimen rakennetta sekä kahden johtavan erillisnäytönohjainvalmistajan grafiikkapiirien arkkitehtuuria. Seuraavaksi tutkinnan alla on GPGPU-tekniikka ja työssä tutkitaan, mitä erilaisia käyttökohteita tekniikalla on ja mitä etuja saavutetaan, kun laskentaa siirretään prosessorilta näytönohjaimelle.

Opinnäytetyössä tutkitaan myös tällä hetkellä markkinoilla olevia ohjelmointikirjastoja: CUDAa ja siihen perustuvaa PhysX:ää, OpenCL:ää, DirectComputea, ATI Streamia sekä Havok Physicsiä. Lopussa verrataan suorituskykyä ja tehonkulutusta, kun laskenta suoritetaan pelkästään prosessorilla ja kun laskentaa avustetaan näytönohjaimella.

GPGPU – REPLACING THE CPU WITH A GPU IN GENERAL-PURPOSE COMPUTING

Suvanto, Ville
Satakunnan ammattikorkeakoulu, Satakunta University of Applied Sciences
Degree Programme in Information Technology
Commissioned by io-media
Supervisor: Ekholm, Ari
April 2011
Number of pages: 43
Appendices: 1

Key words: GPGPU, graphics card, GPU, processor

The purpose of the thesis was to become acquainted with the GPGPU technology which can be used with graphics cards in general-purpose computing, thus redeploying computing away from CPU alone. This technology has shown its potential during the last few years since modern graphics cards have been released to the market. Those graphics cards are based on numerous small-scale processors which have proven their strength particularly in parallel and vector processing.

The work begins with an examination of the history of graphics cards, present market situation between different manufacturers and shipments, structure of modern graphics cards and the architectures of graphics processing units of two leading manufacturers of discrete graphics cards. Next the GPGPU technology is reviewed and clarified what different applications it has and what the benefits are when computing is redeployed from the CPU to a GPU.

In the thesis different programming libraries available in the market are studied. These include CUDA, PhysX which is based on CUDA, OpenCL, DirectCompute, ATI Stream and Havok Physics. In the end performance and power consumption are compared when computing is powered by the CPU only and when it is supported by a graphics card.

LYHENTEET

ALU	Arithmetic Logic Unit
AMD	Advanced Micro Devices, Inc.
APEX	Adaptive Physics Extensions
BOINC	Berkeley Open Infrastructure for Network Computing
CAL	Compute Abstraction Layer
CUDA	Compute Unified Device Architecture
FFT	Fast Fourier Transform
FPU	Floating Point Unit
GPC	Graphics Processing Cluster
GPGPU	General-Purpose Computing on Graphics Processing Unit
GPU	Graphics Processing Unit
FLOPS	Floating Point Operations Per Second
HAL	Hardware Abstraction Layer
HLSL	High Level Shader Language
IBM	International Business Machines
MDA	Monochrome Display Adapter
OpenCL	Open Computing Language
PC	Personal Computer
PCI	Peripheral Component Interconnect
PCI Express	Peripheral Component Interconnect Express
PPU	Physics Processing Unit
SDK	Software Development Kit
SFU	Special Function
SM	Streaming Multiprocessor
SVGA	Super Video Graphics Array
VGA	Video Graphics Array

SISÄLLYS

1 JOHDANTO 6
2 NÄYTÖNOHJAIMET 7
2.1 Näytönohjainten historia 7
2.2 Näytönohjainten markkinatilanne 9
2.3 Näytönohjaimen rakenne 10
2.4 Grafiikkapiirien arkkitehtuurit 12
2.4.1 NVIDIAn Fermi-arkkitehtuuri 13
2.4.2 AMD:n TeraScale 2 -arkkitehtuuri 17
3 GPGPU 19
3.1 GPGPU:n käyttökohteet 19
3.2 Näytönohjaimen edut prosessoriin verrattuna 21
4 OHJELMOINTIKIRJASTOT 23
4.1 NVIDIA CUDA 23
4.1.1 NVIDIA PhysX 24
4.1.2 PhysX:n arkkitehtuuri 25
4.2 Khronos Group OpenCL 26
4.3 Microsoft DirectCompute 28
4.4 AMD ATI Stream 28
4.5 Havok Physics 29
5 TESTIT 30
5.1 Testikokoonpano 30
5.2 PowerDVD 10 32
5.3 MediaShow Espresso 6.5 33
5.4 DirectCompute & OpenCL Benchmark 0.45 35
LÄHTEET 38
LIITTEET

1 JOHDANTO

Opinnäytetyössä tutustutaan GPGPU-tekniikkaan, jolla näytönohjaimia voidaan hyödyntää yleishyödyllisessä laskennassa ja näin siirtää laskentaa pois pelkästään prosessorin harteilta. Tekniikka on ehtinyt osoittaa potentiaalinsa viimeisen parin vuoden aikana, kun markkinoille on julkaistu sitä hyödyntäviä näytönohjaimia ja ohjelmistovalmistajat ovat havahtuneet sen käyttömahdollisuuksista.

Työn alussa tutkitaan näytönohjainten historiaa, nykyistä markkinatilannetta eri valmistajien ja toimitusmäärien valossa, nykynäytönohjaimen rakennetta sekä kahden johtavan erillisnäytönohjainvalmistajan grafiikkapiirien arkkitehtuuria. Seuraavaksi tutkinnan alle otetaan GPGPU-tekniikka ja tutkitaan, mitä erilaisia käyttökohteita sillä on ja mitä etuja saavutetaan, kun laskentaa siirretään prosessorilta näytönohjaimelle.

Opinnäytetyössä tutkitaan myös tällä hetkellä markkinoilla olevia ohjelmointikirjastoja ja lopussa verrataan suorituskykyä sekä tehonkulutusta, kun laskenta suoritetaan pelkästään prosessorilla ja kun laskentaa avustetaan näytönohjaimella.

Opinnäytetyö on tehty pitkäaikaiselle työantajalleni, io-medialle, jossa työn vastuuhenkilönä on toiminut Sampsa Kurri. Io-media vastaa Suomen suurimpiin tietokoneaiheisiin sivustoihin lukeutuvan Muropaketti.comin sisällöntuotannosta, ja opinnäytetyö tullaan julkaisemaan sivustolla artikkelina.

2 NÄYTÖNOHJAIMET

2.1 Näytönohjainten historia

Näytönohjainten historia ulottuu vuoteen 1981, jolloin IBM julkaisi ensimmäisen IBM PC -tietokoneen myötä MDA-näytönohjaimen, joka oli nykypäivän laitteisiin verrattuna hyvin alkeellinen. Tekstipohjainen laite oli varustettu neljän kilotavun muistilla ja kykeni esittämään ruudulla yhdellä värillä 80 kappaletta ja 25 riviä. 1980-luvun loppuun mennessä näytönohjaimet olivat kehittyneet siinä määrin, että ne kykenivät tuottamaan SVGA-tasoista 1024x768-resoluutioista kuvaa 8-bittisenä eli 256 värillä. (Answers.com)

1984 IBM julkaisi yhden markkinoiden ensimmäisistä näytönohjaimista, joka kykeni kiihdyttämään 2D-tilan ohella myös 3D:tä. Näytönohjain tuki 256 väriä ja 640x480-resoluutiota 60 hertsin virkistystaajuudella. Yli 4000 tuhannen dollarin hinta ja heikko ohjelmistotuki eivät kuitenkaan nostaneet sitä suureen suosioon.

1991 S3 julkaisi 86C911-grafiikkapiirin, joka tuki ensimmäisenä yhden piirin ratkaisuna 2D-kiihdytystä, ja vuonna 1995 esiteltiin ensimmäiset kuluttajille suunnatut näytönohjaimet, jotka hallitsivat myös 3D-materiaalin kiihdyttämistä. Aallonharjalla kulkivat muun muassa ATI, S3, Matrox ja Cirrus Logic, mutta kaksi vuotta myöhemmin 3dfx-niminen yritys loi perustan nykyaikaisille näytönohjaimille Voodoo-tuoteperheellään. Suosion saattelemana 3dfx:n varpaille astui samaan markkinarakoon myös muun muassa NVIDIA TNT- ja TNT2-tuoteperheiden näytönohjaimillaan. 3D-grafiikkapiirien yleistymistä vauhdittavat uudet kuluttajille suunnatut pelikonsolit, kuten Sony Playstation ja Nintendo 64, jotka tukivat 3D-grafiikan kiihdyttämistä.

Siirryttäessä kohti vuosituhannen vaihdetta näytönohjainten merkitys varsinkin pelikäytössä korostui entisestään ja niiden kehittyessä käytössä olleen PCI-väyläarkkitehtuurin 133 Mt/s:n kaistanleveydestä tuli rajoittava tekijä. Intel ratkaisi

ongelman kehittämällä 266 Mt/s:n kaistanleveyden tarjonneen AGP 1.0:n. AGP-standardi päivittyi neljään otteeseen ja viimeisimmäksi jäänyt 3.5-versio tarjosi 2133 Mt/s:n kaistanleveyden. 2000-luvun alussa NVIDIA dominoi erillisnäytönohjain-markkinoita GeForce-tuoteperheen näytönohjaimilla, mutta nykypäivänä tilanne on hyvin tasainen NVIDIAn ja sen pahimman kilpakumppanin AMD:n kesken. Edeltävässä kappaleessa mainittu ATI siirtyi yritysoston myötä prosessoreita valmistavalle AMD:lle vuonna 2006 ja NVIDIA osti 3dfx:n konkurssista jääneet rippeet vuonna 2002. (Wikipedia. 2010)

Nykypäivän näytönohjaimet ovat kehittyneet entistä enemmän GPGPU-suuntaan tarjoamalla massiivisen määrän Shader-prosessoreita, joita NVIDIA kutsuu CUDA-ytimiksi ja AMD Stream-prosessoreiksi. Molemmat valmistajat tarjoavat näytönohjaimia, jotka tukevat uusimpia DirectX 11- ja OpenGL 4.0 -rajapintoja. Väyläarkkitehtuuri on siirtynyt AGP:stä sarjatyyppiseen PCI Expressiin, jonka nykyisin yleisimmin käytettävä 2.0-standardi tarjoaa PCI Express x16 -liitännän myötä molempiin suuntiin kahdeksan gigatavun kaistanleveyden sekunnissa.

Rajapintojen osalta Khronos Groupin kehittämä ilmainen OpenGL oli aluksi 1990-luvun alussa käytössä ainoastaan ammattikäytössä, mutta vuosikymmenen loppua kohti rajapinnan käyttö yleistyi myös kuluttajapuolella. Vuosikymmenen lopulla OpenGL sai kilpailijan Microsoftin DirectX:stä, joka on kokoelma ohjelmarajapintoja ja käsittää tuen muun muassa 2D- ja 3D-grafiikkakiihdytykselle. Tähän päivään mennessä OpenGL on edennyt 4.0-versioon ja DirectX 11-versioon. DirectX:n eri versiot ovat onnistuneet 2000-luvun puolella haalimaan itselleen suurimman osan markkinoista.

2.2 Näytönohjainten markkinatilanne

Taulukko 1. Grafiikkapiirien toimitukset ja kasvut vuosina 2006-2011 (kuva: JPR)

		2006	2007	2008	2009	2010
Total Graphics Chips CAGR '06-'11:	5.53%	316.5	351.7	373.1	414.2	432.2

Grafiikkapiirien toimitusmäärät ovat kasvaneet jatkuvasti ja yli kaksinkertaistuneet vuodesta 2003 (Jon Peddie Research. 2010) vuoteen 2010 mennessä (Taulukko 1). Siinä missä vuonna 2003 piirejä toimitettiin kaikkien eri valmistajien toimesta yhteensä 217,1 miljoonaa kappaletta, vuonna 2006 piirejä toimitettiin jo yli 300 miljoonaa kappaletta ja vuonna 2010 yli 400 miljoonaa kappaletta. (Jon Peddie Research. 2011)

Jon Peddie Research -tutkimusyhtiön julkaisemat grafiikkapiirimarkkinoiden prosentuaalista kasvua osoittavat lukemat kertovat markkinoiden kasvavan vuosi vuodelta. Vuodesta 2003 lähtien ainoastaan vuonna 2008 kasvu jäi alle 10 prosentin, mikä on seurausta kyseisen vuoden finanssikriisistä. Markkinat ovat elpyneet nopeasti ja vuonna 2009 kehitystä tapahtui 11,0 ja sitä seuraavana vuonna 4,3 prosenttia.

