GPGPU – Új eszközök az információbiztonság területén

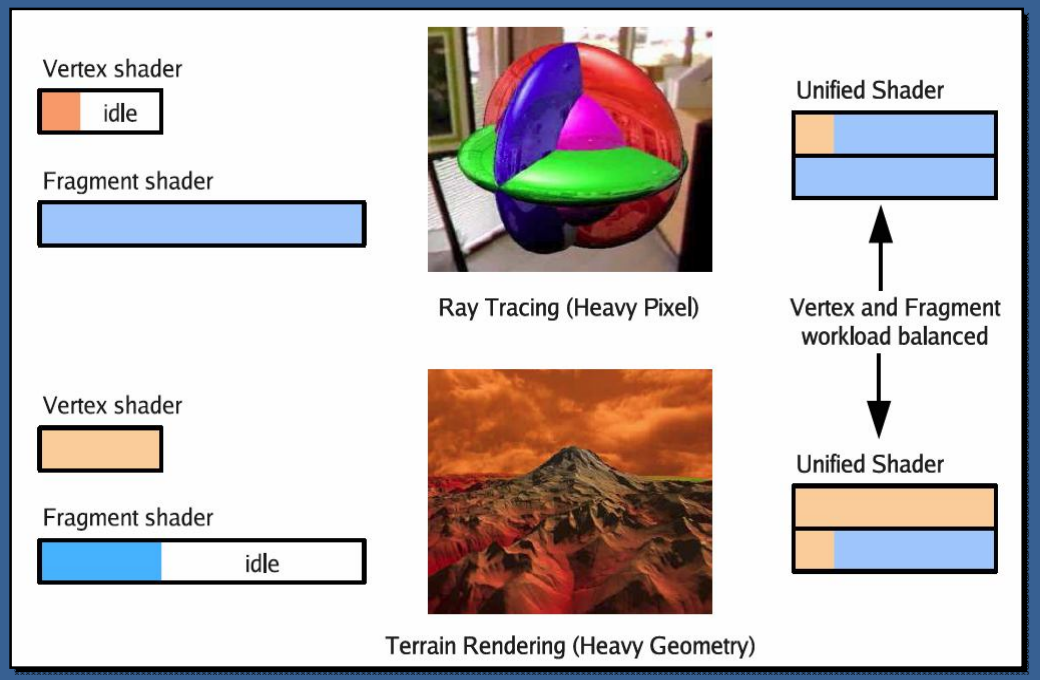
Szénási Sándor

ZMNE doktorandusz

szenasi.sandor@nik.bmf.hu

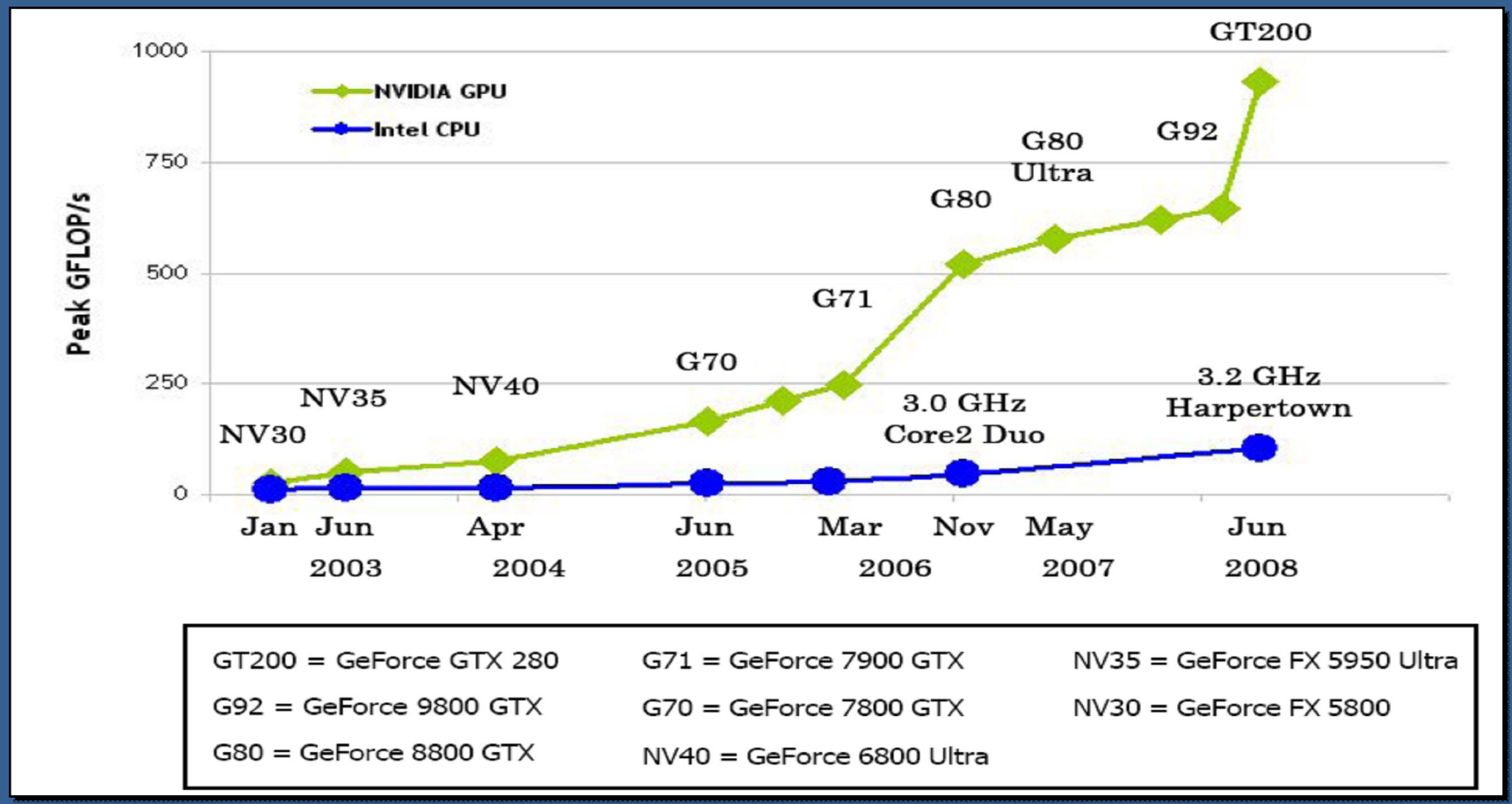
Unified shader model

- Grafikus kártyák 3D gyorsítási lehetőségekkel
- Különböző egységek végzik a különféle feladatokat:
  - pixel shader
  - vertex shader
  - geometry shader
- Egyenetlen terhelés
- Unified shader model kialakítása



