Fehér Krisztián

Hogyan írj extrém gyors programot?

- Bevezetés a CUDA programozásba -

BBS-INFO Kiadó, 2019.

A könyv nagyobb mennyiségben megrendelhető a kiadónál:
BBS-INFO Kiadó, www.bbs.hu Tel.: 407-17-07

ISBN 978-615-5477-78-2
E-book ISBN 978-615-5477-79-9

Kiadja a BBS-INFO Kft. Budapest
Felelős kiadó: a BBS-INFO Kft. ügyvezetője
Nyomdai munkák: Biró Family Nyomda
Felelős vezető: Biró Krisztián

TARTALOMJEGYZÉK

1. Előszó

Izgalmas technológiai áttörésekről olvashatunk nap mint nap, melyek mesterséges intelligenciáról, önvezető autókról és hasonló technológiákról szólnak. A számos úttörő megoldás jelentős hányadának hátterében az NVIDIA vállalat CUDA platformja áll, mely egyedülálló hardveres és szoftveres megoldásaival forradalmi lehetőségeket kínál a jelen és a jövő kutatóinak és a technológia hétköznapi felhasználóinak is.

Ez a könyv a CUDA technológia programozásának alapjairól szól.

1.1. A téma aktualitása

Az emberiség által kitermelt digitális adatmennyiség napjainkra gigászi méreteket ölt. A tárolás mellett most már az adatok elemzése, feldolgozása is egyre inkább előtérbe kerül, melynek igényét a mesterséges intelligenciával és a különböző autonóm rendszerekkel kapcsolatos kutatások csak tovább fokozzák.

Kb. 10 évvel ezelőttig érvényben volt az ún. Moore törvény, mely kimondta, hogy a számítógépek számítási kapacitása évente megduplázódik. Ezt a szabályt még az iskolákban is tanították, annyira szemléletes volt. A CPU-k fejlődése kb. 10 évvel ezelőtt azonban megtorpant és azóta lényegében stagnál. Minimális előrelépések történnek csak, mivel a jelenlegi architektúrák és a fizika törvényei már nem teszik lehetővé a számítási teljesítmény további növelését. Moore törvénye nem érvényes többé!

Több kutatás is folyik alternatív, gyökeresen új felépítésű architektúrák kifejlesztése érdekében, de ezek belátható időn belül nem lesznek elérhetők, pláne bárki által programozhatóak. Megoldásokra márpedig szükség van már most is. Így terelődik a figyelem egyre inkább a videokártyák párhuzamos számítási képességeinek általános célokra történő kihasználására.

Elérkezett a GPU-k korszaka, melyet angol szóval 'GPU computing', vagy 'High Performance Computing' elnevezésekkel is szoktak illetni. Szintén elterjedt kifejezés a GPGPU is (General-Purpose Com-

puting on Graphics Processing Units). A GPU-k a CPU-kkal szemben eleve párhuzamosított, aszinkron adatfeldolgozásra lettek kitalálva. A jelen GPU-k a párhuzamosított adatfeldolgozás terén messze túlszárnyalják a CPU-kat, akár 10 -, vagy 100 szorosan is, nagyon alacsony fogyasztás és nem utolsó sorban alacsony ár mellett. Ezek olyan fegyvertények, amelyek mellett az informatika iránt érdeklődők nem mehetnek el csukott szemmel.

Természetesen a CPU-k sosem fognak eltűnni, de ahol kiugróan nagy számítási teljesítményre lesz szükség, ott a GPU-k használata egyre erőteljesebb lesz. A nagyteljesítményű adatfeldolgozás köré felépülő technológiák a számítástechnika új, napjainkban zajló forradalmát képviselik, ezért mindenképpen ajánlott a folyamatot megismerni: miről szól ez az egész, milyen elvek szerint működik, hol lehet bekapcsolódni ebbe, stb.

1.2. A könyv felépítése

A könyv egyedülálló módon egyszerre ismeretterjesztő kiadvány és programozási segédlet is egyben.

A könyv első felében a CUDA technológia történetéről, jelenlegi felhasználási területeiről szólunk.

A könyv második fele a technológiát a gyakorlatban is felhasználni kívánó, a programozás iránt érdeklődő olvasóknak szól.

A függelék további hasznos témaköröket tárgyal.

Minden, a könyvben bemutatott kód korlátozás nélkül felhasználható, bármilyen célra. A fő forráskódok letölthetőek a kiadó weboldaláról is.

A könyv egyes algoritmusainak kidolgozásához a szerző sok inspirációt merített Dmitry V. Sokolov tinyrenderer publikus raszterizációs algoritmusaiból: https://github.com/ssloy/tinyrenderer.

1.3. Kinek szól a könyv?

Az elsődleges célközönség a C programozásban jártas olvasó, aki rendelkezik tapasztalattal Windows alkalmazások fejlesztésében.