Taulukko 2. Grafiikkapiirivalmistajien markkinaosuudet Q4/2010 (kuva: JPR)

Vendor	This Quarter Market share	last Quarter Market share	Unit Growth Qtr-Qtr	This quarter last year Market share	Growth Yr-Y
AMD	24.2%	23.0%	2.3%	21.7%	11.2%
Intel	52.5%	55.2%	-7.3%	51.1%	2.9%
Nvidia	22.5%	21.0%	4.1%	26.5%	-15.1%
Matrox	0.1%	0.1%	0.0%	0.0%	30.1%
SiS	0.0%	0.0%	-100.0%	0.0%	-100.0%
VIA/S3	0.8%	0.8%	-1.9%	0.7%	19.5%
Total	100.0%	100.0%	-2.6%	100.0%	0.0%

Grafiikkapiirivalmistajista Intel oli vuoden 2010 viimeisellä neljänneksellä markkinoiden suurin tekijä 52,5 prosentin markkinaosuudella (Taulukko 2). AMD oli listan toinen 24,2 prosentin osuudella ja NVIDIA kolmas 22,5 prosentin osuudella. Kolme suurinta valmistajaa hallitsevat markkinoita käytännössä täysin, sillä Jon Peddie Re-

searchin julkaisemista tiedoista selviää, että Matroxin, SiS:n ja VIA/S3:n osuudet jäävät ainoastaan 0,1; 0,0 ja 0,8 prosenttiin.

Tilannetta vääristää hieman se seikka, että Intel saavuttaa markkinajohtajan aseman omiin piirisarjoihinsa ja prosessoreihinsa integroiduilla grafiikkaohjaimilla. Kyseiset grafiikkaohjaimet tarjoavat huomattavasti heikomman suorituskykytason kuin erillisnäytönohjaimet, joista AMD:n ja NVIDIAn markkinaosuudet kertyvät suurimmaksi osin. Pelkästään erillisnäytönohjainmarkkinoita tutkiessa markkinoilla on ainoastaan kaksi suurta valmistajaa, AMD ja NVIDIA.

2.3 Näytönohjaimen rakenne

Kuva 1. NVIDIA GeForce GTX 480 -näytönohjain

Markkinoilta löytyvien näytönohjainten rakenne on nykypäivänä valmistajasta riippumatta hyvin samanlainen. Edestäpäin katsottuna piirilevystä ei ole näkyvissä kuin yläpuolen NVIDIA SLI- tai AMD CrossFireX -liittimet, joiden avulla useampi näytönohjain voidaan kytkeä toimimaan rinnakkain suorituskyvyn parantamiseksi, sekä alhaalla PCI Express x16 -liitin.

Piirilevyn peittää yleensä kahden korttipaikan korkuinen jäähdytysratkaisu, jonka siiliosa on valmistettu alumiinista tai kuparista. Lämmönleviämistä siilin koko alueelle tehostavat lämpöputket, jotka siirtävät energiaa suljetussa kierrossa olevan nesteen yhtäjaksoisella haihtumis- ja lauhtumisprosessilla. Lämpöputkiin tuleva lämpö haihduttaa työaineena olevan nesteen ja syntynyt höyry kulkeutuu lämpöputkien viileämpiin osiin, joissa se tiivistyy nesteeksi luovuttaen lämpöä seinämien läpi. Neste palaa takaisin haihduttimelle ja uusi kierto alkaa. (Wikipedia. 2010)

Nykypäivän näytönohjaimet tuottavat siinä määrin lämpöä, että siili ei yksinään kykene viilentämään piirejä tarpeeksi ja jäähdytyksen tehostamiseksi apuna on yksi tai useampi tuuletin. Tuulettimen tarkoituksena on työntää viileää ilmaa jäähdytysratkaisuun, jolloin se sitoo siilien lämpöä ja poistuu näytönohjaimesta lämminneenä.

PCI Express x16 -liitin kykenee toimittamaan maksimissaan 75 wattia tehoa, mutta ainoastaan harva nykypäivän näytönohjain kuluttaa maksimissaan edellä mainitun arvon verran tehoa. Suurimmasta osaa näytönohjaimia löytyy yksi tai kaksi suoraan virtalähteeseen yhteydessä olevaa lisävirtaliitintä, joista yleisimmin käytettävät kuusipinniset liittimet kykenevät toimittamaan kukin 150 wattia tehoa. (Intel. 2010)

Näytönohjaimien edestä löytyvät korttipaikkaliittimet, jotka näkyvät tietokonekotelon ulkopuolelle ja joihin liitetään näyttölaitteiden kaapeleita. Tällä hetkellä yleisimmin käytössä ovat kaksilinkkiset DVI-I-liittimet, jotka kykenevät siirtämään näytölle maksimissaan 2560x1600-resoluutiosta kuvaa. Lisäksi monista näytönohjaimista löytyy HDMI-, mini-HDMI-, DisplayPort- ja Mini DisplayPort -liittimiä, joiden resoluutiot vaihtelevat HDMI-liittimien maksimista 1920x1080:stä DisplayPort-liittimien maksimiin 2560x1600:aan.

Kuva 2. GeForce GTX 480 ilman jäähdytintä

Jäähdytysjärjestelmän alla oleva paljas piirilevy on komponenttitäytteinen ja suurin osa komponenteista on pienikokoisia pintaliitoskomponentteja. Merkittävin piiri on grafiikkapiiri, jonka piisiru on GeForce GTX 480:n tapauksessa suojattu kuparista valmistetulla shimmillä. Grafiikkapiirin ympäriltä löytyvät muistipiirit ja näytönohjaimen takaosassa on kondensaattoreita, mosfetteja ja ohjainpiirejä vastaamassa viransyötöstä.

2.4 Grafiikkapiirien arkkitehtuurit

Näytönohjaimen suorituskykyyn vaikuttavia tekijöitä on useita ja niistä merkittävimpiä ovat grafiikkapiirin, Shader-prosessoreiden ja muistipiirien kellotaajuus sekä muistien kaistanleveys ja kapasiteetti. Pelkästään edellä mainittuja seikkoja parantamalla raja tulee hyvin nopeasti vastaan. Tästä syystä valmistajat suunnittelevat jatkuvasti uusia arkkitehtuureja, jotka eivät ole edellä mainittujen seikkojen tavoin yhtä näkyviä käyttäjälle, mutta suorituskyvyn ja ominaisuuksien parantamisen kannalta elintärkeitä.

Näytönohjainvalmistajat tuovat markkinoille karkeasti otettuna 12-18 kuukauden välein uuteen arkkitehtuuriin perustuvia näytönohjaimia eri tuotekategorioihin. Nykyaikaiset arkkitehtuurit ovat joustavia sen suhteen, että yhdestä suunnittelusta saadaan

helposti jalostettua eri hintaluokan näytönohjaimia kytkemällä tiettyjä grafiikkapiirin osia, kuten Shader-prosessoreita, pois käytöstä.

Grafiikkapiirin arkkitehtuuri on verrattavissa ihmisen tapauksessa luurankoon. Arkkitehtuuri on grafiikkapiirin pohjalla, mutta lopullisen suorituskyvyn määrittelee se, kuinka pitkälle näytönohjainvalmistaja hyödyntää sitä. Ihmisellä luuranko määrittelee perusmuodot, mutta lihakset jne. antavat lopullisen muodon kokonaisuudesta. Näytönohjaimien tapauksessa lihakset vastaavat kellotaajuuksia ja muun muassa muistikapasiteettia.

NVIDIA ja AMD ovat kaksi ehdottomasti suurinta valmistajaa erillisnäytönohjainten saralla ja tutustumme yritysten uusimpiin arkkitehtuureihin, joka NVIDIAn tapauksessa tunnetaan nimellä Fermi ja AMD:n tapauksessa TeraScale 2.

2.4.1 NVIDIAn Fermi-arkkitehtuuri

Kuva 3. Kaaviokuva Fermi-arkkitehtuurista (kuva: NVIDIA)

Fermi on toistaiseksi markkinoiden monimutkaisin arkkitehtuuri, mikä osaltaan siivitti sen useisiin myöhästymisiin. NVIDIAn toimitusjohtaja Jen-Hsun Huang esitteli Fermi-arkkitehtuurin jo vuoden 2009 lokakuussa, mutta varsinaiset näytönohjaimet julkaistiin vasta puoli vuotta myöhemmin. Kyseessä on 40 nanometrin prosessilla valmistettava GF100-koodinimellinen grafiikkapiiri, joka rakentuu 3,0 miljardista transistorista ja itse piisirulla on kokoa 525 neliömillimetriä.

Kuva 3 esittää Fermi-arkkitehtuuriin perustuvaa GF100-grafiikkapiiriä ja kuvassa näkyy selvästi neljä GPC:tä, joissa kussakin on neljä SM-moduulia. Kun siirrytään edelleen lähempään tutkintaan, kunkin SM:n sisällä on 32 CUDA-ydintä, joten arkkitehtuuri tarjoaa maksimissaan 512 CUDA-ydintä. GPC-yksiköiden välissä on sinisellä värillä korostettu L2-välimuisti, jota on koko grafiikkapiirissä 768 kilotavua. L2-välimuisti kykenee käsittelemään luku- ja kirjoitusoperaatioita ja sen vastuulla ovat GF100:n lataus-, tallennus- sekä tekstuuripyynnöt. (NVIDIA. 2009, 7-8)

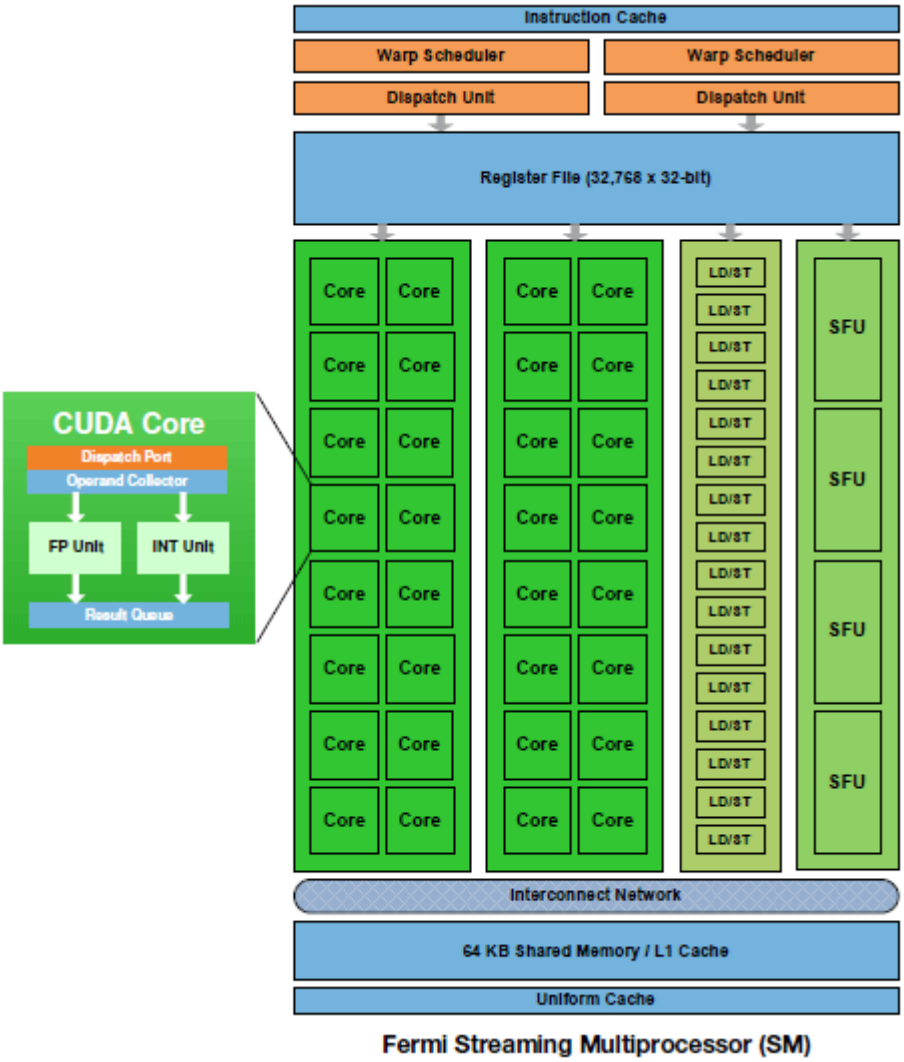
Kaaviokuvan reunoilla näkyy kuusi muistiohjainta (Memory Controller), jotka tukevat GDDR5-muisteja ja ovat 64-bittisiä, joten Fermi tarjoaa yhteensä 384-bittisen väylän muistipiireille. Ylimpänä oleva Host Interface vastaanottaa prosessorikäskyt ja oranssina kuvassa näkyvän GigaThread Enginen tehtävä on noutaa dataa keskusmuistista ja kopioida se ruutupuskuriin. Lisäksi se luo ja lähettää säikeet SM-moduuleille.

Kunkin muistiohjaimen vastuulla on kahdeksan ROP-yksikön nippu, joten yhteensä Fermissä on 48 ROP-yksikköä hoitamassa reunojenpehmennystä sekä muun muassa pikseleiden sekoittamista. Lisäksi Fermi tarjoaa neljä rasterointiyksikköä, 16 geometriayksikköä sekä 64 tekstuuriyksikköä.