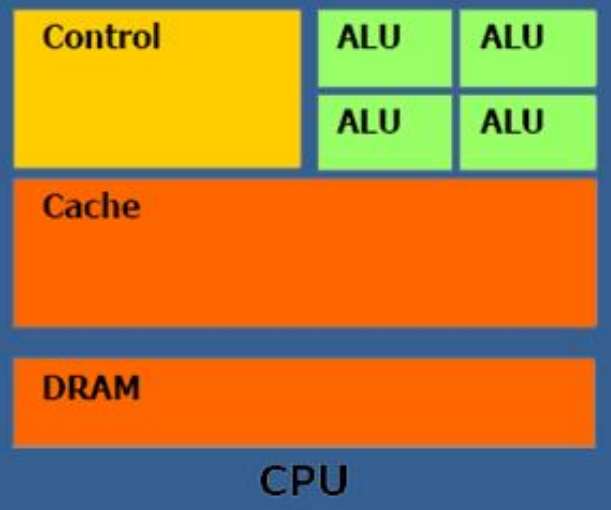
GPGPU megjelenése

- GPGPU - Új lehetőségek a HPC területen

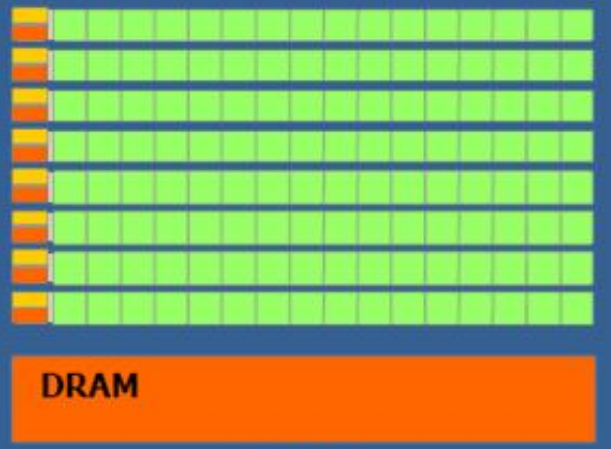


Az új technológia jellemzői

- Az eszközök előnyei
  - Nagyszámú végrehajtó egység
  - SIMT végrehajtás
  - Hardver alapú szálütemezés
  - Összetett memória hierarchia
- Az eszközök hátrányai
  - Adatpárhuzamos végrehajtás
  - Külön memóriatartomány
  - Fejlesztési nehézségek



