Tampereen ammattikorkeakoulu

Näytönohjainarkkitehtuurit

Jouni-Junior Salo

OPINNÄYTETYÖ
Helmikuu 2019

Tieto- ja viestintätekniikan koulutus
Sulautetut järjestelmät

TIIVISTELMÄ

Tampereen ammattikorkeakoulu
Tieto- ja viestintätekniikan koulutus
Sulautetut järjestelmät

SALO JOUNI-JUNIOR
Näytönohjainarkkitehtuurit

Opinnäytetyö 39 sivua
Maaliskuu 2019

Tässä opinnäytetyössä on perehdytty Yhdysvaltalaisen grafiikkasuorittimien valmistajan Nvidian historiaan ja tuotteisiin. Nvidia on toinen maailman suurimmasta grafiikkasuorittimien valmistajasta. Tässä työssä tutustutaan tarkemmin Nvidian arkkitehtuureihin, Fermiin, Kepleriin, Maxwelliin, Pascaliin, Voltaan ja Turingiin.

Opinnäytetyössä tutkittiin, mistä asioista Nvidian arkkitehtuurit koostuvat ja miten eri komponentit kommunikoivat keskenään. Työssä käytiin läpi jokaisen arkkitehtuurin julkaisuvuosi ja niiden käyttökohteet. Työssä huomattiin kuinka paljon Nvidian teknologia on kehittynyt vuosien varrella ja kuinka Nvidian koneoppimiseen tarkoitettuja työkaluja on käytetty.

Nvidia Fermi Kepler Maxwell Pascal Volta Turing rtx näytönohjain gpu

ABSTRACT

Tampere University of Applied Sciences
Information and communication technologies
Embedded systems

SALO JOUNI-JUNIOR
GPU architectures

Bachelor's thesis 39 pages
March 2019

This thesis focuses on the history and products of an American technology company Nvidia Corporation. Nvidia Corporation is one of the two largest graphics processing unit designers and producers. This thesis examines all of the following Nvidia architectures, Fermi, Kepler, Maxwell, Pascal, Volta and Turing.

The goal of the thesis was to find out what parts Nvidia’s architectures consists of and how different components communicate with each other. All of the architectures release years and the usages were studied in this thesis. On the study of the architectures it was discovered how much Nvidia’s technology has improved and how all the tools are used in deep learning and artificial intelligence.

Nvidia Fermi Kepler Maxwell Pascal Volta Turing rtx gpu

SISÄLLYS

ERITYISSANASTO

CPU	Central Processing Unit, suoritin, prosessori
CUDA	Nvidian kehittämä ohjelmointirajapinta
DirectX	Peleille tarkoitettu ohjelmointirajapinta 3D-grafiikkaa varten
DRAM	Dynamic Random Access Memory, dynaaminen RAM-muisti
FMA	Floating Multiply Add, Liukuva yhteenlaskuoperaatio
FPS	Frames Per Second, kehysnopeus
GPC	Graphics Processing Cluster, osa GPU:n rakennetta
GPU	Graphics Processing Unit, grafiikkasuoritin
QPI	QuickPath Interconnect, Intelin väyläteknologia
RAM	Random Access Memory, käyttömuisti, keskusmuisti
ROP	Raster Operations pipeline, tekstuuriyksikkö
SM	Streaming Multiprocessor, Fermi grafiikkapiirin osa
SMM	Streaming Multiprocessor Maxwell, Maxwell grafiikkapiirin osa
SMX	Next generation Streaming Multiprocessor, Kepler grafiikkapiirin osa

1 JOHDANTO

Tässä opinnäytetyössä tutustutaan maailman suurimpaan grafiikkasuorittimien valmistajaan Nvidiaan. Ensimmäiseksi käydään lävitse näytönohjaimen käyttötarkoituksia. Tämän jälkeen siirrytään Nvidia Corporationin historiaan.

Luvussa 3.2 käsitellään Nvidian grafiikkasuorittimien arkkitehtuureja. Työssä tutustutaan seuraaviin Nvidian arkkitehtuureihin: Fermi, Kepler, Maxwell, Pascal, Volta ja Turing. Näistä arkkitehtuureista käydään lävitse niiden rakenne ja merkittävimmät uudistukset.

Viimeisessä kappaleessa käydään läpi Nvidian CUDA-arkkitehtuurin toiminta. Kappaleessa kerrotaan, millä ohjelmointikielillä CUDA-arkkitehtuuria voidaan ohjelmoida ja soveltaa.

Työn aihe on pitkäaikainen harrastukseni ja kiinnostuksen kohteeni ja myöskin se liittyy vahvasti omaan työhöni, joka on tällä hetkellä mousesports nimisen saksalaisen organisaation edustaminen Rainbow Six Siege -nimisessä pelissä ammattilaistasolla. Tällä alalla pitää pysyä ajan tasalla juurikin komponenteissa, koska nykypäivänä melkein kaikki pelit ovat todella vaativia juurikin grafiikkasuorittimille.

2 NÄYTÖNOHJAIN

Näytönohjain on tärkeä komponentti tietokoneessa, koska näytönohjain huolehtii grafiikan tulostamisesta näytölle. Näytönohjain on yleisesti erillinen komponentti pöytäkoneissa, mutta suurimmassa osassa prosessoreita eli CPU:ita on myös integroitu näytönohjain. Näytönohjaimia valmistavat yritykset lisensoivat grafiikkasuoritintekniikan edellä mainituilta suoritinvalmistajilta, eli Nvidia:lta ja AMD:ltä. Kuluttajille suunnatut kymmenen suosituinta näytönohjainvalmistajaa ovat:

- Asus
- MSI
- Gigabyte
- EVGA
- Zotac
- Galax
- PNY
- Palit
- PowerColor
- Sapphire

Näistä kuitenkin tunnetuimmat ja suosituimmat valmistajat ovat Asus, MSI, ja EVGA näiden luotettavan toimivuuden ja luotettavan takuupalvelun takia. (Top Graphics Card Manufacturers & Brands for Nvidia & AMD GPUs 2018.)

2.1 Grafiikan piirto

Nykypäivän grafiikkasuorittimet sisältävät tuhansia CUDA-prosessoreita eli varjostinprosessoreita. Nämä prosessorit tuottavat grafiikan näytölle. Grafiikkasuoritin käskee varjostinprosessoreita tekemään eri työt. Esimerkiksi jos pelissä on vettä, prosessori kertoo varjostinprosessoreille, että nyt piirretään vettä ja jos pelissä on metsää, varjostinprosessorit tietää piirtävänsä metsää.

Ennen CUDA-prosessoreita näytönohjaimet tuottivat monikulmioita ja kortin nopeus laskettiin, kuinka monta polygonia pystyy kortti piirtämään sekunnissa. Nykypäivän kortit käyttävät hyväkseen DirectX 12 -ohjelmointirajapintaa.

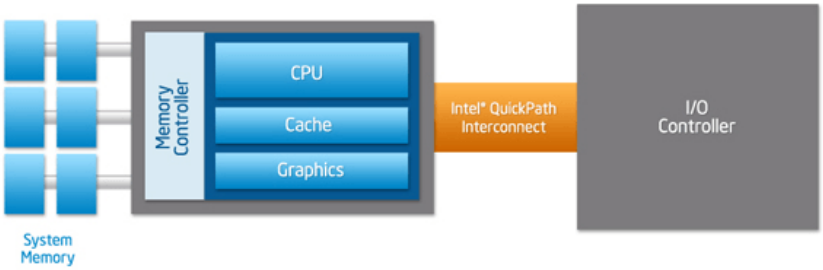
> Uusimman DirectX 12 -rajapinnan avulla kehittäjät voivat nyt käyttää tehokasta näytönohjaingrafiikkaa innovatiivisimmissa peleissään. Uusimmat GeForcen näytönohjaimet tukevat kyseisen rajapinnan ominaisuuksia ja nämä elävöittävät pelaamista uusilla visuaalisilla tehosteilla ja piirtotekniikoilla. (Nvidia Corporation 2019.)

2.2 Erilaiset näytönohjaintyypit

2.2.1 Integroitu näytönohjain

Integroidut näytönohjaimet ovat riittäviä tavalliseen internetin selailuun, HD elokuvien katseluun ja mahdollisesti myös vanhempien ja vähemmän vaativien pelien pelaamiseen. Ennen vuotta 2010 prosessoreissa ei vielä ollut integroitua näytönohjainta, vaan näytönohjain oli integroitu tietokoneen emolevyyn. Vuonna 2010 Intel julkaisi uuden prosessorin käyttäen Clarkdale-arkkitehtuuria. Tämä arkkitehtuuri käyttää Intelin 32nm teknologiaa. Piirilevyn DDR3-muisti ja grafiikkasuoritin on rakennettu käyttäen 45nm transistoreita. Nämä kommunikoivat keskenänsä käyttäen high-speed QPI linkkiä. (bit-tech.net 2009.)

QPI eli Intelin kehittämä Quick Path interconnect on spesifikaatio kaikkien johdonmukaisten rajapintojen perheelle, joka sisältää kaikki kannettavat tietokoneet, työpöytäkoneet ja palvelimet, jotka tarvitsevat erilaisia ominaisuuksia. Tämä QPI siis korvaa sitä vanhemman suoritinväyläteknologian. Väylä on julkaistu vuonna 2007 ja Intel Xeon-malleissa se on otettu käyttöön vuonna 2009. (Intel’s Quick Path Evolved 2011.)



KUVA 1. Intelin Quick Path teknologian toimintaperiaate (intel 2011)

Intelin QuickPath teknologia käyttää Intelin QuickPath intercorrectia, joka sallii high-speed, point-to-point linkit suorittimen sisä -ja ulkopuolella. Kontrastissa rinnakkaisiin portteihin nämä linkit nopeuttavat datansiirtoa kytkemällä yleisen jaetun muistin, ytimet, I/O hubin ja muut Intelin suorittimet toisiinsa. (kuva 1.)
Integroidut muistiohjaimet ja Intelin QuickPath teknologia toimivat yhdessä toimittaakseen seuraavat ominaisuudet:

- Skaalautuvan, high-performance kommunikoinnin
- Muistin suorituskyvyn ja joustavuuden tukea johtavia muistiteknologiota
- Tiiviisti integroidun interconnectin luotettavuuden, saatavuuden ja huollettavuuden.

(Maximizing multicore processor performance 2019)

2.2.2 Dedicated GPU - näytönohjain

Dedicated GPU eli selkokielellä näytönohjain on integroitua näytönohjainta todella paljon tehokkaampi. Kun tehdään raskaampaa työtä esimerkiksi: pelataan tai esimerkiksi lähetetään livekuvaa videopelistä internettiin, tarvitaan tietokoneeseen erillinen näytönohjainkortti.(kuva 2.)

KUVA 2. Kaksi erillistä näytönohjainta kytketty toisiinsa Multi-GPU modella. (guru3d.com 2016)

Ulkoisen näytönohjaimen tärkeimmät osat ovat kortin grafiikkasuoritin(GPU), näyttömuisti(GDDR), jäähdytyslaitteisto ja ulostuloportit näytöille. Näytönohjain kytketään emolevyltä löytyvään PCIe x16 -tiedonsiirtoväylään. Kytkemällä näytönohjain kyseiseen tiedonsiirtoväylään sallitaan emolevyn ja näytönohjaimen kommunikointi keskenään suurimmalla mahdollisella siirtonopeudella. (guru3d.com 2016.)