Jokaisella SM-moduulilla on oma PolyMorph Engine, joka on NVIDIAn skaalautuva geometriayksikkö. Yksiköt toimivat viiden vaiheen perusteella, jotka ovat Vertex Fetch, Tessellator, Viewport Transform, Attribute Setup ja Stream Output. Vaiheet käsitellään edellä mainitussa järjestyksessä ja kunkin vaiheen välissä tulokset toimitetaan SM-moduulille, joka suorittaa shader-käskyn ja palauttaa saadun tuloksen seu-

raavalle vaiheelle. Kaikkien vaiheiden läpikäynnin jälkeen tulos siirretään Raster Enginelle.

Kullakin GPC:llä on käytössä yksi oma Raster Engine, joka rakentuu Edge Setup-, Rasterizer- ja Z-Cull-vaiheista. Ensimmäisessä vaiheessa verteksien sijainnit haetaan ja kolmion reunojen yhtälöt lasketaan. Tämän jälkeen poistetaan kolmiot, jotka eivät ole kohti ruutua. Rasterizer-yksikkö laskee pikseleiden peiton ja sen lopputuote lähetetään Z-Cull-yksikölle, joka poistaa ruutupuskurissa olevien pikseleiden takana piilossa olevat pikselit. (Kurri S. 9.11.2010, 2)



Kuva 4. Kaaviokuva Fermi-arkkitehtuurin SM-moduulista (Kuva: NVIDIA)

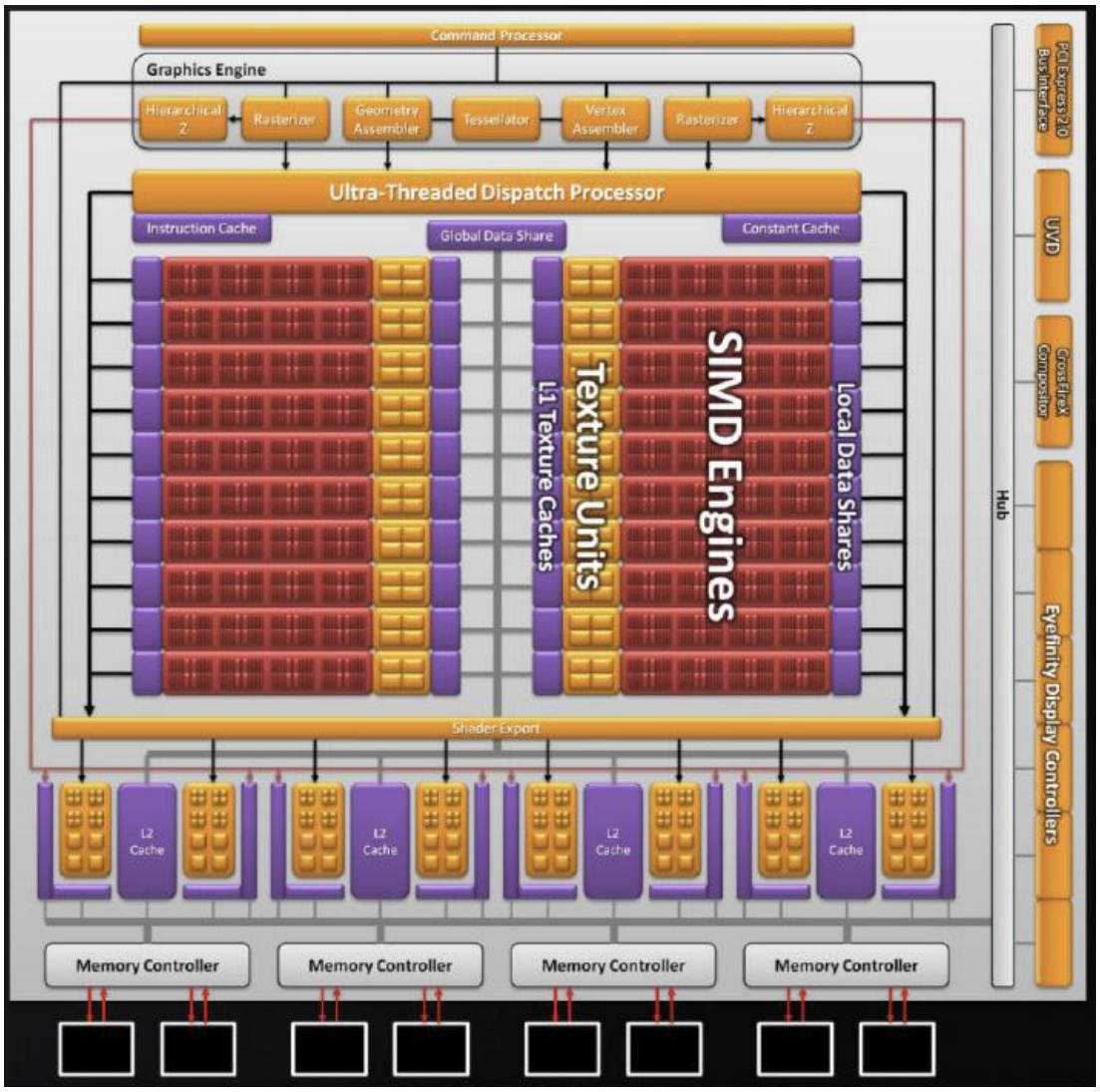
Jokaisessa SM-moduulissa on 32 kappaletta kuvassa kirkkaan vihreällä pohjalla näkyvää CUDA-ydintä, joissa on oma liukuhihnoitettu ALU- ja FPU-yksikkö. Lisäksi

kussakin SM-moduulissa on 16 lataus- ja tallennusyksiköitä (LD/ST), jotka mahdollistavat 16 säikeelle per kellojakso lähde- ja kohdeosoitteiden laskennan, sekä neljä SFU-yksikköä transkendenttikäskyjä varten.

GigaThread Enginen lähettämät säikeet käsitellään SM-moduulien Warp Scheduler -moduuleissa, joiden tehtävänä on jakaa 32 säikeen ryhmät CUDA-ytimille ja muille suoritusyksiköille. Kullakin SM-moduulilla on käytössä 64 kilotavun suuruinen L1-välimuisti, joten yhteensä ensimmäisen tason välimuistia on maksimissaan 1024 kilotavua eli yksi megatavu. L1-välimuisti voidaan konfiguroida 48 kilotavun jaetuksi muistiksi ja 16 kilotavun L1-välimuistiksi tai vaihtoehtoisesti 16 kilotavun jaetuksi muistiksi ja 48 kilotavun L1-välimuistiksi.

Kussakin SM-moduulissa on myös neljä tekstuuriyksikköä, joiden tehtävänä on kussakin kellojaksossa laskea tekstuuriosoite ja hakea neljä tekstuurinäytettä. Tulokset voidaan palauttaa bilineaari-, trilineaari- tai anisotrooppisella suodatuksella sekä ei-suodatettuna. (Kurri S. 9.11.2010, 2)

2.4.2 AMD:n TeraScale 2 -arkkitehtuuri



Kuva 5. Kaaviokuva TeraScale 2 -arkkitehtuurista (kuva: AMD)

AMD julkaisi TeraScale 2 -arkkitehtuurin syyskuussa 2009 samalla, kun se julkaisi Cypress-koodinimelliseen grafiikkapiiriin perustuvat näytönohjaimet. Cypress-grafiikkapiirit valmistetaan 40 nanometrin prosessilla ja ne rakentuvat 1,19 miljardista transistorista, joten NVIDIAn GF100:aan verrattuna transistoreita on huomattavan paljon vähemmän. 525 neliömillimetrin pinta-alallisen piisirun sijaan Cypressin pinta-ala on kuitenkin ainoastaan 334 neliömillimetriä, mikä tekee siitä helpomman ja halvemman valmistaa.

Kaaviokuvaan (Kuva 5) on merkitty punaisella SIMD-moottorit, joita on yhteensä 20 kappaletta. Kussakin vaakatasossa olevassa SIMD-moottorissa on 16 säieprosessoria,

joksi AMD kutsuu viiden Stream-prosessorin nippua. Kaiken kaikkiaan TeraScale 2 -arkkitehtuuri tarjoaa yhteensä 1600 Stream-prosessoria. Jokaisessa SIMD-moottorissa on lisäksi neljä keltaisella kuvattua tekstuuriyksikköä, joiden yhteismäärä nousee näin ollen 20 kappaleeseen, sekä kahdeksan kilotavua L1-välimuistia ja 32 kilotavua paikallista datamuistia. Kaikilla SIMD-moottoreilla on käytössä 64 kilotavun suuruinen datasäilö. (Kurri S. 27.9.2009, 2)

Kaaviokuvan (Kuva 5) yläosassa on keltaisilla laatikoilla kuvattu grafiikkamoottoria, jossa on käytössä kaksi rasterointiyksikköä tehostamassa polygonimallien muuntamisnopeutta pikseleiksi. TeraScale 2 -arkkitehtuurin merkittävin uudistus on tesselointiyksikkö, jonka myötä vähän polygoneja käsittävistä pinnoista, hahmoista sekä esimerkiksi animaatioista kyetään luomaan alkuperäistä yksityiskohtaisempia kasvattamalla polygonimäärää. Rasterointi- ja tesselointiyksiköiden ohella grafiikkamoottorista löytyy myös kaksi hierarkia-Z-yksikköä sekä yksi geometriakääntäjä- ja vertekskääntäjäyksikkö.

SIMD-yksiköiden välimuistien ohella TeraScale 2 käsittää myös L2-välimuistia, jota on pyhitetty kullekin muistiohjaimelle 128 kilotavua. GDDR5-muisteja tukevia muistiohjaimia on neljä kappaletta ja kukin niistä on 64-bittinen, joten muistipiireillä on käytössä 256-bittinen väylä. (Wasson, S. 23.9.2009, 5-6)

3 GPGPU

GPGPU on tekniikka, jonka tarkoituksena on siirtää aiemmin pelkästään prosessorin vastuulla olevaa laskentaa myös näytönohjaimelle. Näytönohjainten grafiikkapiirit on suunniteltu aiemmin huomattavasti rajoitetumpaan käyttöön kuin prosessorit, mutta näytönohjainkehitys on jo muutaman vuoden ajan suunnannut kohti yleiskäyttöisempää linjaa. Prosessoreihin verrattuna nykypäivän grafiikkapiirit ovat vahvoilla esimerkiksi rinnakkais- ja vektorilaskentaan liittyvillä osa-alueilla. Pienien käyttäjäryhmien lisäksi kiinnostus on herännyt peruskuluttajien ja tutkijoiden keskuudessa, kun markkinoille on saapunut helposti saataville GPGPU:ta tukevia laitteita ja ohjelmistoja.

Suorituskykyyn liittyvien seikkojen ohella GPGPU on herättänyt kiinnostusta myös kustannusten saralla, sillä parhaimmissa tapauksissa muutaman kymmenen euron hintainen näytönohjain kykenee tarjoamaan huomattavasti paremman vastineen rahoille kuin tuhannen euron hintainen prosessori.

3.1 GPGPU:n käyttökohteet

Fysiikkamallinnuksessa laskut ovat usein yksinkertaisia, mutta riippuvat toisista laskutoimenpiteistä. Uudet näytönohjaimet kykenevät suorittamaan valtavan määrän laskutoimituksia sekunnissa ja rinnakkaisuuteen perustuvien arkkitehtuurien myötä niissä riittää resursseja laskemaan laskutoimenpiteitä suurilla nopeuksilla ja ne kykenevät mukautumaan toisten ytimien tuottamiin tuloksiin.

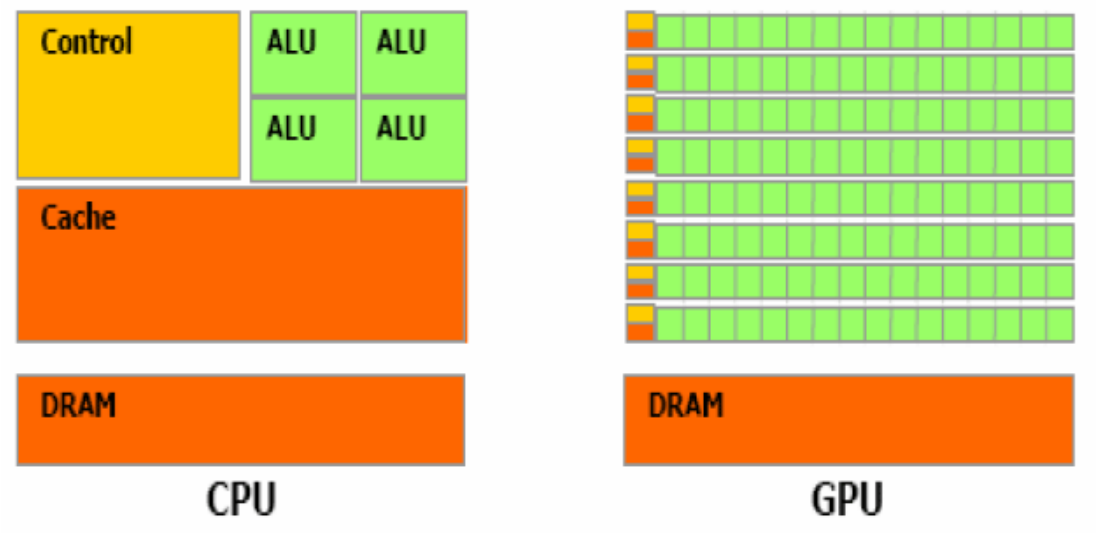
Nykyiset grafiikan osalta huippuunsa hiotut pelit kuluttavat prosessoritehoa pelin perustoimintojen lisäksi muun muassa mahdollisimman realistisen tekoälyn ja fysiikanmallinnuksen luomiseen. Rajallisten resurssien vuoksi useat paljon fysiikkamallinnusta vaativat efektit, kuten räjähdykset ja realistisen näköiset vesivirtaukset, ovat loistaneet poissaolollaan.