CPU



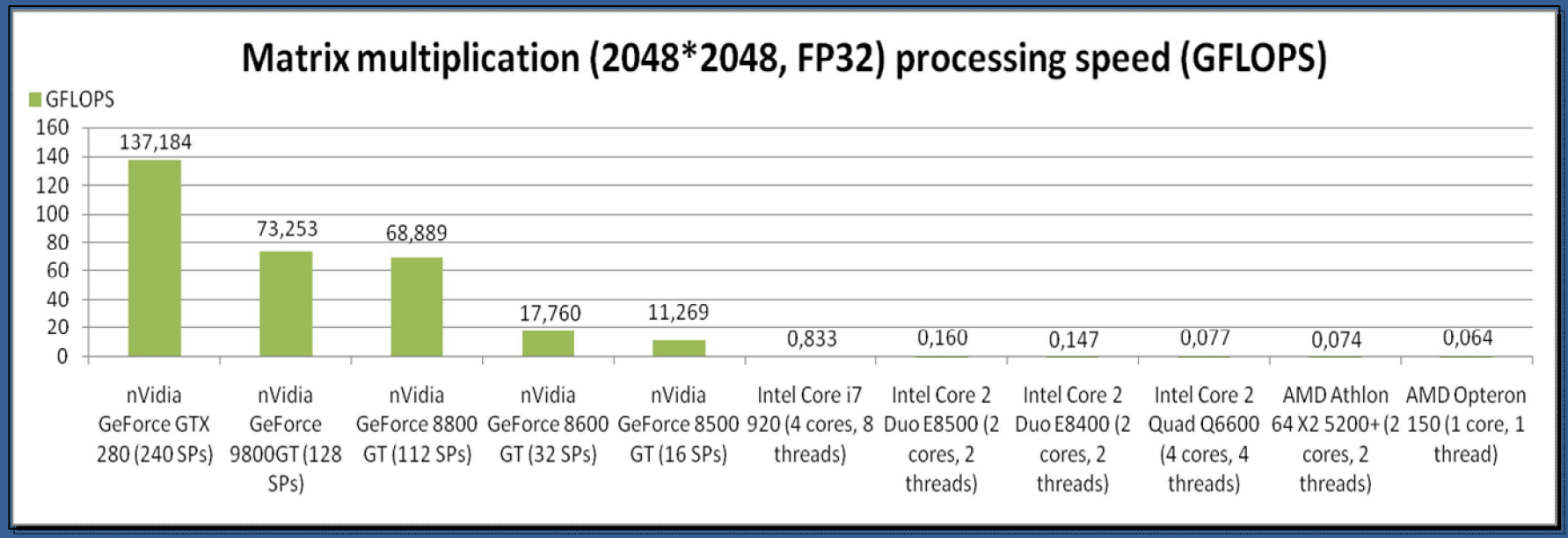
GPU

NVIDIA 280GTX

Mag	GT200
Megjelenés ideje	6/08
IC technológia	65 nm
Tranzisztorok száma	1400 mtrs
Lap méret	576 mm^2
Mag frekvencia	602 MHz
Feldolgozó egységek száma	240
Shader frekvencia	1.296 GHz
FP32 utasítás/órajel	3
Csúcs FP32 teljesítmény	933 GLOPS
Csúcs FP64 teljesítmény	77/76 GLOPS
Mem. órajel	2214 Mhz
Mem. interfész	512-bit
Mem. sávszélesség	141.7 GB/s
Mem. méret	1.0 GB
Mem. típus	GDDR3
Mem. csatorna	8*64-bit
Összekapcsolási lehetőség	SLI
Interfész	PCIe 2.0x16
MS Direct X	10.1 subset

