Tárgyleírás

Szak megnevezése: Programtervező informatikus BSc és MSc

Oktatás nyelve: magyar

Tárgy neve: GPU programozás/GPU Programming

Tárgyfelelős neve: Dr. Hajder Levente

Tárgyfelelős tudományos fokozata: habilitált doktor

Tárgyfelelős MAB szerinti akkreditációs státusza: AT

Az oktatás célja:

tudás:

A kurzus elmélyíti és kibővíti a hallgatók elméleti és gyakorlati ismereteit a párhuzamos programozás és a GPU architektúrák területén. Az előadások során a hallgatók megismerkednek a modern grafikus processzorok felépítésével, a párhuzamos számítási modellekkel, valamint az OpenCL és CUDA API-k működésével és elméleti hátterével. Külön hangsúlyt kapnak a párhuzamos tervezési minták, a párhuzamosíthatóság vizsgálata, valamint a blokk- és adatpárhuzamos programozás alapelvei. Külön hangsúlyt fektetünk a kurrens grafikus hardverek ISA assembly szintű vizsgálatára. Emellett a hallgatók elsajátítják a robusztus numerikus számítások megvalósításához elengedhetetlen IEEE-754/ISO 60559 lebegőpontos szabvány releváns részleteit, illetve betekintést kapnak a hibamentes transzformációk (EFT) témájába.

A gyakorlatokon a hallgatók valós GPU programozási feladatok megoldásán keresztül sajátítják el az OpenCL és CUDA használatát. A feladatok során a párhuzamos tervezési minták optimális implementációjára, matematika és egyéb területek (jelfeldolgozás, gépi látás, számítógépes grafika stb.) feladatainak párhuzamos megoldására, valamint a fejlett GPU képességek (memóriamodellek, szinkronizáció, optimalizációs technikák) bevonására helyeződik a hangsúly.

képesség:

A kurzus fejleszti az alábbi képességeket:

- párhuzamos algoritmikus gondolkodás, párhuzamos programtervezési minták felismerése és implementálása;
- numerikusan robosztus algoritmusok implementációja a hardverben megvalósított lebegőpontos számábrázolásra tekintettel;
- Algoritmusimplementációk kerekítési hibájának szisztematikus vizsgálata;
- a GPU architektúrák működésének megértése, OpenCL és CUDA programozási modellek alkalmazása;
- teljesítményoptimalizálási stratégiák megvalósítása, memóriahozzáférés és szinkronizáció hatékony kezelése;
- nagy számításigényű feladatok GPU-n történő implementációja és skálázhatóságának értékelése.

Ezek a képességek különösen értékesek a nagy teljesítményű számítástechnika (HPC), a mesterséges intelligencia, a számítógépes grafika, a jelfeldolgozás és más számításigényes iparágak szakemberei számára.

attitűd:

- Vállalja és hitelesen képviseli informatikai szakterülete szakmai alapelveit.
- Nyitott a képesítésével, szakterületével kapcsolatos szakmai, technológiai fejlődés és innováció megismerésére és befogadására.
- Reflektív módon tekint saját szakmai kompetenciáira és tevékenységére.
- Törekszik a folyamatos szakmai képzésre és általános önképzésre.
- Elfogadja az informatikai szakma munka- és szervezeti kultúra szabályait, etikai elveit.

autonómia, felelősség:

- Felelősséget vállal szakmai tevékenységéért.
- Törekszik a hatékony és minőségi munkavégzésre.
- Felelősséggel vállalja részfeladatok megoldását komplex, párhuzamos szoftverfejlesztési feladatok megoldásában.

Az oktatás tartalma:

- Párhuzamos architektúrák áttekintése: CPU és GPU összehasonlítása, SIMD és SIMT modellek.
- Modern GPU architektúrák felépítése: streaming multiprocesszorok, memóriahierarchia, szálblokkok és rácsok.
- Gépi számábrázolás és az IEEE 754/ISO 60599 lebegőpontos szabvány
- Algoritmusok kerekítési hibaanalízise és hibamentes transzformációk (EFT) alkalmazása
- Alacsonyszintű GPU architektúrák konkrét modelleken és GPU-k ISA assembly szintű karakterizációja
- OpenCL és CUDA programozási modellek elmélete és alapfogalmai.
- Szálkezelés és szinkronizáció: globális, lokális és privát memóriák; memória-hozzáférés optimalizáció.
- Párhuzamos tervezési minták: adatpárhuzamos és blokkpárhuzamos feladatmegoldás, redukciók, prefix összegek, szórás-gyűjtés minták.
- Párhuzamos algoritmusok GPU-s implementációja: mátrixműveletek, konvolúció, keresési és rendezési algoritmusok.
- Optimalizációs stratégiák: memória-koaleszálás, bankütközések elkerülése, divergens vezérlési szerkezetek kezelése.
- Többszálú programok hibakeresése és teljesítményprofilozás GPU környezetben.
- Fejlett GPU képességek: megosztott memória hatékony használata, konstans memória, textúramemória.
- Heterogén programozási környezetek: CPU–GPU együttműködés, aszinkron adatátvitel, több-GPU-s futtatás.
- Párhuzamos algoritmusok alkalmazási területei: gépi tanulás, képfeldolgozás, tudományos szimulációk, valós idejű rendszerek.