Kuluttajien kannalta yksi hyvin näkyvä osa-alue, joka hyötyy GPGPU:sta, on teräväpiirtoelokuvia toistavat multimediasovellukset. Täyden 1080p-resoluution (1920x1080) elokuvien toistaminen vaatii paljon suorituskykyä ja pelkällä prosessorilla toistaessa se asettaa kohtuullisen kovat vaatimukset laitteistolle. Suurimpien ohjelmistotalojen multimediatoisto-ohjelmat ovat tukeneet jo jonkin aikaa näytönohjainkiihdytystä, mikä käytännössä tarkoittaa tekniikkaa tukevan näytönohjaimen omistajalle, ettei prosessorilta vaadita elokuvien toistamiseen juuri lainkaan suorituskykyä. Toistoon saattaa riittää pelkästään emolevyn piirisarjaan integroitu grafiikkaohjain ja erillisnäytönohjaimen osalta kaikki nykyään markkinoilta saatavilla olevat mallit jaksavat toistaa sulavasti raskaimpiakin elokuvia.

Ammattikäytössä GPGPU:n käyttökohteita rajoittaa ainoastaan mielikuvitus ja tekniikkaa voidaan hyödyntää esimerkiksi matemaattisten laskutoimitusten suorittamisessa, sääennusteiden laskemisessa, moottoreiden sekä nesteiden käyttäytymisen mallintamisessa, kryptografiaan, kuvankäsittelyyn, säteenseurantaan, kasvojen ja erilaisten objektien tunnistukseen sekä kyseenalaisena alueena salausten purkamiseen.

Valmiita sovelluksia löytyy jo aiemmin mainittujen pelien ja multimediatoisto-ohjelmien ohella myös ohjelmistotalo Adobelta, jonka Photoshop-kuvankäsittelyohjelma ja Acrobat-PDF-ohjelma tukevat näytönohjainkiihdytystä. Myös hajautettua laskentaa hyväksi käyttävä Folding@home-ohjelmaa voi käyttää näytönohjainkiihdytteisesti ja ohjelman tarkoituksena on laskea proteiinien laskostumista ja auttaa lääketeollisuutta uusien hoitomenetelmien kehittämisessä. Yleisesti käytössä oleva numeerisen laskennan MATLAB-ohjelmisto, CAD-ohjelmisto AutoCAD sekä muun muassa 3D-mallinnusohjelma 3ds Max ovat ainoastaan muutamia esimerkkejä käyttökohteista, joissa näytönohjaimella kyetään nopeuttamaan ohjelman toimintaa pelkkään prosessoriin verrattuna.

3.2 Näytönohjaimen edut prosessoriin verrattuna



Kuva 6. Kaaviokuvat prosessorin ja grafiikkapiirin pääeroista (Kuva: AMD)

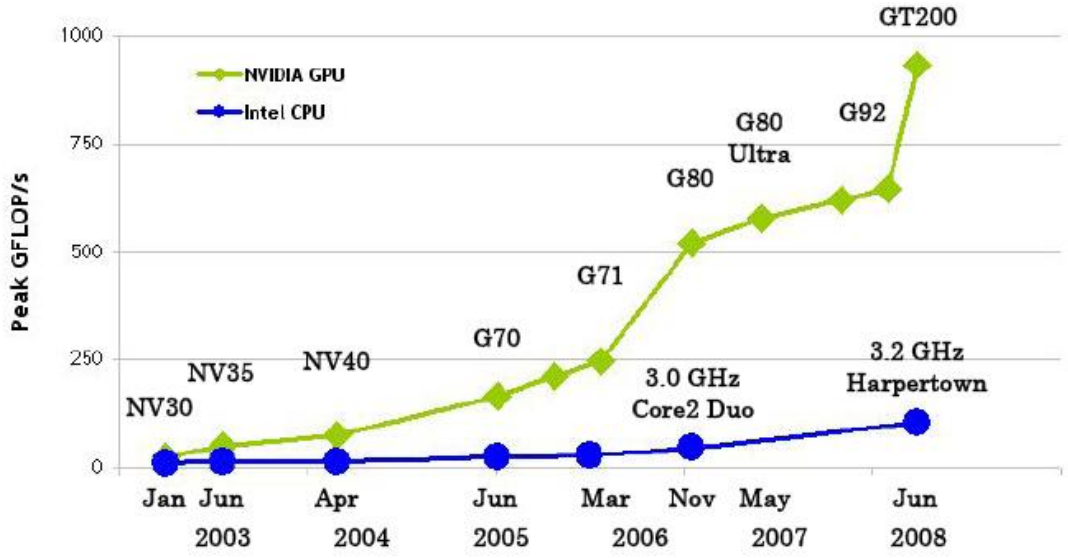
Prosessorin ja grafiikkapiirin alkuperäiset käyttötarkoitukset ovat hyvin erilaiset ja prosessorin englanninkielinen nimi Central Processing Unit antaa osviittaa siitä, että kyseessä on tietokoneen yksi keskeisimmistä komponenteista. Alun perin, vuosikymmeniä sitten, prosessorin rinnalla ei ollut erillisiä näytönohjaimia, mutta käyttöjärjestelmien 2D-ominaisuuksien kasvaessa rinnalle tarvittiin erikoistuneempi ratkaisu kiihdyttämään käyttöliittymää. 3D-sovellusten kehittyessä myös näytönohjaimet ovat kehittyneet ja nykypäivänä ne ovat prosessorin rinnalla välttämätön osa tietokonetta, kun esimerkiksi Microsoftin uusimman Windows 7 -käyttöjärjestelmän Aero-käyttöliittymä jo yksistään asettaa kovatasoiset vaatimukset näytönohjaimelle.

Ylempi kaaviokuva (Kuva 6) selventää karkeasti tilannetta prosessorin ja grafiikkapiirin välillä. Prosessorin ohjausyksikkö, laskentayksiköt ja välimuistit ovat merkittävästi suuremmat eli suorituskykyisemmät kuin grafiikkapiirillä, mutta DRAM-muistien koot ovat yhdenvertaisia.

Siinä missä prosessorissa ytimiä on muutama kappale, grafiikkapiirissä niitä on nykypäivänä satoja (NVIDIA. 2010). Grafiikkapiirin käsittelemät säikeet ovat erittäin kevyitä ja täyden suorituskyvyn saavuttamiseksi niitä tarvitaan tuhansia, kun taas prosessorin tapauksessa käsiteltäviä säikeitä on yhtä aikaa ajossa vain muutamia. Prosessorin strategiana on saavuttaa mahdollisimman hyvä suorituskyky yhtä säiettä

suoritettaessa, kun taas grafiikkapiirillä täysi suorituskyky saavutetaan suurella määrällä säikeitä. (Buck, I. 2007, 15-16)

Arkkitehtuurilliset eroavaisuudet ovat johtaneet siihen, että suorituskykyisimmissä näytönohjaimissa grafiikkapiireillä on käytössä GDDR5-muistia, kun taas prosessoreilla on käytössä DDR3-muistia. Muistien kaistanleveyksissä erot ovat moninkertaiset, sillä GDDR5-muisteilla kyetään saavuttamaan näytönohjaimilla maksimissaan hetkellisesti yli 200 gigatavun tiedonsiirtonopeuksia sekunnissa, kun DDR3:lla nopeudet jäävät alle 20 Gt/s:iin. Toisaalta prosessoreiden käytössä olevat muistit ovat tehokkaampia viiveissä. (Microsoft. 2010)



Kuvio 1. Näytönohjainten ja prosessoreiden kehitys (kuva: NVIDIA)

Intelin kuusiytiminen Core i7-980X Extreme Edition -prosessori kykenee Hyper-Threading-teknologian myötä käsittelemään yhtäaikaisesti maksimissaan 12 säiettä ja sen suorituskyvyn yleisenä mittarina käytettävä GFLOPS-arvo on 108. (Kurri S. 13.3.2010, 2)

NVIDIAn 480 CUDA-ytimellä varustetulla, Fermi-arkkitehtuuriin perustuvalla GeForce GTX 480 -näytönohjaimella arvo on 1,35 TFLOPSia eli 1382 GFLOPSia. Kun edellä mainitun näytönohjaimen sekä edellisessä kappaleessa mainitun prosessorin GFLOPS-lukuja verrataan keskenään, näytönohjaimelle ilmoitetun arvon voidaan todeta olevan yli 10-kertainen. (Wasson, S. 31.3.2010, 1)

Prosessoreilla suorituskyvyn kasvattaminen FLOPS-arvon valossa on huomattavasti hankalampaa kuin näytönohjainten grafiikkapiireillä niiden monimutkaisuuden takia. Siinä missä markkinoiden suorituskykyisimmissä työpöytäkäyttöön suunnatuissa prosessoreissa on fyysisiä ytimiä maksimissaan kuusi kappaletta, GeForce GTX 480:n tapauksessa CUDA-ytimiä on 480 kappaletta. (NVIDIA. 2010)

FLOPS on suorituskyvyn vertailussa käytettävä suure, jossa verrataan liukulukulaskennan suorituskykyä. Varsinkin tieteellisessä laskennassa läsnäolollaan loistava suure ilmaisee lukuarvon, kuinka monta liukulukuoperaatiota laite kykenee suorittamaan sekunnissa. G-etuliite tarkoittaa kerrannaisyksikköä giga (10^9) ja T puolestaan teraa (10^{12}).

4 OHJELMOINTIKIRJASTOT

4.1 NVIDIA CUDA

CUDA on NVIDIAn kehittämä arkkitehtuuri rinnakkaislaskentaan, joka pyrkii hyödyntämään grafiikkapiiriä yleiskäyttöisessä laskennassa. CUDAa hyödyntäviä ohjelmia ohjelmoidaan C-kielellä, mutta mukana on NVIDIAn kehittämiä laajennuksia ja tiettyjä rajoituksia.

CUDAn käyttö on yleistynyt paljon pelipuolella lähinnä fysiikkamallinnuksissa, mutta lisäksi se on löytänyt tiensä moniin ei-graafisiin sovelluksiin. Esimerkkejä käyttökohteista löytyy muun muassa laskennallisista biologia- ja kryptografiasovelluksista sekä yhtenä tunnetuimpana esimerkkinä Berkleyn yliopiston ylläpitämästä BOINC-infrastruktuurista. Kyseessä on infrastruktuuri, jonka avulla voidaan ajaa useita hajautettuja laskentaprojekteja samanaikaisesti ja ideana on, että tutkijat voivat käyttää ympäri maailmaa olevien kotitietokoneiden joutilasta laskentatehoa. BOINCin käytössä on nykypäivänä satoja tuhansia tietokoneita, joilla sen laskentateho ylittää viiden petaFLOPSin rajan. (BOINCstats. 2010)

CUDA ei ole tuettuna kaikilla vanhemmilla NVIDIAn näytönohjaimilla, sillä tuki rajoittuu työpöytäpuolelle GeForce 8 -tuoteperheeseen ja sitä uudempiin malleihin. Lisäksi tuki kattaa tiettyjä Tesla-työasema- ja datapalvelinratkaisuja, Quadro-työpöytä- ja mobiilinäytönohjaimia, GeForce-mobiilinäytönohjaimia sekä ION-alustat. (NVIDIA. 2010)

NVIDIA julkaisi CUDAn ensimmäisen SDK-paketin helmikuussa 2007 Windowsille ja Linuxille, ja Mac OS X -tuki tuli mukaan 2.0-versiossa vuotta myöhemmin. CUDAn uusin versio on syyskuussa 2010 julkaistu 3.2, jonka SDK on aiempien versioiden tavoin saatavilla ilmaiseksi.