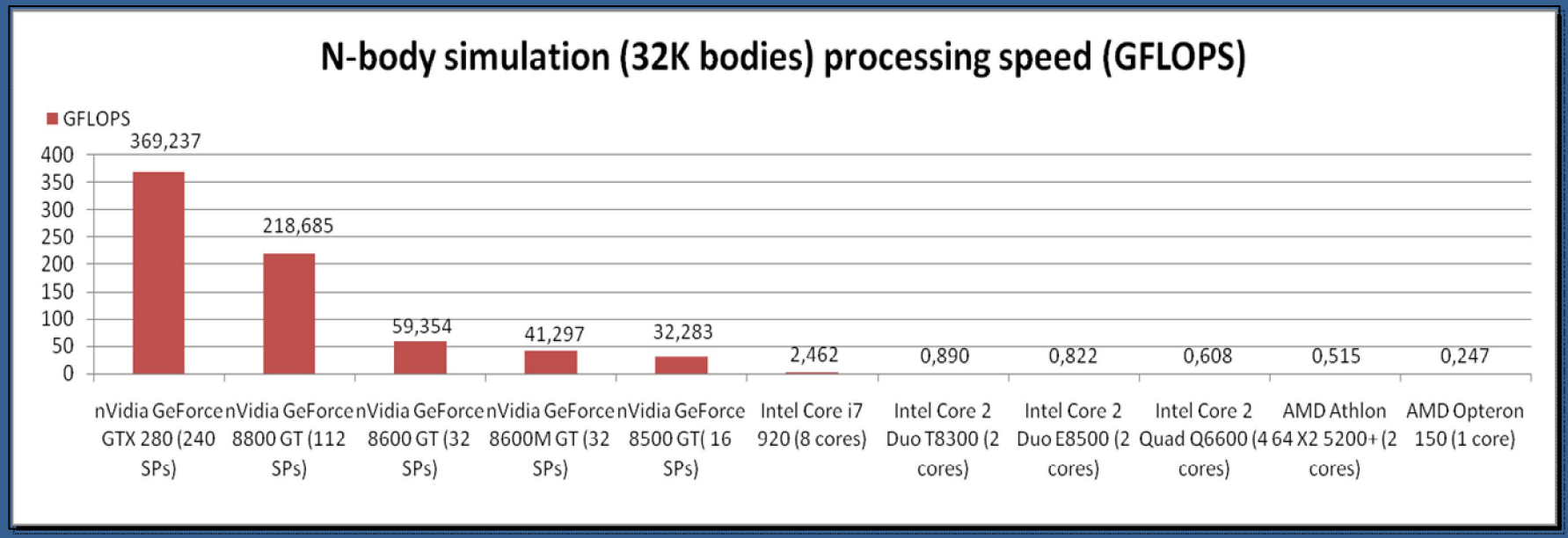
Mátrix szorzás vizsgálata

- Számításigény $O(n^3)$ / Memória igény $O(n^2)$
- A magok számával szinte lineárisan arányosan növekszik az elérhető teljesítmény (jól párhuzamosítható feladat)
- A memória átvitel azonban meglehetősen sok időt igényel, ez jelentősen rontja a GPGPU eredményeket