A számonkérés és értékelés rendszere:
A hallgatók a kurzust kombinált vizsgával teljesíthetik, A végső jegy megszerzéséhez félévközi beadandó feladatokat (gyakorlati értékelés) kell teljesíteniük.

Irodalom: [csak magyar nyelvű képzés tárgya esetén töltendő]

- Randal E. Bryant and David R. O'Hallaron - Computer Systems: A Programmer's Perspective (ISBN 013409266X)

- Jean-Michel Muller et al. - Handbook of Floating-Point Arithmetic (ISBN 3319765256)
- NVIDIA, NVIDIA CUDA Toolkit documentation, http://docs.nvidia.com/cuda/
- The open standard for parallel programming of heterogeneous systems. https://www.khronos.org/opencl/

Az oktatás célja angolul / Aim of the subject:

Knowledge

Through this course, students enhance their theoretical and practical expertise in parallel programming and GPU architectures. The lectures introduce students to GPU architectures, parallel computing models, and the theory and operation of OpenCL and CUDA APIs. Special emphasis is placed on parallel design patterns, the examination of parallelizability, and the principles of block and data-parallel programming. We place particular emphasis on the ISA assembly-level examination of current graphics hardware. In addition, students learn the relevant details of the IEEE-754 / ISO 60559 floating-point standard that are essential for implementing robust numerical computations, and they are introduced to the topic of Error-Free Transformations (EFT). During the practice sessions, students will gain knowledge of the use of OpenCL and CUDA through solving GPU programming tasks. The focus will be on the optimal implementation of parallel design patterns, the parallel solution of tasks from mathematics and other areas (signal processing, machine vision, computer graphics, etc.), and the incorporation of advanced GPU capabilities (memory models, synchronization, optimization techniques).

Abilities:

The following skills are developed by the course:

- parallel algorithmic thinking, recognition and implementation of parallel program design patterns;
- design and implementation of numerically robust algorithms that account for the properties and limitations of hardware-implemented floating-point arithmetic.
- the ability to derive rounding error bounds of algorithms;
- understanding the operation of GPU architectures and applying the OpenCL and CUDA programming models;
- implementing performance optimization strategies, efficient management of memory access and synchronization;
- implementing computationally intensive tasks on GPUs and evaluating their scalability.

These skills are particularly valuable for professionals in high-performance computing (HPC), artificial intelligence, computer graphics, signal processing, and other computation-intensive industries.

Attitude:

- Commitment and authentic representation of the professional principles of the IT profession.
- Open to learn about and embrace professional and technological developments and innovations in his/her field of competence and specialization.
- Reflective about his/her own professional competences and activities.
- Pursues continuing professional development and general self-learning.
- Accepts the rules of work and organizational culture and ethical principles of the IT profession.

Autonomy, responsibility:

- Takes responsibility for his/her professional activities.
- Strives to work efficiently and with quality.
- Assumes responsibility for sub-tasks in complex and parallel software development tasks.

Az oktatás tartalma angolul / Major topics:

- Overview of Parallel Architectures: comparison of CPUs and GPUs; SIMD and SIMT models.
- Modern GPU Architectures: structure of streaming multiprocessors, memory hierarchy, thread blocks and grids.
- OpenCL and CUDA Programming Models: theory and fundamental concepts.
- Thread Management and Synchronization: global, local, and private memory; optimization of memory access.
- Number representation and the IEEE 754/ISO 60559 floating-point standard. Introduction to machine number formats, precision and range, and the essential aspects of the IEEE 754 / ISO 60559 standard for floating-point arithmetic.
- Rounding-error analysis of algorithms and the use of error-free transformations (EFT). Techniques for analyzing rounding errors in numerical algorithms, and the application of EFTs to improve robustness and reproducibility.
- Low-level GPU architectures and ISA-level characterization. Study of specific GPU hardware models, including instruction set architecture (ISA)–level characterization and analysis for understanding performance and numerical behavior.
- Parallel Design Patterns: data-parallel and block-parallel solutions, reductions, prefix sums, scatter–gather patterns.
- Implementation of Parallel Algorithms on GPUs: matrix operations, convolution, search and sorting algorithms.
- Optimization Strategies: memory coalescing, avoiding bank conflicts, handling divergent control flows.
- Debugging and Performance Profiling of Multithreaded Programs in GPU Environments.
- Advanced GPU Capabilities: efficient use of shared memory, constant memory, texture memory.
- Heterogeneous Programming Environments: CPU–GPU cooperation, asynchronous data transfer, multi-GPU execution.
- Application Areas of Parallel Algorithms: machine learning, image processing, scientific simulations, real-time systems.

A számonkérés és értékelés rendszere angolul / Requirements and evaluation:
Students can complete the course by combined examination. To get the final grade, they complete mid-term assignments (practical assessment).

Irodalom / Literature: [csak angol nyelvű képzés tárgya esetén töltendő / in case the language of the subject is ONLY English]