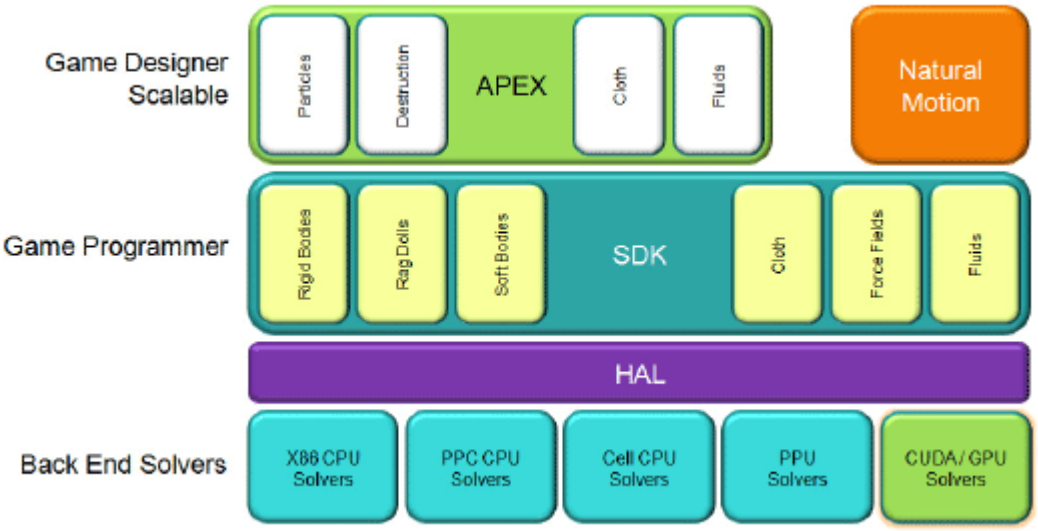
4.1.1 NVIDIA PhysX

PhysX ei ole varsinaisesti ohjelmointikirjasto, vaan peleihin implementoitava reaaliaikainen fysiikkamoottori, joka käyttää fysiikan kiihdyttämisessä apuna näytönohjainta.

NVIDIA julkaisi vuoden 2008 helmikuussa lehdistötiedotteen (NVIDIA. 2008), jossa se ilmoitti ostaneensa AGEIA Technologiesin. Yritys hehkutti tuolloin kykenevänsä yritysoston myötä tuomaan GeForce-tuoteperheen näytönohjaimilla kiihdytettävät PhysX-ominaisuudet satojen miljoonien kuluttajien saataville. NVIDIAn toimitusjohtaja Jen-Hsun Huang paljasti yrityksen liittävänsä PhysX:n osaksi CUDAa, minkä myötä mikä tahansa CUDAa tukeva näytönohjain tukisi PhysX:ää. (Suvanto V. 2008, 1)

PhysX:n nitominen CUDAan ja näytönohjainajureihin tapahtui virallisesti alkuvuodesta 2008. Nykyisellään PhysX on tuettuna GeForce 8- ja sitä uudempien tuoteperheiden kaikissa näytönohjaimissa ja ajurit itsessään ovat saatavilla 32- ja 64-bittisille Windows XP-, Vista- ja 7- sekä Linux- ja Mac OS X -käyttöjärjestelmille. Lisäksi PhysX:ää käytetään Microsoft Xbox 360-, Sony PlayStation 3- ja Nintendo Wii -pelikonsolien peleissä. (Suvanto V. 2008, 2)

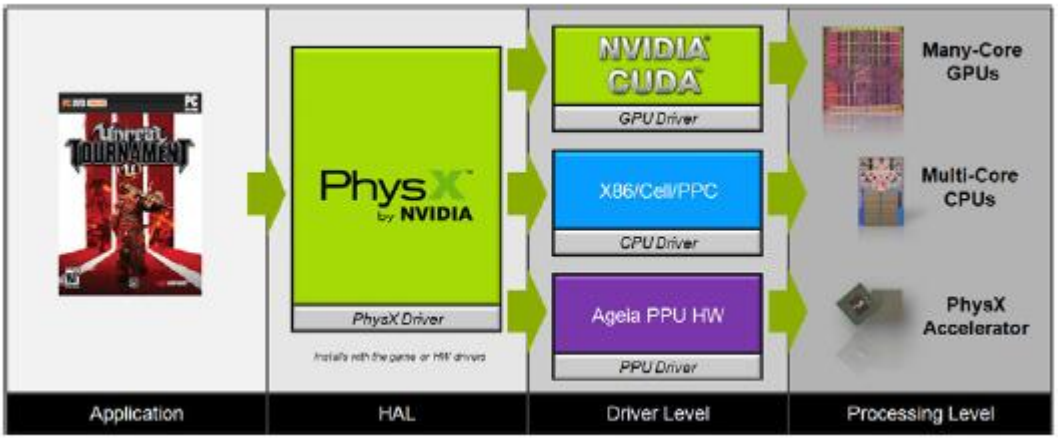
4.1.2 PhysX:n arkkitehtuuri



Kuva 7. Kaaviokuva PhysX-arkkitehtuurista (kuva: NVIDIA)

PhysX:ssä keskeisenä osana on arkkitehtuurikaaviossa keskellä oleva SDK-lohko, jota ohjelmistokehittäjät käyttävät implementoidessaan PhysX:ää. Lisäelementit NVIDIAlta, kuten APEX, tai kolmansilta osapuolilta mahdollistavat kirjastojen käyttämisen PhysX-efektien kanssa ja helpottavat fysiikan luomista ja implementointia peleihin aiempaa helpommin. APEXin ohella Natural Motion on yksi suuri tekijä, jonka kirjasto on suurilta osin vastuussa esimerkiksi Grand Theft Auto IV -pelin efekteistä.

PhysX:ää tukevassa pelissä PhysX:n laskutoimenpidekyselyt menevät HAL-tasoon, joka on esimerkiksi DirectX:n kaltainen rajapinta. HALin myötä fysiikkaa voidaan ajaa erilaisella raudalla, mikä PC:llä tarkoittaa joko x86-arkkitehtuurin mukaista prosessoria, PPU:ta tai NVIDIA PhysX:n myötä GeForce-näytönohjaimen grafiikkapiiriä. NVIDIAn mukaan erillisiä PPU-kortteja voidaan käyttää ainoastaan alhaisen tason fysiikkalaskentaan, kun taas näytönohjaimissa riittää suorituskykyä monimutkaisempaan fysiikkalaskentaan. (Suvanto V. 2008, 2)



Kuva 8. Kaaviokuva PhysX:n toiminnasta ohjelmasta rautatasolle (kuva: NVIDIA)

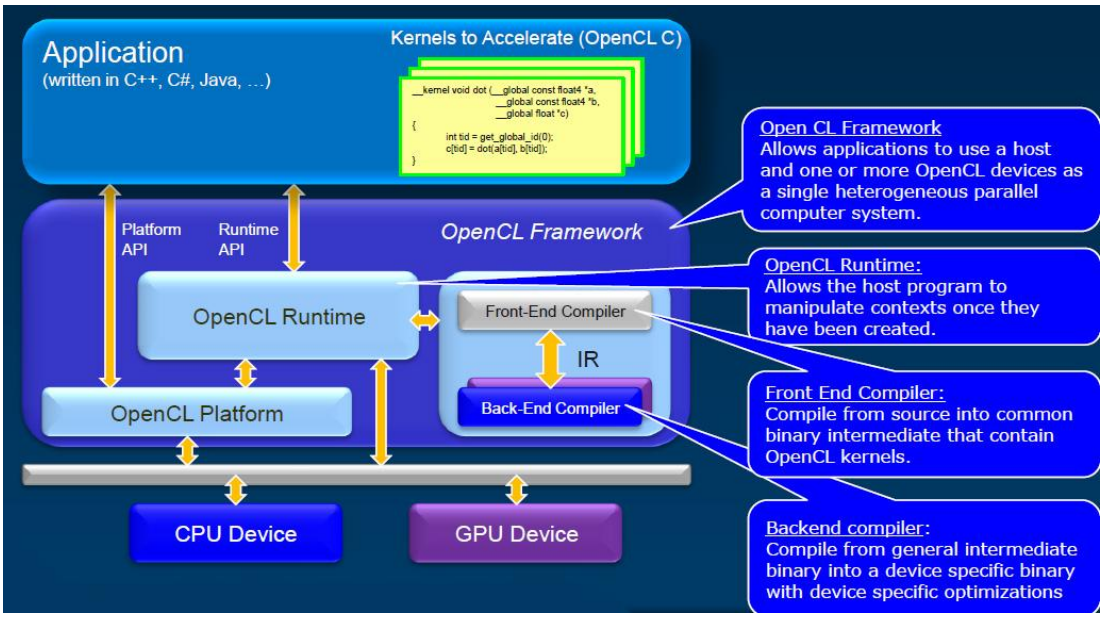
NVIDIAn esimerkkikaaviossa (Kuva 8) on käytetty aikanaan eniten hypeä osakseen saanutta PhysX-peliä, Unreal Tournament 3:a, joka sijoittuu kaaviossa kaikkein vasemmalle. Seuraavana pelin käskyt ohjautuvat HAL-tasoon, joka käytännössä on PhysX-ajuri, ja siitä edelleen ajureiden kautta rautaan, jossa laskutoimitukset suoritetaan. (Suvanto V. 2008, 2)

4.2 Khronos Group OpenCL

OpenCL on alun perin Applen, mutta nykyään Khronos Group -konsortion kehittämä avoin, rojaltivapaa standardi usealle eri alustalle ja rinnakkaisohjelmointiin moderneille prosessoreille, joita löytyy työpöytäkoneista, palvelimista sekä mobiililaitteista. Khronos Groupin taustalla toimivat muun muassa AMD-, Apple-, ARM-, IBM-, Intel-, Nokia-, NVIDIA-, Samsung- ja Texas Instruments -yritykset.

OpenCL:n perimmäinen tarkoitus on hyvin samanlainen kuin NVIDIAn CUDAn eli valjastaa grafiikkapiirit 3D-laskennan ohella yleishyödylliseen laskentaan. Näytönohjainpohjainen yleishyödyllinen laskenta on vasta viime vuosina nostanut päätään, mikä näkyy myös OpenCL:n historiassa. Standardin 1.0-versio julkaistiin vuoden 2008 joulukuussa ja uusin 1.1-versio kesäkuussa 2010. OpenCL:n taustalla ovat suuret erillisnäytönohjainvalmistajat, AMD ja NVIDIA, minkä myötä standardi on tuettuna molempien edellä mainittujen valmistajien näytönohjaimilla.

C-ohjelmointikieleen perustuva OpenCL-standardi määrittelee muun muassa C99-ohjelmointikielen näytteen rinnakkaisuuteen kehitetyille laajennuksilla, rajapinnan koordinoimaan datan ja tehtäväperusteisen rinnakkaislaskennan useilla erilaisilla epäsymmetrisillä prosessoreilla, IEEE 754 -standardiin perustuvia lukuisia vaatimuksia sekä tehokkaan yhteistoiminnan OpenGL:n, OpenGL ES:n ja muiden rajapintojen kanssa. (Khronos Group. 2010)



Kuva 9. Kaaviokuva OpenCL-arkkitehtuurista (kuva: Haifux)

OpenCL:ssä Framework toimii ohjelman ja prosessorin sekä näytönohjaimen välissä, ja antaa ohjelmille mahdollisuuden käyttää isäntää eli prosessoria sekä yhtä tai useampaa OpenCL-laitetta yhtenä heterogeenisenä rinnakkaislaskentajärjestelmänä.

Framworkin sisällä toimiva runtime antaa isäntäohjelmalle mahdollisuuden muokata kontekstejä niiden luomisen jälkeen. Edustakääntäjä kääntää lähdekoodista binääriksi, joka sisältää OpenCL-kernelit, ja taustakääntäjä kääntää binäärit edelleen laitekohtaiseksi binääriksi niille suunnitelluilla optimoinneilla. (Rosenberg, O. 2008, 20)

4.3 Microsoft DirectCompute

DirectCompute on Microsoftin käsialaa ja se kilpailee samalla alalla NVIDIAn CUDAn ja Khronos Groupin OpenCL:n kanssa. Kyseessä on ohjelmointirajapinta, joka on kehitetty grafiikkapiirien GPGPU-käyttöä ajatellen ja tuettuina ovat yrityksen Windows 7- ja Windows Vista -käyttöjärjestelmät. DirectCompute on osa Microsoftin vuosien varrella massiiviseksi paisunutta DirectX-kokoelmaa. DirectCompute on tuettuna kolmen eri version, 10:n, 10.1:n ja 11:n, turvin, joista jälkimmäinen julkaistiin DirectX 11:n kanssa ja on taaksepäin yhteensopiva aiempien versioiden kanssa. (Wikipedia. 2010)

DirectCompute-sovellusten ohjelmoiminen tapahtuu HLSL-kielellä, joka on C:n kaltainen kieli, mutta keskeisimpinä eroina HLSL:ssä ei ole käytössä osoittimia, vaan siinä on sisäänrakennetut muuttujat ja tietotyypit sekä varusfunktiot. (Microsoft. 2010)

Ensimmäiset DirectComputea tukevat ohjelmat ovat keskittyneet lähinnä pelien efektien parantamiseen, kohdistuen jälkikäsittelyyn ja suodattimiin, Deferred Shading -tekniikkaan sekä realistisempiin varjoihin Ambient Occlusion- ja Order Independent Transparency -tekniikoilla. Käytännön esimerkkejä löytyy markkinoilta muun muassa Aliens vs Predator- ja Dirt 2 -peleistä. (Hanley A. 2009, 5)

4.4 AMD ATI Stream

ATI Stream jatkaa samalla linjalla kuin aiemmin käsitellyt CUDA-, OpenCL- ja DirectCompute-teknologiat. Kyseessä on nippu AMD:n rauta- ja ohjelmistoteknologioita, jotka toimivat prosessorin rinnalla ja kiihdyttävät näytönohjaimella muitakin kuin pelkästään 3D-sovelluksia.