Näytönohjaimen keskeisin osa eli grafiikkasuoritin hallitsee kuvan piirtoa näytölle. Esimerkkinä Nvidian uusin lippulaivakortti GeForce RTX 2080 Founders edition (kuva 3) sisältää 2944 NVIDIA CUDA -ydintä eli varjostinta. Suorittimen peruskellotaajuus on 1515 MHz ja pelikäytössä kortti vahvistaa kellotaajuutensa jopa 1800 MHz:iin. RTX 2080 kortissa käytetään GDDR6-muistipiirejä, jotka toimivat 14 Gbps:n nopeudella ja muistin kaistanleveys on 448 GB/s. Kyseinen Nvidian kortti tukee jopa 7680x4320 resoluutiota ja siihen voidaan liittää enintään 4 näyttöä. Vakionäyttöliittiminä toimii DisplayPort, HDMI ja USB Type-C (kuva 4). (Nvidia Corporation 2019[1].)

KUVA 3. Nvidian lippulaivamalli GeForce RTX 2080 Founders Edition (Nvidia Corporation 2019[1])

KUVA 4. Nvidia RTX 2080 -kortin näyttöliittimet (Nvidia Corporation 2019[1])

Virtansa näytönohjaimet saavat tietokoneen virtalähteen näytönohjaimille tarkoitetuista pinneistä. Vanhemmat ja pienempitehoiset näytönohjaimet toimivat yhdellä kuuden pinnin virtaliittimellä, mutta uudemmat ja suurempitehoisemmat kortit vaativat kaksi kahdeksanpinnistä virtaliitintä. Suositeltavaa on, että tietokoneen virtalähteenä olisi vähintään 500 W, mutta uudempien ja tehokkaampien kokoonpanojen kanssa kannattaa valita 750 W:n virtalähde. NVIDIA ilmoittaa RTX 2080:n tehoksi 225 W ja suosittelee kortin kanssa 650 W virtalähdettä. (Nvidia Corporation 2019[1])

2.2.3 Multi GPU Mode - usean näytönohjaimen kokoonpanot

Nvidian SLI ja AMD:n Crossfire teknologia sallii jopa neljän saman generaation grafiikkakortin yhdistämisen toisiinsa. Tämän teknologian avulla raaka peliteho pystytään tehokkaasti tuplaamaan, kolminkertaistamaa tai jopa nelinkertaistamaan, mutta vain teoriassa.

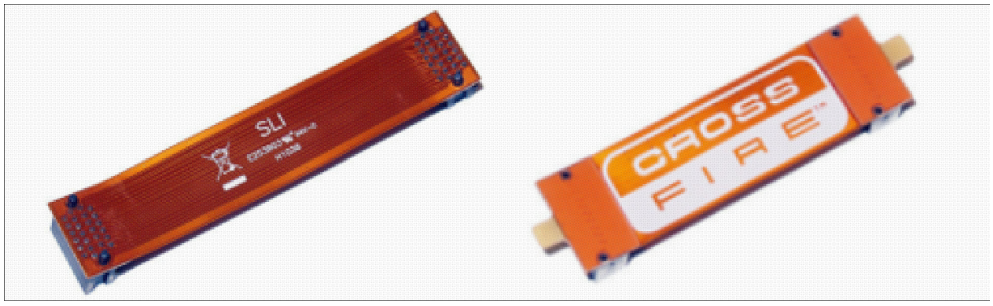
Alkaen Nvidian Pascal arkkitehtuurin näytönohjaimista Nvidia tukee suurimmaksi osaksi kahden grafiikkakortin kokoonpanoja. Asia on näin, koska mitä enemmän grafiikkakortteja kokoonpano sisältää, sitä huonommaksi skaalautuminen menee. Tämä johtaa siihen, että grafiikkaohjainten kanssa ilmenee enemmän ongelmia. Nykypäivänä kahden grafiikkakortin kokoonpano on ideaali, jos haluaa hyödyntää Multi GPU -teknologiaa. (Multi GPU Mode Explained, 2016.)

Useamman grafiikkakortin yhdistämiseen tarvitaan näytönohjaimen valmistajasta riippuen joko AMD:n Crossfireä tai Nvidian SLI:tä tukevaa emolevyä.

Melkein kaikki emolevyt tukevat AMD:n Crossfire-ominaisuutta. Ainoana vaatimuksena AMD:n korteille on PCIe x16 -liitin emolevyllä. Nvidian kanssa asia ei ole niin yksinkertainen. Nvidian SLI-teknologiaa tukevat vain seuraavat emolevyt:

- P55-sarjan emolevyt
- P67-sarjan emolevyt
- Z68-sarjan emolevyt
- X58-sarjan emolevyt
- Z77-sarjan emolevyt
- Z87-sarjan emolevyt
- X79-sarjan emolevyt
- Z79-sarjan emolevyt
- X99-sarjan emolevyt.

Kytkettäessä ja otettaessa käyttöön useampi grafiikkakortti, tarvitaan myös AMD:n Crossfireiitin (kuva 5) tai Nvidian tapauksessa SLI-silta (kuva 6).

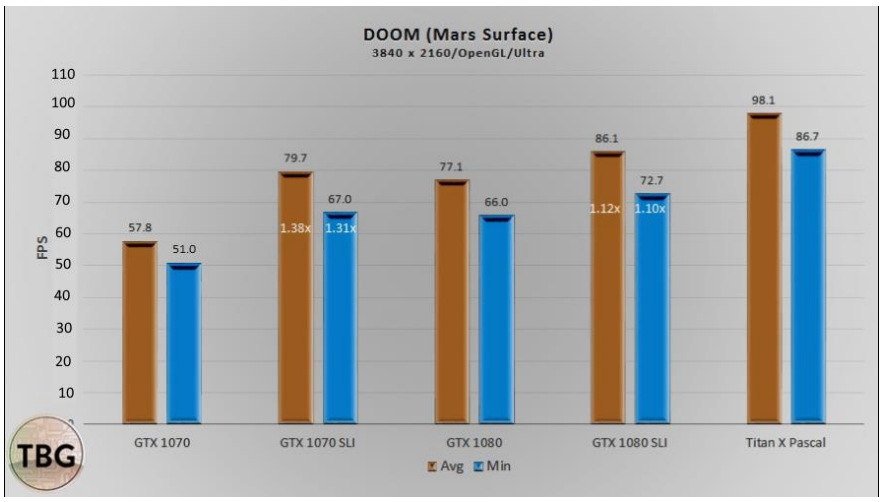


KUVA 5. AMD:n Crossfireliitin (AMD 2019)

KUVA 6. Nvidian SLI-sillat kahdesta neljälle kortille. (Nvidia Corporation 2019[2])

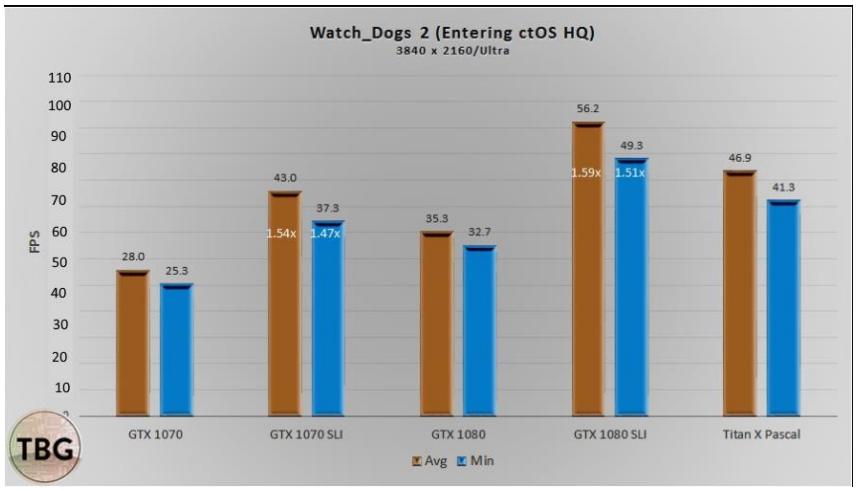
Vertaillessa Nvidian valmistamia GTX 1070- ja GTX 1080-grafiikkakorttien suoritustehoja useamman grafiikkakortin kokoonpanossa, pystytään toteamaan, että suorituskyky ei tosiasiassa kaksinkertaistu peleissä, vaikka niin voisi kuvitella

Vertailussa käytettiin peliä DOOM 3840x2160 resoluutiolla ja pelin maksimi grafiikka-asetuksilla (kuva 7). Kuvasta voidaan todeta, että kytkemällä kaksi GTX 1070 -grafiikkakorttia toisiinsa SLI-teknologiaa hyödyntäen, pelin maksimi FPS (Frames Per Second) kasvaa vain noin 20% ja saman verran, kun kytketään kaksi hieman tehokkaampaa GTX 1080 -korttia toisiinsa. Kuvassa näytetään myös yhden Nvidian Titan X Pascalin suorituskyky, joka on tehokkaampi, kuin kaksi GTX 1080 korttia SLI:ssä.



KUVA 7. Nvidian Grafiikkakorttien vertailu DOOM -pelissä (The 4k Gaming Showdown 2016)

Joidenkin pelien optimointi kuitenkin tukee todella hyvin SLI/Crossfire -teknologiaa, jolloin kahden kortin yhdistämisestä saadaan suurempi etu, kuten seuraavasta kuvasta 8 huomataan, kun analysoitiin peliä Watch Dogs 2 3840x2160 resoluutiolla



KUVA 8. Watch Dogs 2 pelin suorituskyky eri Nvidian korteilla(The 4k Gaming Showdown 2016)

Kuvasta 8 voidaan todeta, kuinka kahden kortin SLI-teknologia antaa lähes 1.54-1.59 kertaisen edun verrattuna vain yhteen grafiikkakorttiin ja myöskin voittaa edellisessä vertailussa voittaneen Titan X Pascalin. Pelien optimointi siis vaikuttaa huomattavasti SLI:n/Crossfiren vaikutukseen pelejä pelatessa. Tavalliselle kuluttajalle suositeltavaa olisi hankkia yksi tehokas kortti kahden halvemman kortin SLI:n/Crossfiren hinnalla

3 NVIDIA

3.1 Yleistä Nvidiasta

NVIDIA Corporation on Yhdysvalloissa perustettu grafiikkateknologia-alan yhtiö. Yhtiö tunnetaan parhaiten heidän grafiikkasuorittimiensa ansiosta. Nvidia on myös valmistanut mobiililaitteille tarkoitettuja mikropiirejä sekä ammattilaiskäyttöön tarkoitettuja tuotteita. Yhtiön tunnetuimpiin tuotteisiin kuuluu GeForce-sarjan grafiikkaprosessorit. Ammattilaiskäyttöön Nvidia myös valmistaa Tesla-sarjan mikroarkkitehtuureja ja Quadro-sarjan arkkitehtuuria, joka perustuu GeForce -piireihin.

Nvidian tuotteet keskittyvät pelaamiseen, ammattilaistason suunnittelun ja suurta suorituskykyä vaativaan laskentaan joihin Nvidia tarjoaa prosessorit, vaadittavat ohjelmat ja työkalut.