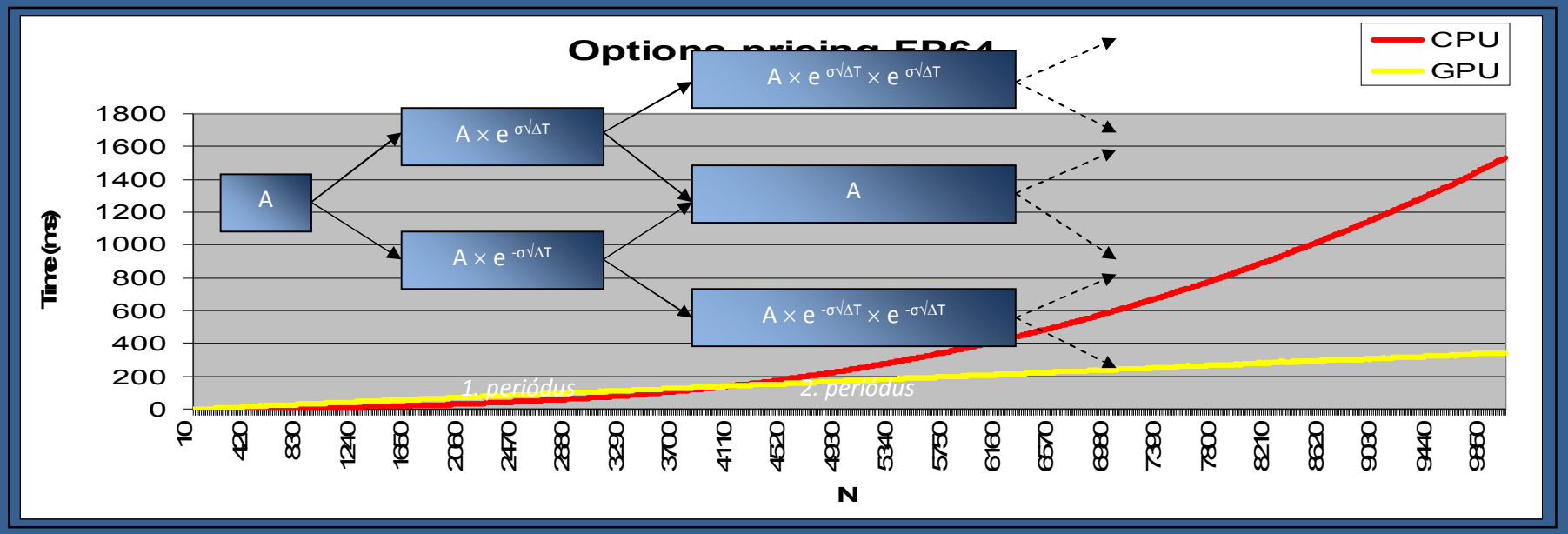
N-test szimuláció vizsgálata

- Számításigény $O(n^2)$ / Memória igény $O(n)$
- A feladat kimondottan számításigényes, hiszen minden test esetén figyelembe kell venni az összes többi test hatását
- A testek adatai azonban meglehetősen kis memóriaterületet foglalnak el, így az átvitel sem jelentős



Binomiális fa feldolgozása

- Számításigény $O(n^2)$ / Memória igény $O(1)$
- A memóriaátvitel minimális, mivel csak a fa felépítéséhez szükséges adatokat, illetve a végeredményt kell mozgatni
- A fa egy szintje párhuzamosan feldolgozható, a szintek között azonban szinkronizációra van szükség



Összefoglalás

- Az eszközök programozása egyelőre nehézkes
  - Új szoftver/hardver környezet
  - Új programozási paradigma
- Optimalizáció fontos
  - Különböző hardver felépítések
  - Fordítóprogramok kezdetlegesek
- Folyamatos fejlesztések
  - Fizetőképes kereslet rendelkezésre áll
  - Új generációs videokártyák (Fermi/Radeon HD 5800)

GPGPU vonal előtérbe helyezése

- A grafikus kártya gyártók tudatosan erősítik a GPGPU lehetőségeket termékeikben
- Példaképp az NVIDIA Fermi fejlesztései:
  - Több végrehajtó egység/memória
  - Dupla pontosságú aritmetika gyorsítása
  - ECC támogatás
  - Konfigurálható cache hierarchia kialakítása
  - Gyorsabb atomi műveletek
  - Párhuzamosan több kernel futtatása

Felhasznált irodalom

[1]: Patidar S. & al., “Exploiting the Shader Model 4.0 Architecture, Center for Visual Information Technology, IIIT Hyderabad, http://research.iiit.ac.in/~shiben/docs/SM4_Skp-Shiben-Jag-PJN_draft.pdf [2009.11.24.]

[2]: Nvidia CUDA Compute Unified Device Architecture Programming Guide, Version 2.0, June 2008, Nvidia

[3]: S. Dezső: GPUs/Data Parallel Accelerators, http://nik.bmf.hu/sima/200809springnotesMPPBSc.htm [2009.11.24.]

[4]: Cs. Csillingh, Cs. Iványi, S. Szénási: Multiplatform Performance Analysis of Parallel Workloads on Graphics Hardware, Morgan Stanley-BME Financial Innovation Centre Kick-off & Workshop, 2009

[5]: NVIDIA CUDA Programming Guide v2.0 http://www.nvidia.com/cuda [2009.11.24.]

Köszönöm a figyelmet!