Käytännössä ATI Stream rakentuu kahdesta komponentista: kehitys- sekä suoritusympäristöstä. Kehitysympäristö, ATI Stream Software Development Kit, sisältää Brook+-ohjelmointirajapinnan, joka on kehitetty versio Stanfordin yliopiston kehit-

tämästä Brookista. Kyseessä on C:n kaltainen avoimeen lähdekoodiin perustuva ohjelmointikieli, joka on optimoitu Stream-laskentaan. ATI Streamia tukevia sovelluksia voi 2.0-ohjelmistoversioita uudemmilla versioilla kehittää myös käyttämällä OpenCL:ää.

Suoritusympäristö tunnetaan nimellä Compute Abstraction Layer ja se on ollut integroituna AMD:n näytönohjainten Catalyst-ajureihin jo parin vuoden ajan. Ohjelmistokehitys ja ohjelmien suorittaminen onnistuu Windows- ja Linux-käyttöjärjestelmillä. (Wikipedia. 2010)

4.5 Havok Physics

Havokin ratkaisut ovat PhysX:n kaltaisia fysiikkamoottoreita, joista ensimmäinen julkaistiin jo vuonna 2000 ja se teki itsensä tunnetuksi muun muassa Half-Life 2 -pelissä. Havok Physicsin uusin versio on 7.1 ja se on tuettuna tietokonepuolella Windows-, Linux- ja Mac OS X -käyttöjärjestelmissä sekä lisäksi Microsoft Xbox 360-, Xbox, Sony PlayStation 3-, PlayStation 2-, PlayStation Portable-, Nintendo Wii- ja GameCube -pelikonsoleissa. Lisenssin ostajille tarjotaan pääsy suurimpaan osaan ohjelmiston C/C++-lähdekoodista, mikä antaa mahdollisuuden ohjelmistokehittäjille vapauden muokata fysiikkamoottorin ominaisuuksia tai kääntää se eri alustoille. (Havok. 2010)

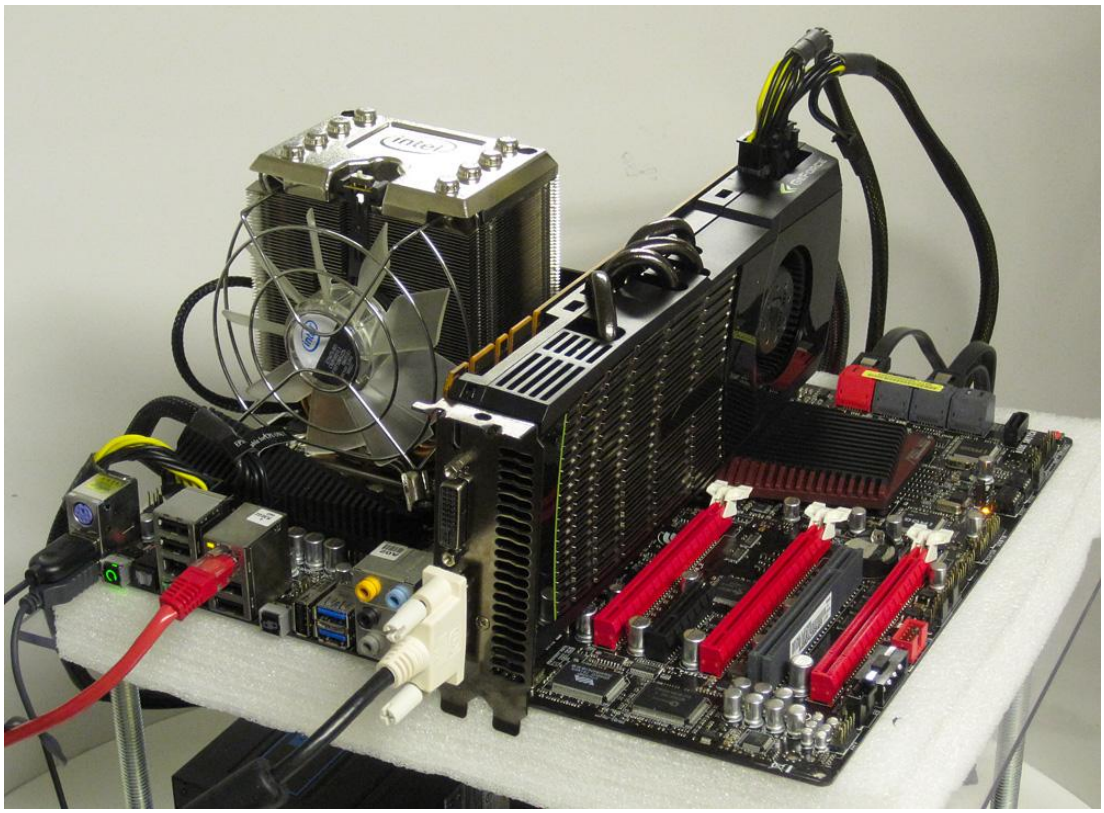
Havok Physicsin keskeisenä ideana on tarjoa pääasiassa peleihin reaaliaikaisia kappaleiden välisiä yhteentörmäyksiä ja jäykkien kappaleiden dynaamisuutta kolmiulotteisena. Käytännössä fysiikkamoottori näkyy peleissä esimerkiksi ragdoll-hahmojen realistisissa ja dynaamisissa liikkeissä sekä ympäröivään peliympäristöön fyysisesti reagoivina kappaleina. (Wikipedia. 2010)

5 TESTIT

Mittauksissa oli tarkoituksena selvittää, miten suorituskyky ja testikokoonpanon tehonkulutus muuttuvat riippuen siitä, käytetäänkö ohjelman suorittamisessa pelkästään prosessoria vai onko apuna myös näytönohjain.

Ensimmäisenä testiohjelmana käytettiin PowerDVD 10 -videotoisto-ohjelmaa, jolla mitattiin prosessorin käyttöastetta. Videoiden muuntamiseen keskittyneellä MediaShow Espresso 6.5 -ohjelmalla mitattiin muuntamiseen kulunut aika ja synteettisellä DirectCompute & OpenGL Benchmark 0.45 -ohjelmalla suorituskykyä DirectComputea ja OpenCL:ää hyödyntämällä sekä pelkästään prosessorilla.

5.1 Testikokoonpano



Kuva 10. Mittauksissa käytetty testikokoonpano

Mittauksissa käytettiin Intel Core i7-980X Extreme Edition -prosessoria, jonka peruskellotaajuus on 3,33 GHz, mutta rasittaessa prosessorin Turbo Mode -teknologia nostaa kellotaajuuden rasitettavien ytimien lukumäärästä riippuen joko 3,46 tai 3,6 GHz:iin. Alustaksi testikokoonpanoon (Kuva 10) valittiin Intel X58 -piirisarjaan perustuva Asus Rampage III Extreme -emolevy, johon asennettiin NVIDIA GeForce GTX 480 -näytönohjain, kaksi OCZ:n kahden gigatavun DDR3-1066-muistikampaa, Corsairin 1000 watin HX1000W -virtalähde sekä Western Digitalin teratavun tallennuskapasiteetillinen Caviar Black WD1001FALS -kiintolevy. Käyttöjärjestelmänä käytettiin 64-bittistä Microsoft Windows 7 Home Premiumia ja näytönohjainta varten asennettiin NVIDIA GeForce 260.99 -ajurit.

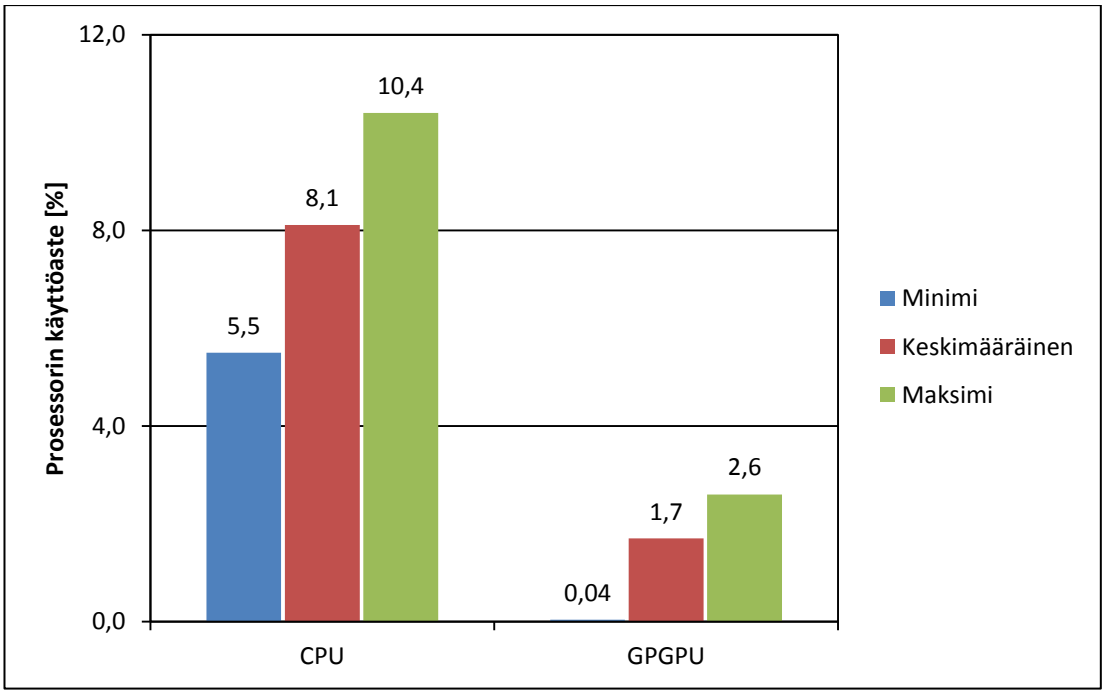
Testikokoonpanon tehonkulutusta seurattiin Etech PM300 -energiamittarilla, jolla mitattiin koko tietokoneen tehonkulutus pois lukien monitori.

Taulukko 3. Testikokoonpanossa käytetyt komponentit, ohjelmat ja mittalaite

Prosessori	Intel Core i7-980X Extreme Edition
Emolevy	Asus Rampage III Extreme
DDR3-muistit	2 kpl OCZ OCZ3B2133LV6GK
Näytönohjain	NVIDIA GeForce GTX 480
Virtalähde	Corsair HX1000W
Kiintolevy	Western Digital Caviar Black WD1001FALS
Käyttöjärjestelmä	Microsoft Windows 7 Home Premium (64-bittinen versio)
Testiohjelmat	CyberLink PowerDVD 10 CyberLink MediaShow Espresso 6.5 DirectCompute & OpenCL Benchmark 0.45
Energiamittari	Etech PM300

5.2 PowerDVD 10

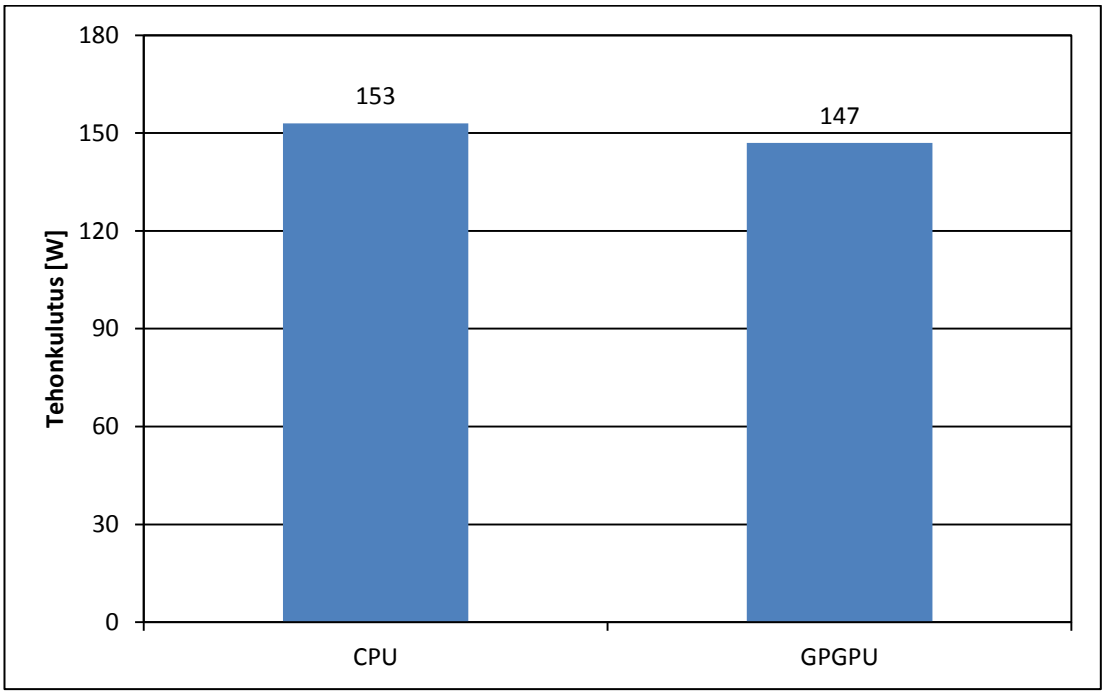
PowerDVD 10 on CyberLink-nimisen yrityksen kehittämä videotoisto-ohjelma, joka on saatavilla sekä Windows- että Linux-käyttöjärjestelmille. Ohjelmasta on saatavilla useita eri versioita ja se tukee laaja-alaisesti eri formaatteja, joista näytönohjaimen hyödyntämistä ajatellen tärkeimmät ovat H.264 (MPEG-4 AVC), VC-1, AVCHD sekä WMV HD, joita kiihdytetään NVIDIAn näytönohjaimilla CUDAn avulla ja AMD:n näytönohjaimilla ATI Streamilla. PowerDVD oli markkinoiden ensimmäinen Blu-ray 3D:tä tukeva ohjelma, joka sai Blu-ray Disc Associationilta sertifioinnin Blu-ray 3D Profile 5.0:lle. (CyberLink. 2010)



Kuvio 2. Prosessorin käyttöaste PowerDVD 10 -ohjelmalla

PowerDVD 10 -ohjelmalla toistettiin I Am Legend -elokuvan MP4-formaatissa olevaa 1920x816-resoluutioista ennakkomainoselokuvaa ja Windows 7 -käyttöjärjestelmän Performance Monitor -työkalulla mitattiin prosessorin minimi-, keskimääräinen ja maksimikäyttöaste. Kuviossa 2 CPU-tulos osoittaa, kun toistossa ei käytetty apuna näytönohjainta ja GPGPU puolestaan näytönohjainavusteisen tuloksen. Kuviossa pienempi tulos on parempi.