Nvidia tarjoaa pelaajille GeForce Experience -ohjelman, joka tukee Nvidian näytönohjaimia. Tällä ohjelmalla voidaan optimoida pelit napin painalluksella omalle kokoonpanolle ja ohjelma huolehtii ajureiden ajan tasalla pysymisen. Ohjelmalla voidaan myös tallentaa ja jakaa videoita, näyttökuvia tai live-striimejä ystäville Youtubessa, Twitchissä tai facebookissa. Ohjelmalla voidaan myös suoratoistaa kuvaa SHIELD -laitteille. (Nvidia Corporation 2019[3].)

Ammattilaistason työskentelyssä käytetään hyväksi Nvidian Quadro -sarjan näytönohjaimia. Näitä näytönohjaimia käytetään esimerkiksi suunnitteluissa, digitaalisen sisällön luonnissa ja esimerkiksi lääketieteellisissä kuvauksissa.

Tesla-näytönohjaimia käytetään tieteellisissä tutkimuksissa. Näillä korteilla voidaan mallintaa aivoja ja lääkkeen löytymistä erilaisia sairauksia vastaan. Nvidian kortteja nähdään useissa maailman nopeimmissa supertietokoneissa. (Nvidia Corporation 2019[4].)

Nvidialla on myös mobiililaitteille tarkoitettu Tegra-mallisto. Tätä mallistoa käytetään puhelimista auton viihdejärjestelmiin. Nvidian tuotteita on käytetty itsestään ajavissa autoissa. (Nvidia Corporation 2016[5].)

3.1.1 Nvidian historiaa

Vuonna 1993 Jen-Hsun Huang, Chris Malachowsky ja Curtis Priem perustivat Nvidia Corporationin Kalifornian Santa Clara:ssa. Yhtiön tunnuslause ”The Way It’s Meant to Be Played” viittaa siihen, että Nvidian grafiikkaprosessorit soveltuvat pelien pelaamiseen. (Nvidia Corporation 2016[6].)

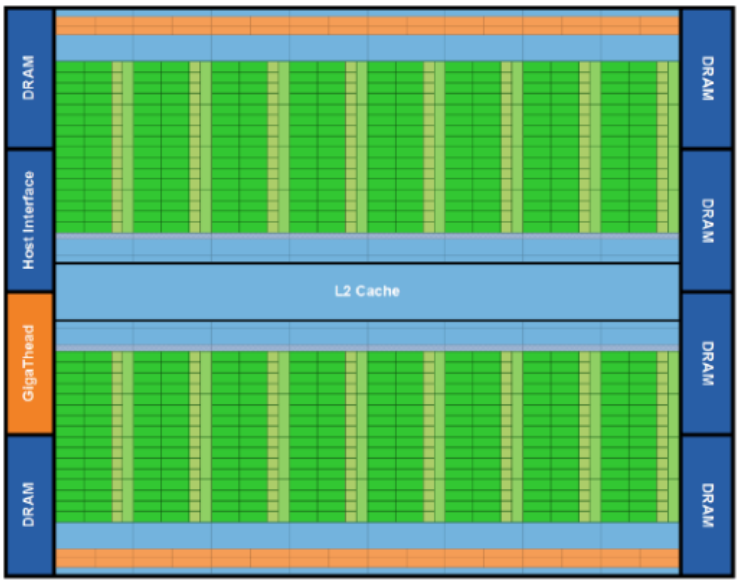
Vuonna 1995 Nvidia julkaisi ensimmäisen tuotteensa, NV1-multimediapiirikortin, joka tuki 2D ja 3D grafiikoita. Segan peli Virtual Fighter oli ensimmäisiä pelejä, joka toimi Nvidian korteilla. Vuoden kuluttua Nvidia julkaisi ensimmäiset Microsoft DirectX -ohjelmointirajapinnan ajurinsa, jotka tukivat Direct3D:tä eli 3D grafiikan renderöintiä. Vuonna 1999 Nvidia kehitti ensimmäisen GPU:nsa GeForce 265:n. Vuonna 2001 Nvidia julkaisi ensimmäisen ohjelmoitavan GPU:n Nvidia GeForce 3:n, joka sallii kehittäjien luoda visuaalisia tehosteita. 2002 Nvidia nimettiin Amerikan nopeimmin kasvavaksi yritykseksi. Vuosi nimeämisen jälkeen yhtiö kehitti tunnetun SLI-teknologian, joka sallii usean grafiikkakortin yhdistämisen toisiinsa. (Nvidia Corporation 2016[6].)

Vuosi 2006 oli merkittävä sillä, silloin Nvidia julkaisi CUDA-ytimet. Fermi-arkkitehtuuri julkaistiin vuonna 2009, Kepler-arkkitehtuuri vuonna 2012, Maxwell-arkkitehtuuri vuonna 2014, Pascal-arkkitehtuuri vuonna 2016, Volta-arkkitehtuuri vuonna 2017 ja vuonna 2018 Nvidia julkaisi Turing-arkkitehtuurin, joka maailman ensimmäisenä ja ainoana arkkitehtuurina tukee RTX:ää eli reaaliaikaista säteenseurantaa peleissä. (Nvidia Corporation 2016[6].)

3.2 Arkkitehtuurit

3.2.1 Fermi-arkkitehtuuri

Vuonna 2009 Nvidia julkisti yhtiön ensimmäisen näytönohjainarkkitehtuurinsa, Fermin. Fermi-arkkitehtuuriin perustuvia näytönohjaimia on GeForce-, Quadro- ja Tesla-tuoteperheissä (Nvidia Corporation 2016[7]). Tämä näytönohjainarkkitehtuuri rakentuu kolmesta miljardista transistorista ja piirit ovat valmistettu 40 nanometrin prosessilla. Kuvassa 9 nähdään Nvidian Fermi-arkkitehtuurin rakenne.



KUVA 9. Fermi arkkitehtuurin rakenne (Nvidia Corporation 2016[7])

Fermi-arkkitehtuurin varjostin -yksikköjen määrä on jopa 512. Nämä yksiköt ovat aikaisemmin mainittuja CUDA-ytimiä. Nämä ytimet ovat jaoteltu 16 Streaming Multiprosessoriin (SM), eli toisin sanoen kussakin streaming multiprosessorissa on CUDA-ytimiä 32 kappaletta.

Lisäksi näillä multiprosessoreilla on myös käytössä 64 kilotavua konfiguroitavaa muistia ja L2-välimuistia 768 kilotavua. Multiprosessoreiden konfiguroitava muisti on mahdollista jakaa 16/48 kilotavun tai 48/16 kilotavun palasiin, joista ensimmäinen puolikas on prosessorin L1-välimuisti ja toinen puolikas jaettua muistia.

Arkkitehtuurin mukana myös päivitettiin GigaThread Engine, joka kykenee siirtämään suuria määriä dataa molempiin suuntiin saman aikaisesti. Fermistä löytyy kuusi 64 -bittistä muistiohjainta, joka tarkoittaa, että muistiväylän suuruus kokonaisuudessaan on 384 bittiä leveä. Arkkitehtuuri tukee GDDR5-muistia. Ensimmäiset Fermi-arkkitehtuuria käyttävät näytönohjaimet julkaistiin vuonna 2009. Näihin näytönohjaimiin lukeutuu muun muassa seuraavat näytönohjaimet:

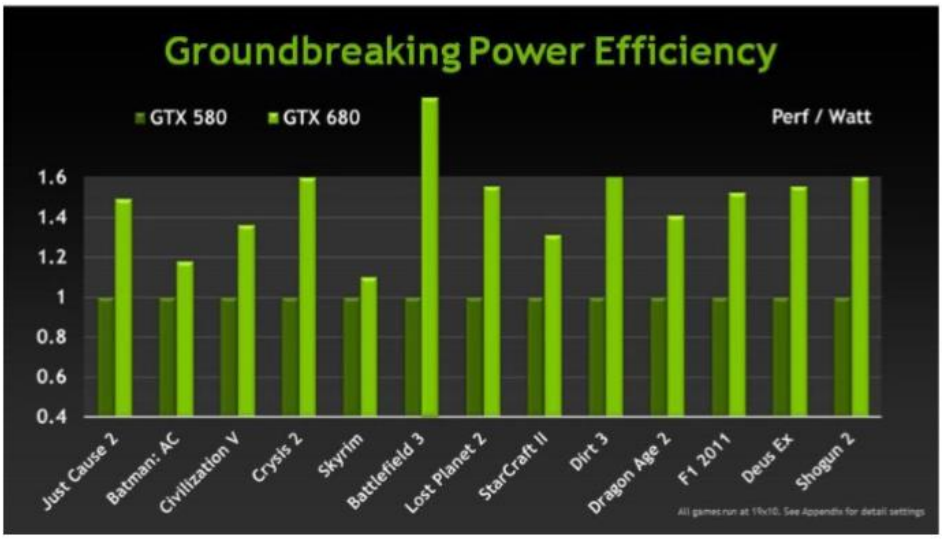
- NVIDIA Geforce 410M
- NVIDIA Geforce 810M
- NVIDIA Geforce GT 415M
- NVIDIA Geforce GT 550M
- NVIDIA Geforce GTX 480
- NVIDIA Geforce GTX 550 Ti
- NVIDIA Geforce GTX 580
- NVIDIA Geforce GTX 670M

3.2.2 Kepler-arkkitehtuuri

Vuonna 2012 Nvidia julkisti toisen näytönohjainarkkitehtuurinsa, Keplerin. Ensimmäinen Nvidian grafiikkasuoritin, joka käytti Kepler-arkkitehtuuria, oli GTX 680. Yhtälailla kuin Kermi-arkkitehtuurissa, Keplerin grafiikkaprosessorit koostuvat GPC:istä, streaming multiprosessoreista ja muistikontrollereista. Kuten kuvasta 11 nähdään GTX 680 -grafiikkasuoritin sisältää neljä GPC:tä, kahdeksan SMX:ää ja neljä muistikontrolleria. (Nvidia Corporation 2016[8].)

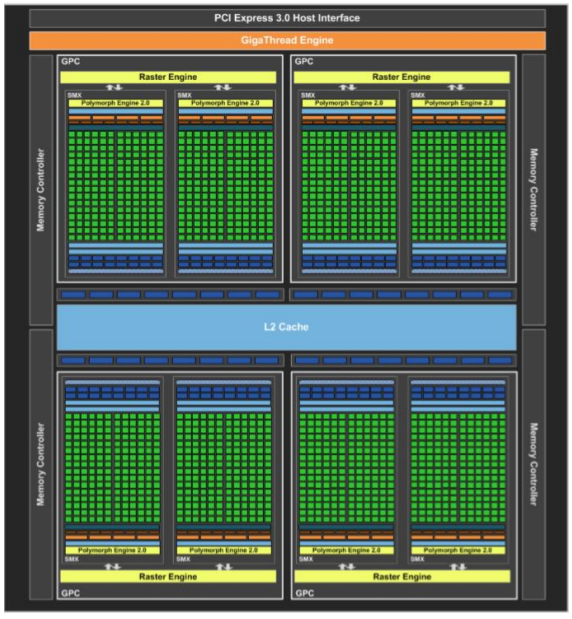
Huomattavin uudistus Kepler-arkkitehtuurissa oli sen merkittävä suorituskyky wattia kohden. Tämä oli mahdollista Nvidian uuden streaming multiprosessorin, SMX:n ansiosta. SMX:ää hyödyntäen Nvidia näki mahdollisuuden vähentää GPU:n virrankulutusta Kepler-arkkitehtuurin kanssa. 1536 CUDA-ytimen streaming multiprosessorin (GK104) avulla, GTX 680 SMX tuottaa kaksinkertaisen suorituskyvyn wattia kohden verrattuna vanhempaan Fermi-arkkitehtuurin streaming multiprosessoriin (GF110). Tämän kaiken ansiosta GeForce GTX 680

mahdollistaa vallankumouksellisen suorituskyvyn wattia kohden verrattuna vanhempaan GTX 580:een (kuva 10)



KUVA 10 GTX 680 ja GTX 580 suorituskyky wattia kohden(Nvidia Corporation 2016[8].)