A WIN32 API, a Direct2D és a Visual Studio mint fejlesztőeszköz alapszintű ismerete elengedhetetlen. Ha ezek a témák ismeretlenül csengenek, a szerzőnek az ajánlott irodalomban található könyvei segítségével pótolhatóak ezen alapismeretek.

1.4. A szerzőről

A szerző hivatásos szoftvertesztelő, minőségbiztosítási tanácsadó, diplomás német irodalmár, a Magyar Térinformatikai Társaság (HUNAGI) egyéni szakértői tagja, független kutató.

Elsődleges szakterülete a digitális grafika programozása, digitális térképalkalmazások készítése és a párhuzamos programozási technikák felhasználási területei.

Kibervédelmi kérdéseket évek óta tudatosan tanulmányoz, több kiberbiztonsággal kapcsolatos előadást is tartott már Magyarországon, sőt vendégoktatóként még oktatási intézményben is. Szakmai munkáját évről évre növekvő érdeklődés és elismerés kíséri.

Gyerekkorában autodidakta módon tanult meg programozni, az évek során számos programozási nyelvvel megismerkedett. Megszerzett tudását előszeretettel használja alternatív, kísérleti alkalmazások készítésére, melyek egy része ingyenesen elérhető, sőt vannak köztük nyílt forráskódúak is. A szerző fejleszt Windows desktop, Android és webes környezetekre is.

Tudását igyekszik minél szélesebb körben megosztani másokkal is. Ennek folyományaként több könyve is megjelent már a hazai könyvesboltokban az elmúlt években, nem egy közülük sikerlisták élére is került.

2. A CUDA madártávlatból

2.1. Mi az a CUDA?

A **CUDA** a **Compute Unified Device Architecture** rövidítése, mely magyarul kb. annyit tesz: **Egységesített Számítási Eszközarchitektúra**.

Az NVIDIA a CUDA-t napjainkban **számítási platformként (CUDA-X)** definiálja, ami jól megragadja a technológia lényegét is.

A CUDA első megjelenése 2006 körülre datálható, a GeForce 8800-as kártya megjelenésére, mely a világon elsőként kínált általános célú számítási feldolgozóegységeket, közismertebb nevükön **CUDA magokat**.

Az azóta eltelt időszakban már több mint 500 millió, CUDA technológiát alkalmazó grafikus kártyát adtak el, ami példátlan siker.

A CUDA segítségével általános feladatokat gyorsíthatunk fel, CPU-alapú megközelítésekhez képest jelentős mértékben.

2.2. CUDA generációk

Az NVIDIA a CUDA architektúráit hagyományosan valamilyen híres tudósról nevezi el.

A legfontosabb architektúrák elnevezése, időben:

- Fermi
- Kepler
- Maxwell
- Pascal
- Volta
- Turing.

Egy-egy új architektúrát jellemzően pár éves időközökben hoznak ki. A Volta részben kivételt képez, ugyanis ebből az architektúrából csupán néhány felső kategóriás gyorsítókártya került ki. Az összes többi esetén azonban GeForce és Quadro kártyák is rendre készültek.

2.3. Mire használják a CUDA-t?

Tekintsük át a legfontosabb szakterületeket, melyeken a CUDA domináns szerepet tölt be!

2.3.1. Mesterséges intelligencia

Az MI kutatások sikere belátható időn belül az adatfeldolgozás gyorsaságán áll vagy bukik. Jelenleg az ún. **mély tanulásos** (angolul: **deep learning**) alapú MI kutatások dominálnak, melyek keretében MI-kezdeményeket tanítanak különböző viselkedésekre, rengeteg mintaadat feldolgozása útján. Éppen ezért belátható, hogy a CUDA az MI kutatások egyik jelentős hajtómotorja.

2.3.2. Adatközpontok

A modern adatközpontok felépítése és üzemeltetése a folyamatosan növekvő igények miatt egyre nagyobb kihívást jelent. A kis helyigény, nagyfokú integráltság és az alacsony energiafelhasználás döntő tényezők szoktak lenni. Ezen a területen az NVIDIA Tesla gyorsítókártyái például kimagaslóan jók.

2.3.3. GPU felhő

Óriási az igény a változtatható számítási kapacitások bérleti alapon történő felhasználására. Ilyet számos vállalat kínál, például az Amazon is. Segítségével alacsony költségek árán lehet óriási számítási teljesítményeket időszakosan kibérelni.

A másik terület a streaming alapú online videojátékipar, mely a játékosoknak alacsony költségen kínál hozzáférést nagyteljesítményű videokártyákhoz.

2.3.4. Dizájn és vizualizáció, fizikai szimulációk

A különböző tervezési feladatoknál nagyon fontos tényező az idő, több szempontból is. Minél gyorsabban előállítható egy-egy terv végleges változata, annál értékesebb egy technológia.

Felbecsülhetetlen értékű például egy olyan technológia, aminek a segítségével például valós időben lehet látni egy még csak tervezési fázisban levő gépjármű megjelenését, vagy akár viselkedését is.