Videota toistettaessa pelkästään prosessorilla, prosessorin käyttöaste oli minimissään 5,5; keskimäärin 8,1 ja maksimissaan 10,4 prosenttia (Kuvio 2). Kun toistoa avustettiin näytönohjaimella, vastaavat lukemat olivat 0,04; 1,7 ja 2,6 prosenttia. Mittaukset osoittivat selvästi, miten näytönohjain avusti videotoistoa poistamalla työtä prosessorilta ja lopputuloksena prosessorin käyttöaste laski näytönohjainta käyttäessä huomattavasti alemmaksi.



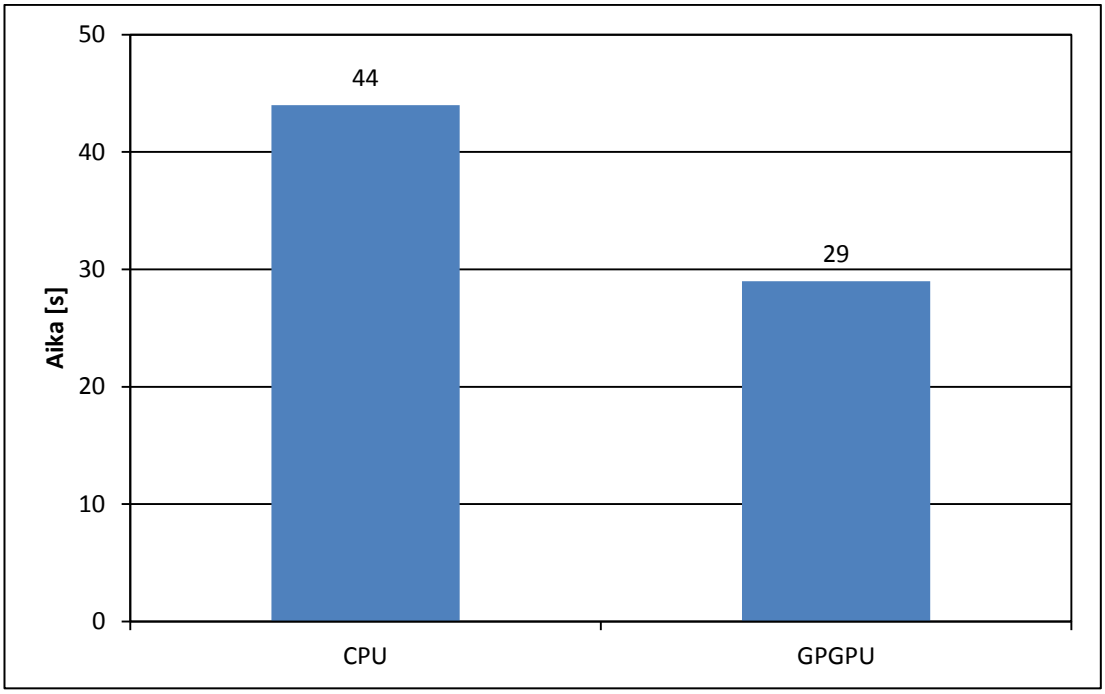
Kuvio 3. Testikokoonpanon tehonkulutus PowerDVD 10 -ohjelmalla

Tehonkulutusmittaukset (Kuvio 3) osoittivat kulutuksen olevan vähäisempää, kun videota toistettiin näytönohjaimella. Pelkästään prosessoria käytettäessä tehonkulutus oli 153 wattia ja näytönohjainavusteisesti 147 wattia. Näytönohjain oli asennettuna testikokoonpanoon molemmissa tapauksissa, joten näytönohjainta käyttäessä tehonkulutus oli kuusi wattia alhaisempi.

5.3 MediaShow Espresso 6.5

MediaShow Espresso 6.5 on PowerDVD 10:n tavoin CyberLinkin käsialaa ja kyseessä on videoiden muuntamiseen keskittynyt ohjelma, jonka toimintaa voidaan kiihdyttää NVIDIA CUDA- ja AMD ATI Stream -teknologiaa tukevilla näytönohjaimilla.

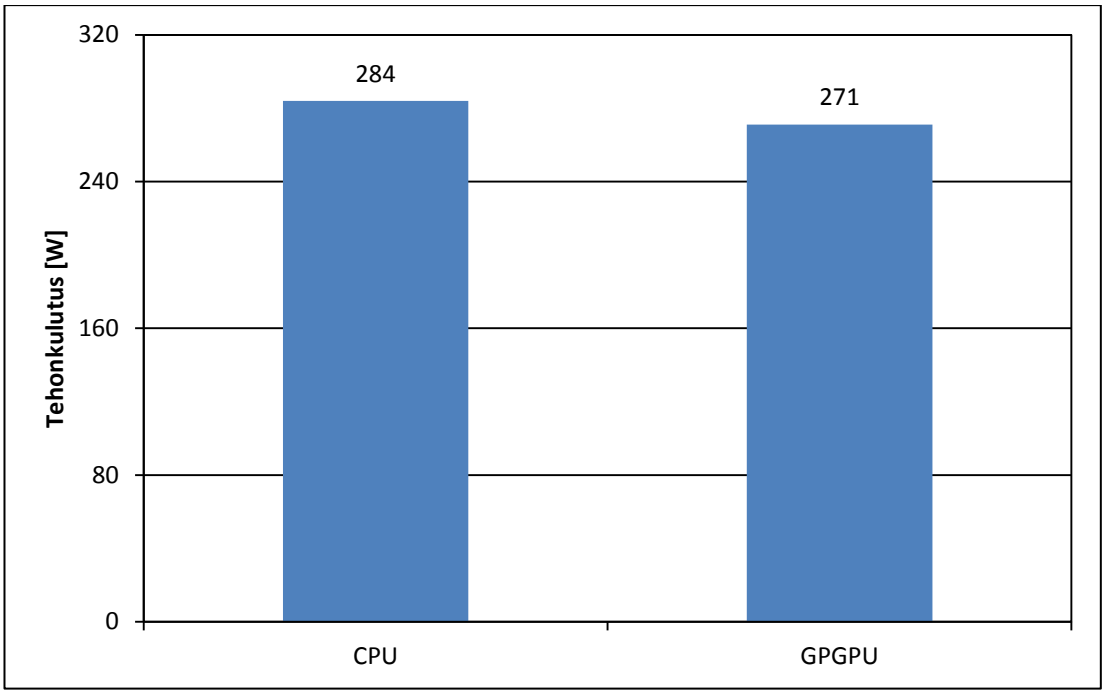
Helppokäyttöisyyteen keskittyneellä ohjelmalla voi muuntaa videoita formaatteihin, joiden toistaminen onnistuu muun muassa Apple iPhonella ja iPadilla sekä muiden valmistajien älypuhelimilla, Sony PSP -käsikonsolilla ja Microsoft Xbox 360 -pelikonsolilla. Tuettuna on myös YouTube-rajapinta, jonka myötä videot on mahdollista siirtää suoraan Googlen suosittuun videopalveluun. Videoita on mahdollista muuntaa H.264- (MPEG-4 AVC), MPEG-2-, MPEG-4-, WMV- ja DivX-formaatteihin (MPEG-4 Part 2). (CyberLink. 2010)



Kuvio 4. Muuntamiseen käytetty aika MediaShow Espresso 6.5 -ohjelmalla

MediaShow Espresso 6.5:llä muunnettiin I Am Legend -elokuvan 121 megatavun kokoinen MP4-formaatissa oleva 1920x816-resoluutioinen ennakkomainoselokuva PlayStation 3 -esiasetuksella M2TS-formaattiin. Kuviossa 4 CPU-tulos osoittaa, kun muuntamisessa ei käytetty apuna näytönohjainta ja GPGPU puolestaan näytönohjainavusteisen tuloksen. Kuvion luvut osoittavat muuntamiseen kuluneen ajan ja pienempi tulos on parempi.

Kun videomuunnoksessa käytettiin pelkästään prosessoria, operaatioon kului 44 sekuntia (Kuvio 4), ja näytönohjaimella avustettaessa muunnokseen kuluva aika tippui 29 sekuntiin.

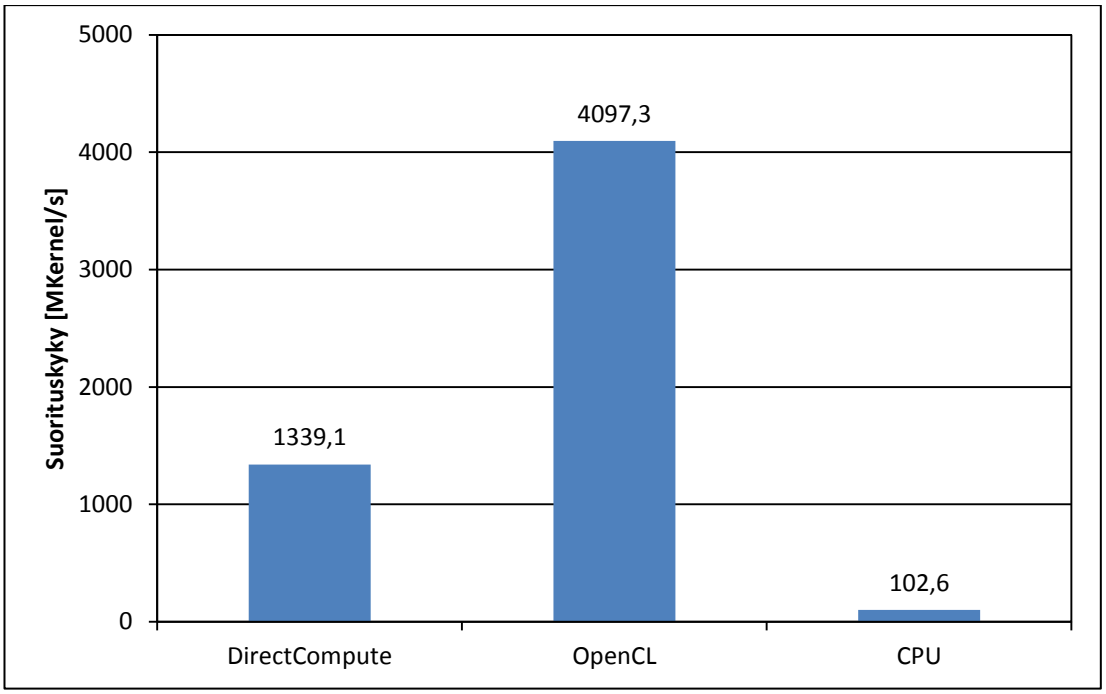


Kuvio 5. Testikokoonpanon tehonkulutus MediaShow Espresso 6.5 -ohjelmalla

MediaShow Espresso 6.5 -ohjelman tehonkulutusmittaukset (Kuvio 5) osoittavat kulutuksen laskeneen, kun muuntamisessa käytettiin apuna näytönohjainta. Pelkästään prosessoria käyttäessä tehonkulutus oli 284 wattia, mutta näytönohjaimella kulutus laski 271 wattiin.