Kuvassa 10 on vertailtu GeForce GTX 680 ja GeForce GTX 580 GPU:ita. Näiden korttien suorituskyky wattia kohden on testattu eri peleillä. Kuvasta nähdään, että GTX 680:n suorituskyky wattia kohden on huomattavasti paljon parempi, kuin GTX 580:n.



KUVA 11. Kepler arkkitehtuuria käyttävän GeForce GTX 680:n rakenne. (Nvidia Corporation 2016[8].)

Uudessa GTX 680:ssa jokaisella GPC:llä on oma rasterimoottori ja kaksi SMX yksikköä. Kuten kuvasta 11 nähdään GTX 680:ssa on yhteensä kahdeksan SMX yksikköä, joissa on yhteensä 1536 CUDA-ydintä.

GeForce GTX 680:n muistin alijärjestelmä uudistettiin totaalisesti, mikä sallii todella paljon suuremmat muistin kellotaajuudet. GTX 680:n muistinopeus oli 6008 MHz, joka siihen aikaan oli nopein muistinnopeus markkinoilla. Nvidia esitteli myös GPU Boostin pelaajille Keplerin myötä, joka säätää GPU:n kellotaajuutta sen tehonkulutuksen mukaan. (Nvidia Corporation 2016[8].)

GTX 680 -suorittimessa on yhteensä neljä muistikontrolleria ja jokaisessa muistikontrollerissa on 128 KB L2-välimuistia ja kahdeksan ROP-yksikköä. Jokainen ROP-yksikkö prosessoi yksittäistä värinäytettä. Näillä neljällä muistikontrollerilla GTX 680 GPU:ssa on yhteensä 512 KB L2-välimuistia ja 32 ROP-yksikköä, eli 32 värinäytettä. (Nvidia Corporation 2016[8].)

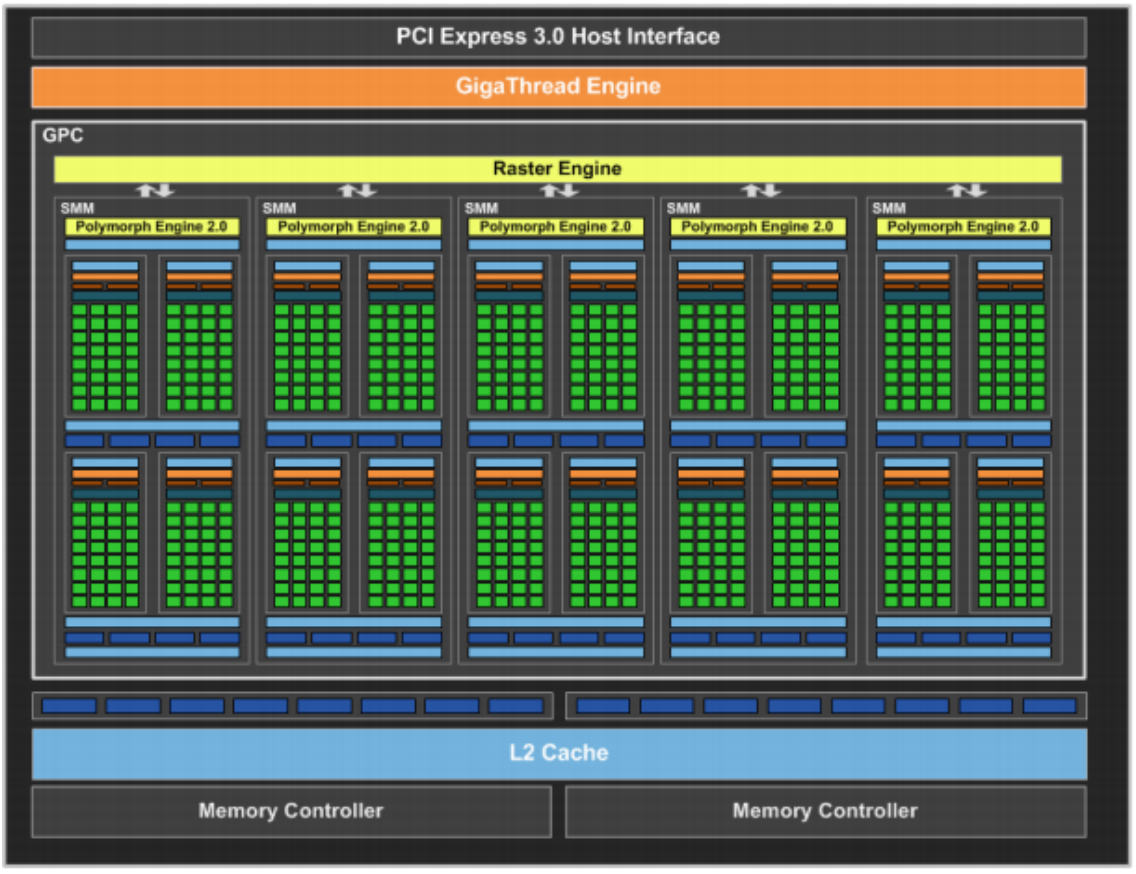
3.2.3 Maxwell-arkkitehtuuri

Vuonna 2014 Nvidia julkisti uuden Maxwell arkkitehtuurinsa, joka toimii GM107 -grafiikkapiirillä. Ensimmäisen sukupolven Maxwell-grafiikkapiireissä keskityttiin optimoimaan suorituskyvyn ja virrankulutuksen hyötysuhdetta. Edeltäjänsä Keplerin tavoin, Maxwell-piirit valmistettiin 28 nanometrin viivanleveydellä. Uudessa Maxwell-arkkitehtuurissa on käytetty uudistettuja streaming multiprosessoreita, jotka parantavat suorittimien energiatehokkuutta huomattavasti. Uusi multiprosessoreiden arkkitehtuuri sallii multiprosessoreiden määrän kasvatuksen viiteen GM107-grafiikkapiirissä, verrattuna vanhemman GK107-grafiikkapiirin kahteen multiprosessoriin. (Nvidia Corporation 2016[9].)

Maxwell arkkitehtuurissa CUDA-ydinten määrä on myös laskettu kahteen. Kuitenkin Maxwellin suoritustehokkuuden ansiosta, suorituskyky yhtä streaming multiprosessoria kohden pyörii 10 % luokilla Kepleriin verrattuna. Tämä tarkoittaa siis sitä, että streaming multiprosessorin parannuksen ansiosta CUDA-ydinten määrä GPU:ta kohden on huomattavasti suurempi, verrattuna Fermi -tai Kepler -piireihin.

Tämä ensimmäinen grafiikkapiiri GM107 on suunnattu kannettaviin tietokoneisiin ja työpöytäpuolella se on käytössä edullisissa GeForce GTX 750- ja GTX 750 Ti -näytönohjaimissa. GeForce GTX 750 -Ti kortissa käytettiin täyttä GM107 -grafiikkapiiriä, joka tarkoittaa sitä, että siinä oli käytössä kaikki 640 CUDA-ydintä ja 40 tekstuuriyksikköä eli ROP:ia. (Nvidia Corporation 2016[10].)

GM107-grafiikkapiiri toimii 1020 MHz:n peruskellotaajuudella ja GPU boostia käyttäessä kellotaajuus nousee 1085 MHz:n taajuuteen. Grafiikkapiirin 128 -bittisen muistiväylän jatkeena on kaksi gigatavua 1350 MHz:in kellotaajuudella toimivaa GDDR5-muistia. (Muropaketti GTX 750 2014)



KUVA 12 GM107-grafiikkapiirin rakenne(Nvidia Corporation 2016[10])

Kuvasta 12 nähdään että, GM107-grafiikkapiiri koostuu yhdestä GPC:stä, viidestä Maxwellin streaming multiprosessorista (SMM) ja kahdesta 64-bittisestä muistikontrollerista. Maxwellin L2-välimuisti on kokonaisuudessaan 2048 KB, joka on melkein 10 kertaa suurempi kuin edeltäjänsä Keplerin, joka oli 256 KB.

Tämä tarkoittaa sitä, että kun grafiikkapiirillä on enemmän välimuistia sitä vähemmän grafiikkakortin DRAM:ille tulee pyyntöjä, joka vähentää kokonaisuudessaan kortin virrankulutusta ja parantaa suorituskykyä. Tämä kaikki tarkoittaa sitä, että Maxwell pystyy antamaan kaksinkertaisen suorituskyvyn wattia kohden verrattuna Kepleriin käyttäen samaa 28 nanometrin tuotantoprosessia. (Nvidia Corporation 2016[10].)

3.2.4 Pascal-arkkitehtuuri

Vuoden 2016 keväällä Nvidia esitteli uuden arkkitehtuurinsa Pascalin. Pascal-arkkitehtuuriin perustuvat grafiikkapiirit valmistetaan uudella 16 nanometrin viivanleveydellä. Suurin grafiikkapiiri GP100 rakentuu 15.3 miljardista transistorista ja sen ominaisuuksiin lukeutuu esimerkiksi 3840 CUDA-ydintä ja 15 gigatavua toisen sukupolven HBM2-muistia. Pienemmän 16 nanometrin valmistustekniikan ansiosta, grafiikkapiiri on pinta-alaltaan 21 % pienempi kuin 28 nanometrin valmistustekniikan piiri. 16 nanometrin valmistustekniikalla valmistetuissa grafiikkapiireissä on kuitenkin 38 % enemmän transistoreita. Virrankulutus uudella tekniikalla on vain 15 wattia korkeampi. (Nvidia Corporation 2016[4].)

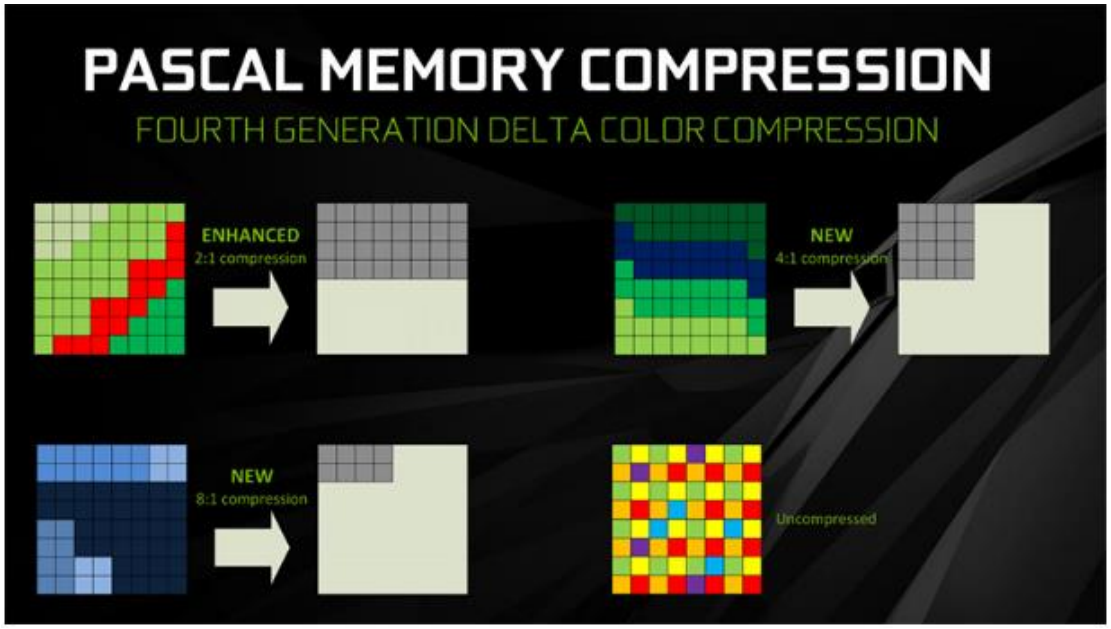
Täysi GP100-grafiikkapiiri koostuu kuudesta GPC:stä, 60:stä Pascalin streaming multiprosessorista, 30 TPC:stä, joista jokainen sisältää kaksi streaming multiprosessoria ja grafiikkapiiri sisältää myös kahdeksan 512-bittistä muistikontrolleria.(kuva 13).

KUVA 13. Pascal GP100:n täysi rakenne.(Nvidia Corporation 2016[4].)

Piirin jokainen GPC sisältää kymmenen streaming multiprosessoria. Jokaisessa multiprosessorissa on 64 CUDA-ydintä ja neljä tekstuuriydintä. Kaikki 60 streaming multiprosessoria yhteen laskettuna GP100 sisältää yhteensä 3840 CUDA-ydintä ja 240 tekstuuriydintä. GP100-grafiikkapiiri sisältää 4096 KB L2-välimuistia. Kuvassa 13 nähdään kokonainen GP100-grafiikkapiiri, joka sisältää kaikki 60 SM-yksikköä. (Nvidia Corporation 2016[4])

Pascal-arkkitehtuurissa sen geometriayksikkö eli Polymorph Engine on päivitetty versioon 4.0. Tämän uutena ominaisuutena siihen on lisätty Simultaneous Multi-Projection -yksikkö, joka sijaitsee geometrialiukuhihnan lopussa ja se kykenee monistamaan yhden geometriadatan 16 eri kuvakulmaan. Tämän avulla voidaan tulostaa usean näytön surround-järjestelmissä kuva oikeassa perspektiivissä kaikille näytöille, koska pelien perspektiivi on yleensä väärentynyt sen takia, koska pelit käyttävät vain yhtä kuvakulmaa kuvan renderöintiin. (Testissä näytönohjaimet… 2016.)

Pascalin muistiarkkitehtuuri on päivitetty neljännen sukupolven Delta Color pakkausalgoritmilla. Tämän uudistuksen myötä pelit vaativat noin 20 % vähemmän muistikaistaa saavuttaakseen saman suorituskyvyn kuin edeltäjällänsä Maxwellilla. (Nvidia Corporation 2016[4].)



KUVA 14. Pascalin uudistettu pakkausalgoritmi(Nvidia Corporation 2016[4])

Grafiikkapiirissä käytetty uusi GDDR5X-standardi mahdollistaa tiedonsiirron jopa 10 Gbps nopeudella. Tämä standardi perustuu vahvasti GDDR5-muisteihin. Yhdistäessä GDDR5X-muistit ja Delta Color pakkausalgoritmin, parantuu muistiväylän kaistanleveys noin 1.7 kertaiseksi verrattuna Maxwellin kaistanleveyteen.

Nvidia julkaisi myös Tesla P100 -laskentakortin. Laskentakortti perustuu GP100-grafiikkapiiriin, toisen sukupolven HM-muisteihin ja uuteen Pascal-arkkitehtuuriin. Tesla P100 -laskentakortti on tarkoitettu suurteholaskentaan ja niitä käytetään supertietokoneissa. Ensimmäinen Nvidian GeForce GTX 1080 -kortti, joka käyttää Pascal-arkkitehtuuria, julkaistiin 27. toukokuuta vuonna 2016. (Nvidia Corporation 2016[4].)

3.2.5 Volta-arkkitehtuuri

Vuonna 2017 Nvidia julkaisi ensimmäisen kuluttajamarkkinoille tarkoitetun näytönohjaimen Volta-arkkitehtuurilla. Titan V:ssä käytettävä Volta GV100 -grafiikkapiiri sisältää 5120 CUDA-ydintä, 640 Tensor-ydintä ja yhteensä 320 tekstuuriyksikköä. Kellotaajuus grafiikkapiirillä on 1.2 MHz ja Boost kellotaajuus jopa 1.45 MHz. (Nvidia Corporation 2019[11].)

GV100-grafiikkapiirin pinta-ala on reilu 815 neliömillimetriä ja transistoreita grafiikkapiiriltä löytyy 21 miljardia. Valmistamiseen on käytetty 12 nanometrin FinFET+ -valmistusprosessia.

Vaikka Titan V -kortti on suunnattu kuluttajamarkkinoille, kortti on tarkoitettu tehokkaan suorituskykynsä takia erityisesti koneoppimiseen ja tekoälysovelluksiin. Titan V:ssä koneoppimiseen tarkoitetut Tensor-ytimet tuottavat 110 teraFLOPSin laskentatehon. Grafiikkaprosessorina grafiikkapiirin teho on noin 15 teraFLOPSia.

KUVA 15 Volta GV100 grafiikkapiiri 84:llä streaming multiprosessorilla.(Nvidia Corporation 2019[11])

Kuvassa 15 nähdään kokonaisen Volta100-grafiikkapiirin rakenne. Täysi grafiikkapiiri sisältää kuusi GPC:tä joista jokainen sisältää seitsemän TPC:tä ja 14 streamin multiprosessoria. Grafiikkapiirissä on myös 84 Volta streaming multiprosessoria, joista jokainen sisältää 64 FP32-ydintä, 64 INT32-ydintä, 32 FP64-ydintä, 8 tensor-ydintä ja neljä tekstuuriyksikköä. Piirissä on kahdeksan 512-bittistä muistikontrolleria, joista koostuu 4096 bittiä.

Kokonainen GV100-grafiikkapiiri 84:llä streaming multiprosessorilla sisältää yhteensä 5376 FP32-ydintä, 5376 INT32-ydintä, 2688 FP64-ydintä, 672 tensor-

ydintä ja 336 tekstuuriyksikköä. Piirin jokaista HBM2 DRAM -pinoa ohjataan muistikontrolleriparilla. Piiri sisältää 6144 KB L2 välimuistia.

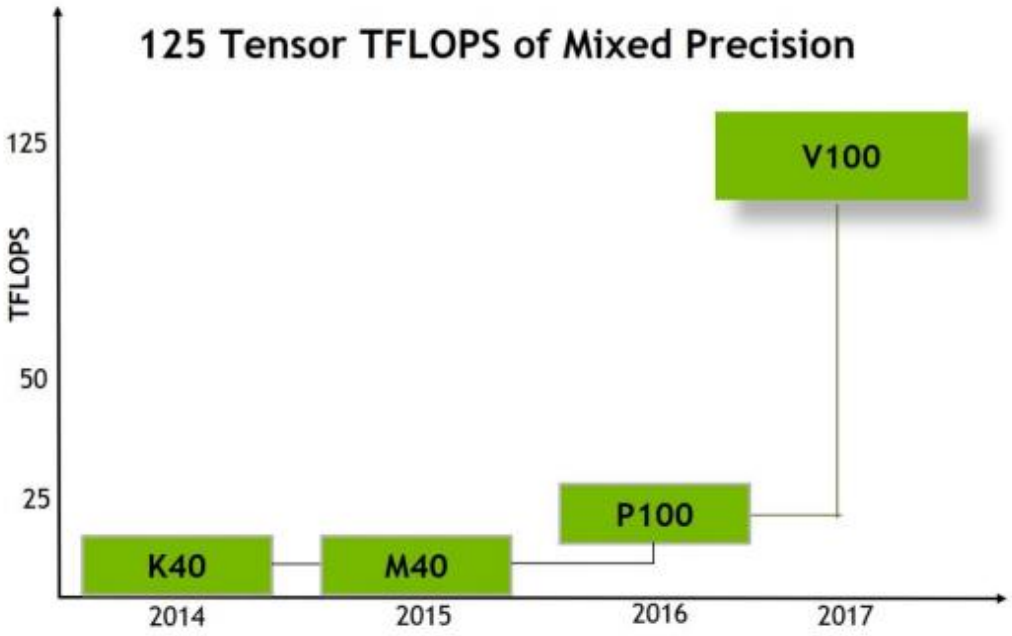
Voltan uudessa streaming multiprosessorissa on lukuisia uudistuksia vanhempaan Pascalin multiprosessoriin. Suurimmat uudistukset multiprosessorissa ovat uudistuneet Tensor-ytimet, jotka tuottavat 12 -kertaisen teraFLOPSin tehon verrattuna Pascalin GP100-grafiikkapiiriin samalla virrankulutuksella. Streaming multiprosessorissa on myös 50 % korkeampi energiatehokkuus. L1-välimuistia on myös paranneltu kyseisessä multiprosessorissa (Nvidia Corporation 2019[11].)

TAULUKKO 1 Tesla-suorittimien vertailu (Nvidia Corporation 2019[11])

Tesla Product	Tesla K40	Tesla M40	Tesla P100	Tesla V100
GPU	GK180 (Kepler)	GM200 (Maxwell)	GP100 (Pascal)	GV100 (Volta)
SMs	15	24	56	80
TPCs	15	24	28	40
FP32 Cores / SM	192	128	64	64
FP32 Cores / GPU	2880	3072	3584	5120
FP64 Cores / SM	64	4	32	32
FP64 Cores / GPU	960	96	1792	2560
Tensor Cores / SM	NA	NA	NA	8
Tensor Cores / GPU	NA	NA	NA	640
GPU Boost Clock	810/875 MHz	1114 MHz	1480 MHz	1530 MHz
Peak FP32 TFLOPS[1]	5	6.8	10.6	15.7
Peak FP64 TFLOPS[1]	1.7	.21	5.3	7.8
Peak Tensor TFLOPS[1]	NA	NA	NA	125
Texture Units	240	192	224	320
Memory Interface	384-bit GDDR5	384-bit GDDR5	4096-bit HBM2	4096-bit HBM2
Memory Size	Up to 12 GB	Up to 24 GB	16 GB	16 GB
L2 Cache Size	1536 KB	3072 KB	4096 KB	6144 KB
Shared Memory Size / SM	16 KB/32 KB/48 KB	96 KB	64 KB	Configurable up to 96 KB
Register File Size / SM	256 KB	256 KB	256 KB	256KB
Register File Size / GPU	3840 KB	6144 KB	14336 KB	20480 KB
TDP	235 Watts	250 Watts	300 Watts	300 Watts
Transistors	7.1 billion	8 billion	15.3 billion	21.1 billion
GPU Die Size	551 mm²	601 mm²	610 mm²	815 mm²
Manufacturing Process	28 nm	28 nm	16 nm FinFET+	12 nm FFN

Taulukossa 1 vertaillaan Nvidian julkaisemien Tesla-tuotteiden eri ominaisuuksia vuosien 2014 ja 2017 välillä. Taulukosta nähdään esimerkiksi, että uusimmassa Volta grafiikkapiirissä on melkein viisi kertaa enemmän streaming multiprosessoreita verrattuna Nvidian ensimmäiseen Tesla -tuotteen Kepler -piiriin. Transistoreiden määrä on myös kasvanut 14 miljardilla. Kuvasta 16 nähdään, että Volta

tarjoaa jopa 125 teraFLOPSin tehon nykypäivän syväoppimiseen uudistuneilla tensor-ytimillä.