5.4 DirectCompute & OpenCL Benchmark 0.45

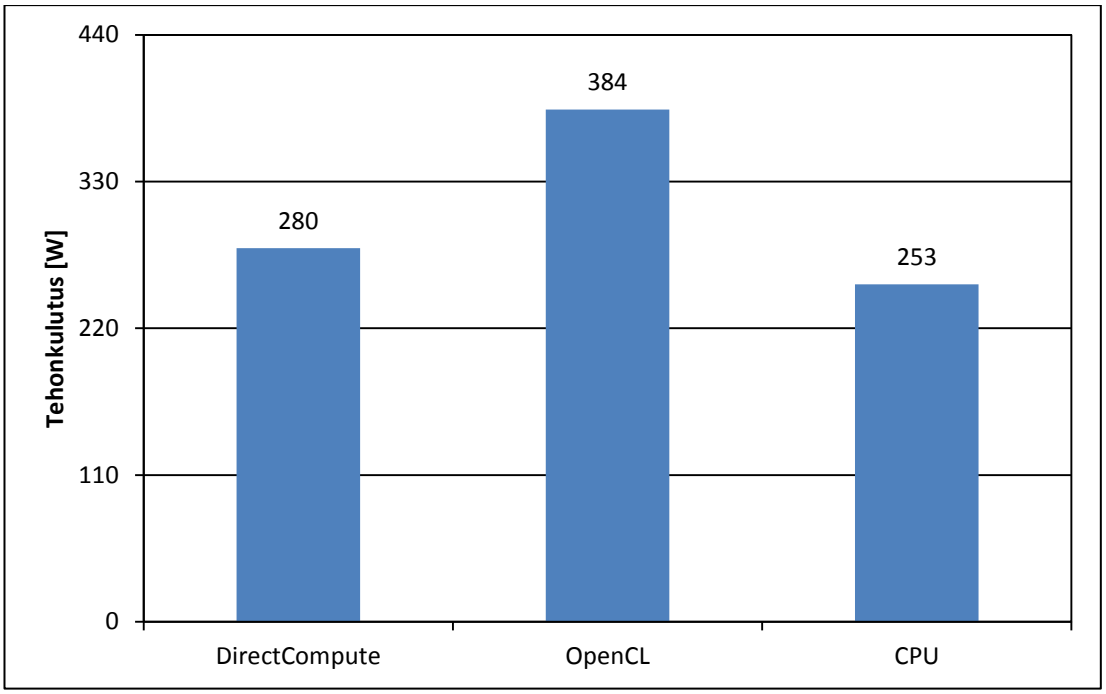
DirectCompute & OpenCL Benchmark on Patryk Dabrowskin kehittämä synteettinen testiohjelma, jolla voidaan mitata ja verrata suorituskykyä käyttämällä pelkästään prosessoria tai näytönohjaimella OpenCL- tai DirectCompute-rajapintaa. Ohjelma suorittaa yksinkertaisia FFT-muunnoksen kaltaisia operaatioita ja ilmoittaa tuloksen suoritettuina megakerneleinä sekunnissa.



Kuvio 6. Suorituskyky DirectCompute & OpenCL Benchmark 0.45 -ohjelmalla

DirectCompute & OpenCL Benchmark osoittaa hyvin, miten näytönohjaimella voidaan saavuttaa moninkertainen suorituskykyparannus, kun operaatio on optimoitu käyttökohteen mukaan. Kuviossa 6 suurempi tulos on parempi.

Pelkkää prosessoria käyttäessä ohjelmalla saavutettiin 102,6 Mkernels/s:n tulos. Kun ohjelman suorittamisessa käytettiin näytönohjainta ja DirectCompute-rajapintaa, tulos nousi 1339,1 Mkernels/s:iin ja OpenCL:llä 4097,3 Mkernels/s:iin.



Kuvio 7. Testikokoonpanon tehonkulutus DirectCompute & OpenCL Benchmark 0.45 -ohjelmalla

Tehonkulutusmittaukset (Kuvio 7) osoittavat merkittäviä eroja tehonkulutuksissa ja vähiten tehoa kului, kun ohjelman suorittamisessa käytettiin pelkästään prosessoria tehonkulutuksen ollessa tällöin 253 wattia. Näytönohjaimella ja DirectCompute-rajapinnalla tehonkulutus nousi 280 wattiin ja OpenCL-rajapinnalla 384 wattiin.

LÄHTEET

Verkkodokumentit

NVIDIA. 2008. NVIDIA to Acquire AGEIA Technologies [verkkodokumentti].
[Viitattu 29.6.2009].
Saatavissa: http://www.nvidia.com/object/io_1202161567170.html

Kurri, S. 2010. Intel Core i7-980X Extreme Edition (Gulftown). Muropaketti [verkkodokumentti].
[Viitattu 7.6.2010].
Saatavissa: http://plaza.fi/muropaketti/artikkelit/prosessorit/intel-core-i7-980x-extreme-edition-gulftown

Wasson, S. 2010. Nvidia's GeForce GTX 480 and 470 graphics processors [verkkodokumentti].
[Viitattu 7.6.2010].
Saatavissa: http://techreport.com/articles.x/18682

NVIDIA. 2010. GeForce GTX 480 [verkkodokumentti].
[Viitattu 7.6.2010].
Saatavissa: http://www.nvidia.com/object/product_geforce_gtx_480_us.html

Answers.com. Timeline of computing 1980-1989 [verkkodokumentti].
[Viitattu 8.6.2010].
Saatavissa: http://www.answers.com/topic/timeline-of-computing-1980-1989

Wikipedia. 2010. Video card [verkkodokumentti].
[Viitattu 8.6.2010].
Saatavissa: http://en.wikipedia.org/wiki/Video_card

Wikipedia. 2010. Graphics processing Unit [verkkodokumentti].
[Viitattu: 8.6.2010].
Saatavissa: http://en.wikipedia.org/wiki/Graphics_processing_unit

NVIDIA. 2009. Fermi Compute Architecture White Paper [verkkodokumentti].
[Viitattu 8.6.2010].
Saatavissa:
http://www.nvidia.com/content/PDF/fermi_white_papers/NVIDIA_Fermi_Compute_Architecture_Whitepaper.pdf

Kurri, S. 2010. NVIDIAn Fermi-GeForcen GF100-grafiikkapiiri. Muropaketti [verkkodokumentti].
[Viitattu 9.6.2010].
Saatavissa: http://plaza.fi/muropaketti/artikkelit/tekniikkakatsaukset/nvidian-fermi-geforcen-gf100-grafiikkapiiri

Wasson, S. 2009. AMD's Radeon HD 5870 graphics processor. The Tech Report [verkkodokumentti].
[Viitattu 14.6.2010].
Saatavissa: http://www.techreport.com/articles.x/17618/

Buck, I. 2007. GPU Computing, Programming a Massively Parallel Processor. NVIDIA [verkkodokumentti].
[Viitattu 14.6.2010].
Saatavissa: http://www.cgo.org/cgo2007/presentations/cgo2007-keynote-buck.pdf

NVIDIA. 2010. What is GPU Computing? [verkkodokumentti].
[Viitattu 14.6.2010].
Saatavissa: http://www.nvidia.com/object/GPU_Computing.html

CyberLink. 2010. PowerDVD 10 [verkkodokumentti].
[Viitattu 17.6.2010].
Saatavissa: http://www.cyberlink.com/products/powerdvd/compare-retail_en_US.html

CyberLink. 2010. MediaShow Espresso 5.5 [verkkodokumentti].
[Viitattu 17.6.2010].
Saatavissa: http://www.cyberlink.com/products/mediashow-espresso/compare_en_US.html

Wikipedia. 2010. Lämpöputki [verkkodokumentti].
[Viitattu: 4.8.2010].
Saatavissa: http://fi.wikipedia.org/wiki/L%C3%A4mp%C3%B6putki

Intel. 2010. Some PCI Express Graphics cards require extra power [verkkodokumentti].
[Viitattu: 4.8.2010].
Saatavissa: http://www.intel.com/support/motherboards/desktop/sb/cs-012073.htm

NVIDIA. 2010. CUDA GPUs [verkkodokumentti].
[Viitattu 5.8.2010].
Saatavissa: http://www.nvidia.com/object/cuda_gpus.html

BOINCstats. BOINC combined [verkkodokumentti].
[Viitattu 5.8.2010].
Saatavissa: http://boincstats.com/stats/project_graph.php?pr=bo

Khronos Group. 2010. OpenCL Overview [verkkodokumentti].
[Viitattu: 9.8.2010].
Saatavissa: http://www.khronos.org/opencl/

Haifux. 2008. OpenCL Overview [verkkodokumentti].
[Viitattu: 9.8.2010].
Saatavissa: http://www.haifux.org/lectures/212/OpenCL_for_Halifux_new.pdf

Microsoft. 2010. DirectX 11 DirectCompute: A Teraflop for Everyone [verkkodokumentti].
[Viitattu: 16.8.2010].
Saatavissa: http://download.microsoft.com/download/4/9/7/49714DA3-1BE8-4750-8566-
D818DD18E4C8/DirectX_11_DirectCompute_A_Teraflop_for_Everyone_US.zip

Wikipedia. 2010. DirectCompute [verkkodokumentti].
[Viitattu: 16.8.2010].
Saatavissa: http://en.wikipedia.org/wiki/DirectCompute

Microsoft. 2010. DirectCompute Lecture Series 101: Introduction to DirectCompute [verkkodokumentti].
[Viitattu: 17.8.2010].
Saatavissa:
http://code.msdn.microsoft.com/Project/Download/FileDownload.aspx?ProjectName=DirectComputeLecture&DownloadId=12810

Hanley, A. 2009. AMD Radeon HD 5800 series technology preview - ATI Stream, DirectCompute, physics, DirectX 11. Elite Bastards [verkkodokumentti].
[Viitattu 19.8.2010].
Saatavissa:
http://elitebastards.com/index.php?option=com_content&view=article&id=820%3Aamd-radeon-hd-5800-series-technology-preview&catid=17%3Apreviews&Itemid=31&limitstart=4

AMD. 2010. ATI Stream Technology [verkkodokumentti].
[Viitattu: 19.8.2010].
Saatavissa: http://www.amd.com/US/PRODUCTS/TECHNOLOGIES/STREAM-TECHNOLOGY/Pages/stream-technology.aspx

AMD. 2010. ATI Stream Software Development Kit (SDK) v2.2 [verkkodokumentti].
[Viitattu: 19.8.2010].
Saatavissa: http://developer.amd.com/gpu/atistreamsdk/pages/default.aspx

Havok. 2010. Havok Physics [verkkodokumentti].
[Viitattu: 20.8.2010].
Saatavissa: http://www.havok.com/uploads/Havok_Physics_Brief_Mar%209.pdf

Wikipedia. 2010. Havok (software) [verkkodokumentti].
[Viitattu: 20.8.2010].
Saatavissa: http://en.wikipedia.org/wiki/Havok_%28software%29

Suvanto, V. 2008. NVIDIA PhysX -tekniikka. Muropaketti [verkkodokumentti].
[Viitattu 4.10.2010].
Saatavissa: http://plaza.fi/muropaketti/artikkelit/tekniikkakatsaukset/nvidia-physx-tekniikka

Wikipedia. 2010. AMD FireStream [verkkodokumentti].
[Viitattu: 4.10.2010].
Saatavissa: http://en.wikipedia.org/wiki/AMD_FireStream

Jon Peddie Resarch. 2011. Jon Peddie Research reports disappointing 4th quarter: PC Graphics shipments down 7.8% year over year [verkkodokumentti].
[Viitattu 7.3.2011].
Saatavissa: http://jonpeddie.com/press-releases/details/jon-peddie-research-reports-disappointing-4th-quarter/

Jon Peddie Resarch. 2010. Jon Peddie Research announces first quarter shipments of PC graphics increase 44% year over year [verkkodokumentti].
[Viitattu 7.3.2011].
Saatavissa: http://jonpeddie.com/press-releases/details/jon-peddie-research-announces-first-quarter-shipments-of-pc-graphics-increa/

DIRECTCOMPUTE & OPENCL BENCHMARK 0.45

DirectCompute	OpenCL	CPU
Pisteet (MKernels/s)	**Pisteet (MKernels/s)**	**Pisteet (MKernels/s)**
1339,1	4097,1	107,2
1341,4	4124,1	102,6
1339,1	4097,3	102,5
Tehonkulutus (W)	**Tehonkulutus (W)**	**Tehonkulutus (W)**
280	384	253

CYBERLINK MEDIAESPRESSO 6

CPU	GPU + CPU
Tulos (sekuntia)	**Tulos (sekuntia)**
44	29
44	30
44	29
Tehonkulutus (W)	**Tehonkulutus (W)**
284	271

CYBERLINK POWERDVD 10

CPU

Tulos (%)

Minimi	Keskimääräinen	Maksimi
5,885	8,240	10,175
5,105	8,101	10,435
5,495	8,110	10,565

GPU + CPU

Tulos (%)

Minimi	Keskimääräinen	Maksimi
0,425	1,923	2,635
0,035	1,611	2,800
0,035	1,727	2,505

CPU	GPU + CPU
Tehonkulutus (W)	
153	147