KUVA 16 Tesla -tuotteiden laskentateho vuosien 2014 ja 2017 välillä (Nvidia Corporation 2019[11])

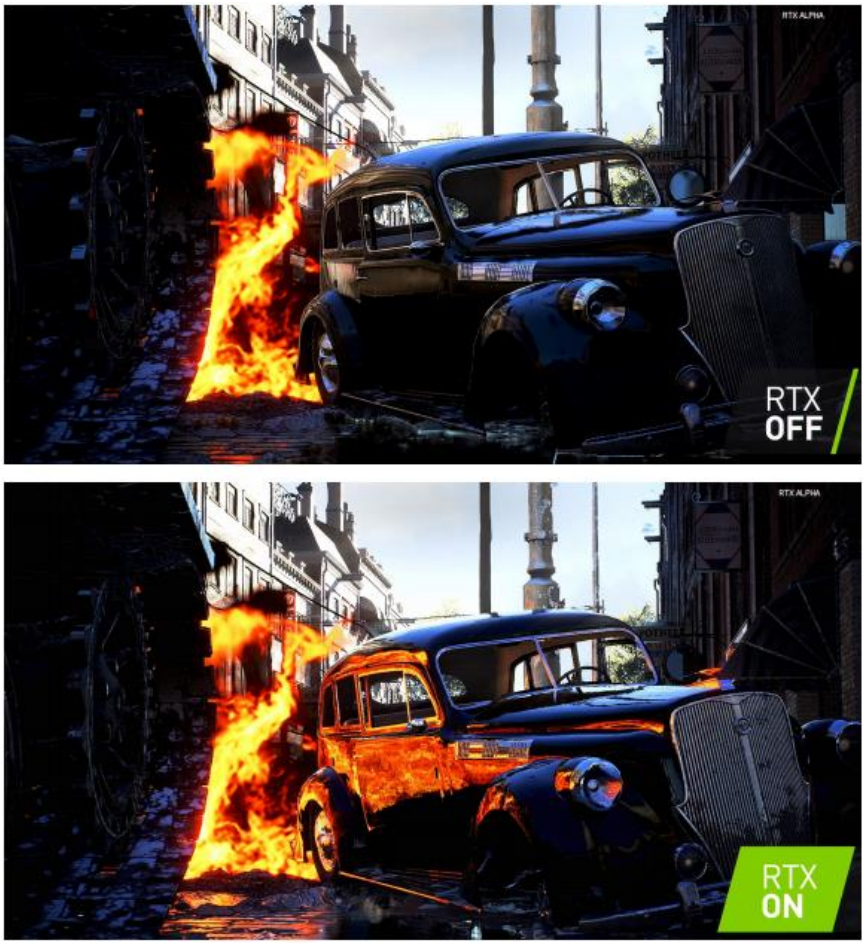
Kuten aikaisemmin mainittiin, Voltan uudistuneet tensor-ytimet ovat avainasemassa arkkitehtuurissa. Näiden avulla Volta GV100 -grafiikkapiiri pystyy toimittamaan tarvittavan suorituskyvyn, jota tarvitaan laajoissa neuroverkoissa. (Nvidia Corporation 2019[11])

Tesla V100 -grafiikkaprosessori sisältää kokonaisuudessaan 640 tensor-ydintä. Grafiikkaprosessorissa tensor-ytimet suorittavat 64 liuku -FMA operaatiota kelloa kohden. Kahdeksan tensor-ydintä, jotka ovat osa streaming multiprosessoria suorittavat yhteensä 512 FMA-operaatiota kelloa kohden.

3.2.6 Turing-arkkitehtuuri

Nvidian uusin Turing-arkkitehtuuri julkaistiin vuoden 2018 toisella puoliskolla. Merkittävin uudistus Turing-arkkitehtuurissa oli reaaliaikainen säteenseuranta peleissä. Uudet RTX-näytönohjaimet saavat tehonsa uudesta Turing-grafiikkasuoritinarkkitehtuurista ja täysin uudesta RTX-alustasta, jotka tuovat peleihin reaaliaikaisen säteenseurannan, tekoälyn ja jopa kuusinkertaisen suorituskyvyn edellisen arkkitehtuurin näytönohjaimiin. (Nvidia Corporation 2019[12].)

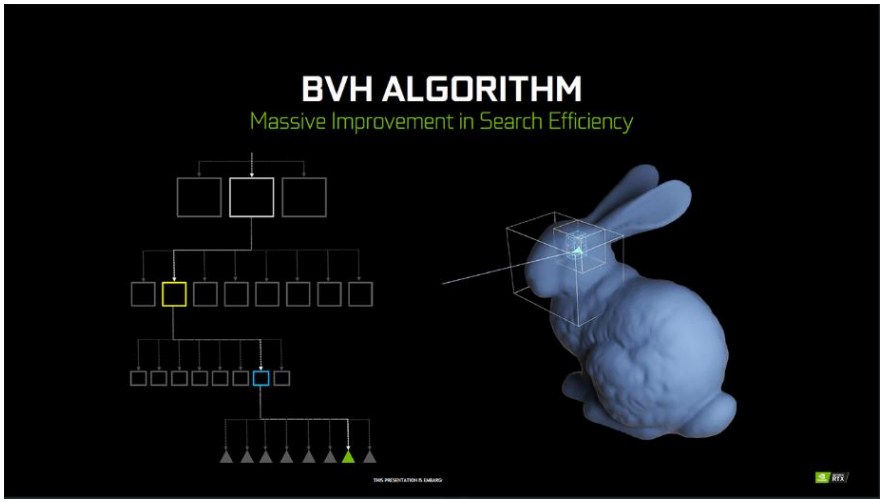
Säteenseurannalle on omistettu erillisiä RT-ytimiä grafiikkapiirissä. Nvidia mainostaa Turingin olevan maailman ensimmäinen säteenseurantagrafiikkapiiri (Nvidia Corporation 2019). Säteenseurannalla saadaan peleissä aikaan elävä valaistus, realistiset heijastukset ja varjot, jotka nostavat todenmukaisuuden huomattavasti perinteisten piirtotekniikoiden yläpuolelle(kuva 17). Turing-arkkitehtuurin sanotaan olevan suurin harppaus sitten CUDA-ytimien julkaisusta.



KUVA 17 Kuvankaappaus pelistä Battlefield V, jossa nähdään säteenseurannan (RTX:n) vaikutus grafiikoihin(Nvidia Corporation 2019[12])

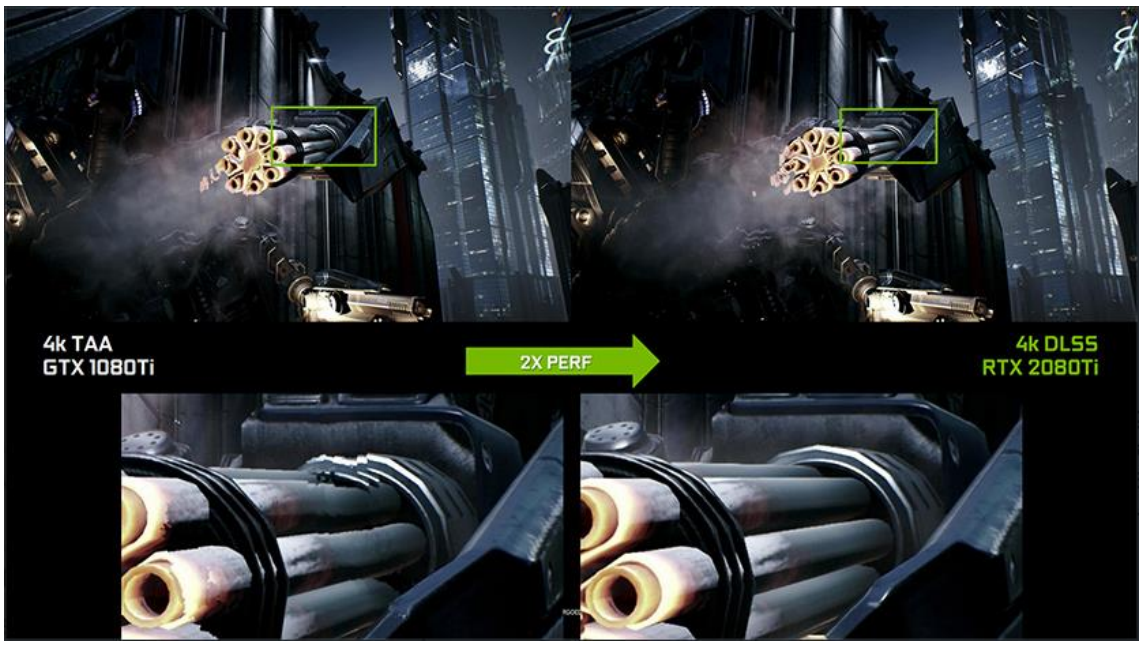
Säteenseuranta voi vaatia suuria määriä laskennallisia tehoja, jotta se voisi tuottaa realistisia näkymiä. Tämä johtuu siitä, kuinka paljon säteitä grafiikkapiiri joutuu piirtämään ja kuinka paljon säteitä heijastuu ja taittuu eri pinnoilta. Säteenseurantaa hyväksikäyttäen voidaan tuottaa kuvia, jotka ovat erottamattomia kuvista, jotka ovat otettu oikeasta elämästä. Säteenseurantatekniikkaa on käytetty elokuvissa, joissa yhdistetään oikeata kuvaa ja tietokoneella luotua kuvaa ja näitä on melkein mahdoton erottaa ihmissilmällä.

Ensimmäinen Turing-arkkitehtuurin uusista teknologioista on säteenseurannan kiihdytys. Arkkitehtuurin streaming multiprosessori -yksikköihin on lisätty uusi RT-yksikkö, joka siis mahdollistaa säteenseurannan kiihdytyksen laskemalla Bounding Volume Hierarchy -hakupuuhun ja säteiden risteämiseen liittyviä laskuja. (kuva 18.)



KUVA 18 BVH -hakupuun hierarkia(Nvidia Corporation 2019[12])

Säteenseurannan jälkeen yksi uusista teknologioista on Turing-arkkitehtuurin Deep Learning Super Sampling (DLSS) -renojenpehmennys. Kuvassa 19 vertaillaan Pascal-arkkitehtuurin GTX 1080 Ti -kortin ja uuden RTX 2080 Ti -kortin reunanpehmennyksiä.

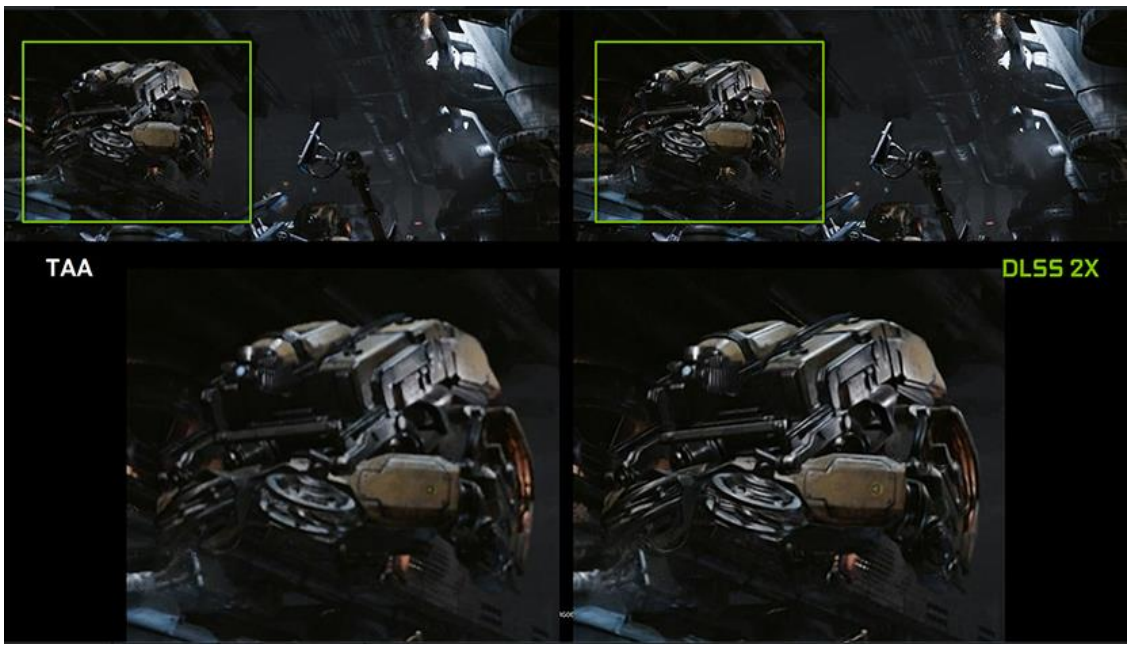


KUVA 19. Reunanpehmennyksen ero vanhempaan arkkitehtuuriin. (Nvidia Corporation 2019[12])

DLSS on uusi tekoälyä hyödyntävä reunojenpehmennysteknologia, joka vaatii erillisen tuen peliltä ja Nvidialta. Tämä tarkoittaa sitä, että kun pelille lisätään DLSS tuki, Nvidia opettaa tekoälylleen kyseisen pelin grafiikan ”perustotuuksia” käyttäen peräti 64 -kertaista super sampling -reunojenpehmennystä. Tämä muutaman megatavun kokoinen tietopaketti jaetaan pelin tai ajureiden mukana ja sitä hyödynnetään lopullisen kuvan muodostamiseen. (Nvidia Corporation 2019[12].)

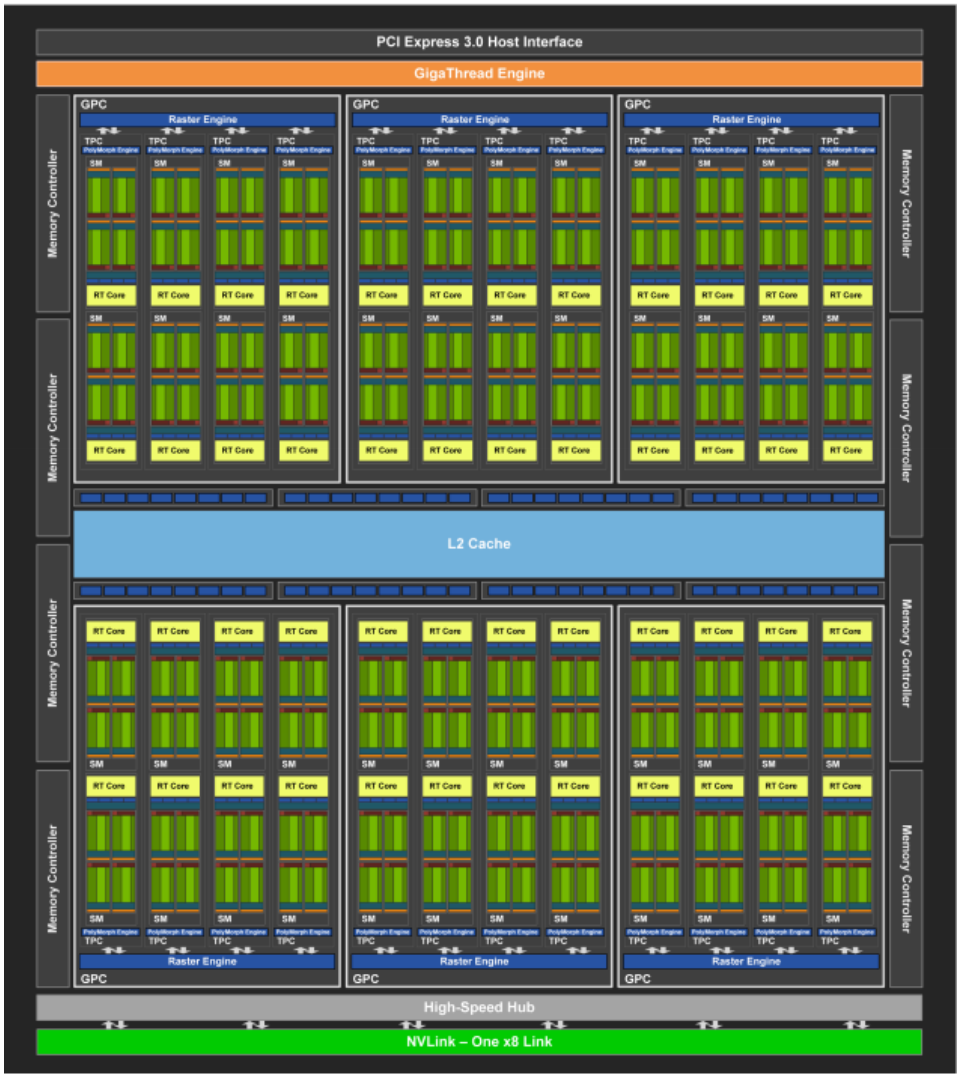
Deep Learning Super Samplingistä on olemassa kaksi versiota, DLSS 2X ja DLSS. Näissä kahdessa versiossa on kuitenkin hieman eroja. DLSS renderöi pelin todellisuudessa matalammalla resoluutiolla, kuin todellisuudessa. Renderöinnin jälkeen tekoäly hyödyntää tensor-yksiköitään ja käyttää opeteltuja perustotuuksia luodakseen lopullisen kuvan halutulla resoluutiolla. (Nvidia Corporation 2019[12].)

DLSS 2X renderöi kuvan halutulla resoluutiolla, mutta parantaa sitä käyttäen opittuja perustotuuksia. DLSS 2X:n reunanpehmennyksen laatu on erittäin korkea ja Nvidian mukaan sitä on lähes mahdotonta erottaa aidosta 64xSSAA-kuvasta. DLSS 2X ei tarjoa etua nopeudessa toisin kuin DLSS. Kuvassa 20 nähdään TAA -reunanpehmennyksen ja DLSS 2X:n erot. (Nvidia Corporation 2019[12].)



KUVA 20 TAA reunanpehmennyksen ja DLSS 2X vertailu (Nvidia Corporation 2019[12].)

Turing-arkkitehtuurin TU104-grafiikkapiiri koostuu kuudesta GPC:stä, 48 streaming multiprosessorista ja kahdeksasta 32 -bittisestä muistikontrollerista. (kuva 21.)



KUVA 21 TU104 -grafiikkapiirin rakenne (Nvidia Corporation 2019[12])

TU 104-grafiikkapiirissä jokainen GPC sisältää yhden rasteri-yksikön ja neljä TPC:tä Jokainen TPC sisältää yhden PolyMorph-moottorin ja kaksi streaming multiprosessoria.

Grafiikkapiirin jokainen streaming multiprosessori sisältää uuden RayTracing -ytimen. Jokainen streaming multiprosessori sisältää myös 64 CUDA-ydintä, 256 KB rekisteritietoja, 96 KB L1-välimuistia ja neljä tekstuuriyksikköä. Täydessä TU104-grafiikkapiirissä on 13.6 miljardia transistoria ja 3072 CUDA-ydintä, 368 tensor-ydintä ja 48 uutta RT-ydintä.

TU104-grafiikkapiiri tukee toisen generaation NVLinkkiä. Piirissä on yksi x8 NVLink, joka tukee 25 GB/s kaistanleveyden jokaiseen suuntaan. Uutta grafiikkapiiriä käytetään esimerkiksi GeForce RTX 2080, Tesla T4, ja Quadro RTX 5000

-korteissa. Taulukossa 2 vertaillaan vanhemman Pascal-arkkitehtuurin ja uuden Turing-arkkitehtuurin ominaisuuksia.

TAULUKKO 2 Pascal- ja Turing -arkkitehtuurien vertailu (Nvidia Corporation 2019[12])

GPU Features	GeForce GTX 1080	GeForce RTX 2080	Quadro P5000	Quadro RTX 5000
Architecture	Pascal	Turing	Pascal	Turing
GPCs	4	6	4	6
TPCs	20	23	20	24
SMs	20	46	20	48
CUDA Cores / SM	128	64	128	64
CUDA Cores / GPU	2560	2944	2560	3072
Tensor Cores / SM	NA	8	NA	8
Tensor Cores / GPU	NA	368	NA	384
RT Cores	NA	46	NA	48
GPU Base Clock MHz (Reference / Founders Edition)	1607 / 1607	1515 / 1515	1607	1620
GPU Boost Clock MHz (Reference / Founders Edition)	1733 / 1733	1710 / 1800	1733	1815
RTX-OPS (Tera-OPS) (Reference / Founders Edition)	8.9 / 8.9	57 / 60	NA	62
Rays Cast (Giga Rays/sec) (Reference / Founders Edition)	0.89	8 / 8	NA	8
Peak FP32 TFLOPS* (Reference / Founders Edition)	8.9	10 / 10.6	8.9	11.2
Peak INT32 TIPS* (Reference/Founders Edition)	NA	10 / 10.6	NA	11.2
Peak FP16 TFLOPS* (Reference / Founders Edition)	NA	20.1 / 21.2	NA	22.3
Peak FP16 Tensor TFLOPS with FP16 Accumulate* (Reference/Founders Edition)	NA	80.5 / 84.8	NA	89.2
Peak FP16 Tensor TFLOPS with FP32 Accumulate* (Reference/Founders Edition)	NA	40.3 / 42.4	NA	89.2
Peak INT8 Tensor TOPS* (Reference / Founders Edition)	NA	161.1 / 169.6	NA	178.4
Peak INT4 Tensor TOPS* (Reference / Founders Edition)	NA	322.2 / 339.1	NA	356.8
Frame Buffer Memory Size and Type	8192 MB GDDR5X	8192 MB GDDR6	16384 GDDR5X	16384 GDDR6
Memory Interface	256-bit	256-bit	256-bit	256-bit
Memory Clock (Data Rate)	10 Gbps	14 Gbps	9 Gbps	14 Gbps
Memory Bandwidth (GB/sec)	320	448	288	448
ROPs	64	64	64	64
Texture Units	160	184	160	192
Texel Fill-rate (Gigatexels/sec)	277.3 / 277.3	314.6 / 331.2	277	348
L2 Cache Size	2048 KB	4096 KB	2048 KB	4096 KB
Register File Size/SM	256 KB	256 KB	256 KB	256 KB

Nvidian uusin julkaisema Tesla T4 on ensimmäinen Turing-arkkitehtuurilla toimiva grafiikkasuoritin, joka on suunniteltu soveltamaan sovelluksia keskus- ja yrityslaitteissa.

Tesla T4:ssä käytetty TU104-grafiikkapiiri sisältää viisi GPC:tä, 20 TPC:tä, 40 streaming multiprosessoria. Grafiikkapiirissä on 2560 CUDA-ydintä ja 320 turing-yksikköä. Grafiikkapiirillä on 256-bittinen muistiliitäntä ja 320 GB/s kaistanleveys.

Turing-arkkitehtuurissa on myös inpainting-ominaisuus, joka osaa hyödyntää tekoälyä täydentääkseen puutteellisia kuvia. Esimerkiksi inpainting:iä voidaan käyttää poistamaan voimalinjoja maisemakuvasta. Inpainting ei ole uusi ominaisuus, mutta ennen tekniikka täytti kuvan, jostakin muusta kohdasta kuvan sisällä. (Nvidia Corporation 2019[13])

Uuden turing-arkkitehtuurin myötä inpainting on kehittynyt niin, että tekoäly oppii useista oikean elämän kuvista syntetisoidakseen uutta sisältöä täyttääkseen puutteet kuvassa. Kuvassa 22 nähdään, erilaisia esimerkkikuvia, joita on käytetty inpaint-teknologian demonstrointiin.

KUVA 22. Inpainting teknologia (Nvidia Corporation 2019[13])

Kuvassa 22 inpainting-teknologia on täyttänyt puutteelliset/pyyhityt osat kuvista käyttäen tekoälyänsä. Tekoäly skannannaa tuhansia erilaisia kuvia oikeasta maailmasta ja käyttää näistä saatua tietoa täyttääkseen puutteelliset kohdat.

4 CUDA-arkkitehtuuri

CUDA on Nvidian rinnakkaislaskennan arkkitehtuurin nimi. CUDA-teknologiaa käytetään kaikissa Nvidian grafiikkapiireissä. Nvidia tarjoaa työkalut, jotka mahdollistavat CUDA-arkkitehtuurin ohjelmoinnin. Nvidian tarjoama toolkit CUDA-arkkitehtuurille sisältää kääntäjä, profiloijan, virheiden jäljittäjän ja kaikki tarvittavat tiedot kehittäjille, jotka ottavat CUDA-arkkitehtuurin käyttöönsä. (Nvidia Corporation 2019[14].)

CUDA on suunniteltu toimimaan seuraavilla ohjelmointikielillä:

- C
- C++
- Fortran

Ohjelmointikielien monipuolisuus auttaa kehittäjien ohjelmointityötä. Ennen CUDAa käytettiin Direct3D:tä ja OpenGL:ää, jotka vaativat kehittyneitä taitoja grafiikkaohjelmoinnissa. CUDA tukee myös OpenACC ja OpenCL viitekehyksiä. (Nvidia Corporation 2019[11].)

CUDAssa on kolme keskeistä asiaa, jolla ydin toimii. Niihin kuuluvat hierarkia säieryhmien välillä, jaettu muisti ja estesynkronointi. Nämä kolme seikkaa tuodaan ohjelmoijalle pieninä ja yksinkertaisina laajennuksina. Tämä mahdollistaa helppokäyttöisen tiedon rinnakkaisuuden hyödyntämisen. Näiden avulla ongelmat voidaan pilkkoa pienemmiksi, joka helpottaa ongelmien ratkaisua. (Nvidia Corporation 2019[14].)

5 YHTEENVETO

Opinnäytetyön tavoitteena oli tutustua tarkemmin grafiikkasuorittimien valmistajaan Nvidiaan. Työssä perehdyttiin aluksi yleisesti näytönohjaimiin ja selitettiin näytönohjaimen perustarkoitus.

Opinnäytetyön pääaiheena käytiin lävitse Nvidian historiaa ja tuotteita. Työssä käytiin läpi Nvidian suunnittelemista arkkitehtuureista jokainen, joka on julkaistu vuoteen 2019 mennessä. Näihin arkkitehtuureihin lukeutui Fermi, Kepler, Maxwell, Pascal, Volta ja Turing.

Opinnäytetyön tavoitteet saavutettiin. Tutkimus Nvidiasta ja sen tuotteista antoi vahvan ymmärryksen, mitä grafiikkasuorittimet sisältävät ja kuinka Nvidian teknologia on kehittynyt vuosien varrella.

Nvidia antaa pelaajille, kuin myös ammattikäyttöön todella hyvät työkalut toteuttaakseen tarvittavat prosessit. Nvidian uusin Turing-arkkitehtuuri antaa 16.3 teraFLOPSin laskentatehon koneoppimiseen. RTX-teknologia antaa pelaajille myös aivan uudenlaisen kokemuksen peleissä hyödyntäen uutta säteenseuranta -teknologiaa.

Opinnäytetyötä tehdessä opin uusia ominaisuuksia, joita Nvidian grafiikkasuorittimet voivat toteuttaa. Työssä opin myös eri Nvidian grafiikkasuorittimien rakenteet ja kuinka suorittimet hyödyntävät eri komponentteja ja esimerkiksi CUDA-ydinten toiminnan ja DirectX-ohjelmointirajapinnan hyödyn.

LÄHTEET

Altman, A. 2016. The 4K Gaming Showdown: Geforce GTX 1070 SLI vs 1080 SLI vs Titan X Pascal. Luettu 14.2.2019.
https://techbuyersguru.com/4k-gaming-showdown-geforce-gtx-1070-sli-vs-1080-sli-vs-titan-x-pascal?page=3

Chiapetta, M. 2013. How to trick out your gaming PC with multiple graphics cards. Luettu 14.2.2019.
https://www.pcworld.com/article/2023630/how-to-trick-out-your-gaming-pc-with-multiple-graphics-cards.html

Hagedoorn, H. 2016. Multi-GPU Mode Explained. Luettu 14.2.2019.
https://www.guru3d.com/articles-pages/geforce-gtx-1070-2-way-sli-review,2.html

Hardwidge, B. 2009. Intel details first CPU with integrated GPU. Luettu 8.2.2019.
https://bit-tech.net/news/tech/cpus/intel-details-first-cpu-with-integrated-gpu/1/

Intel. 2011. Maximizing multicore processor performance. Luettu 14.2.2019.
https://www.intel.com/content/www/us/en/io/quickpath-technology/quickpath-technology-general.html

Kanter, D. 2011. Intel's Quick Path Evolved. Luettu 14.2.2019.
https://www.realworldtech.com/qpi-evolved/

Kurri, S. 2016. Testissä näytönohjaimet: NVIDIA GeForce GTX 1080 & 1070. Luettu 1.3.2019.
https://muropaketti.com/tietotekniikka/testissa-naytonohjaimet-nvidia-geforce-gtx-1080-1070-pascal/

Kurri, S. 2016. Pascal-arkkitehtuuri. Luettu 1.3.2019.
https://muropaketti.com/tietotekniikka/testissa-naytonohjaimet-nvidia-geforce-gtx-1080-1070-pascal/2/

Muropaketti. 2014. NVIDIA GeForce GTX 750 Ti. Luettu 27.2.2019.
https://muropaketti.com/tietotekniikka/nvidia-geforce-gtx-750-ti-maxwell/

Murphy, D. 2010. Intel's 2010 Clarkdale Desktop CPUs: What to Expect. Luettu 8.2019.
https://www.pcworld.com/article/185996/Intel_Clarkdale_Processors.html

Nvidia Corporation. 2019. DirectX 12. Luettu 25.03.2019.
https://www.nvidia.com/fi-fi/geforce/technologies/dx12/

Nvidia Corporation. 2019[1]. RTX 2080. Luettu 21.2.2019.
https://www.nvidia.com/fi-fi/geforce/graphics-cards/rtx-2080/

Nvidia Corporation. 2019[2]. GeForce GTX SLI HB Bridge. Luettu 14.2.2019. https://www.geforce.com/nvidia-sli-bridges

Nvidia Corporation. 2019[3]. GeForce Experience. Luettu 22.2.2019. https://www.nvidia.com/fi-fi/geforce/geforce-experience/

Nvidia Corporation. 2016[4]. Pascal architecture whitepaper. PDF-tiedosto. Luettu 1.3.2019. https://images.nvidia.com/content/pdf/tesla/whitepaper/pascal-architecture-whitepaper.pdf

Nvidia Corporation. 2019[5]. Self-Driving Cars. Luettu 22.2.2019. https://www.nvidia.com/en-us/self-driving-cars/drive-platform/

Nvidia Corporation. 2019[6]. Nvidia History. Luettu 22.2.2019. https://www.nvidia.com/en-us/about-nvidia/corporate-timeline/

Nvidia Corporation. 2009[7]. Fermi whitepapers. PDF-tiedosto. Luettu 24.2.2019. https://www.nvidia.com/content/PDF/fermi_white_papers/NVIDIA_Fermi_Compute_Architecture_Whitepaper.pdf

Nvidia Corporation. 2012[8]. GeForce GTX 680 whitepaper. PDF-tiedosto. Luettu 26.2.2019. https://www.nvidia.com/content/PDF/product-specifications/GeForce_GTX_680_Whitepaper_FINAL.pdf

Nvidia Corporation. 2014[9]. 5 Things you should know about the new Maxwell GPU architecture. Luettu 27.2.2019. https://devblogs.nvidia.com/5-things-you-should-know-about-new-maxwell-gpu-architecture/

Nvidia Corporation. 2014[10]. GeForce GTX 750 Ti whitepaper. PDF-tiedosto. Luettu 27.2.2019. https://international.download.nvidia.com/geforce-com/international/pdfs/GeForce-GTX-750-Ti-Whitepaper.pdf

Nvidia Corporation. 2017[11]. Volta architecture whitepaper. PDF-tiedosto. Luettu 3.3.2019. https://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf

Nvidia Corporation. 2018[12]. Nvidia Turing architecture whitepaper. PDF-tiedosto. Luettu 6.3.2019 https://www.nvidia.com/content/dam/en-zz/Solutions/design-visualization/technologies/turing-architecture/NVIDIA-Turing-Architecture-Whitepaper.pdf

Nvidia Corporation. 2019[13]. Nvidia Turing. Luettu 6.3.2019. https://www.nvidia.com/en-us/design-visualization/technologies/turing-architecture/?nvid=nv-int-sh28-52949

Nvidia Corporation. 2019[14]. CUDA Zone. Luettu 6.3.2019 https://developer.nvidia.com/cuda-zone

Suvanto, V. 2009. NVIDIA esitteli Fermi-näytönohjainarkkitehtuurin. Luettu 24.2.2019.
https://muropaketti.com/tietotekniikka/tietotekniikkauutiset/nvidia-esitteli-fermi-naytonohjainarkkitehtuurin/

Verma, A. 2018. Top Graphics Card Manufacturers & Brands for Nvidia & AMD GPUs. Luettu 8.2.2019.
https://graphicscardhub.com/graphics-card-manufacturers-brands/