Indian Institute of Technology Kanpur

Proposal for a New Course

Course No: CSXXX

Course Title: GPU Architecture and Programming

Credits: 3-0-0-0-[9] (Lectures 3)

Proposing Department/IDP: Computer Science and Engineering

Other Departments which may be interested in the proposed course:
Department of Electrical Engineering, Department of Intelligent Systems

Proposing Instructors:
Mainak Chaudhuri (CSE) Swarnendu Biswas (CSE) Debadatta Mishra (CSE)

Prerequisite: Expertise in programming using C/C++. UG level background on Computer Organization and Architecture, Operating Systems and Compilers. Exposure to low-level programming and parallel computing is desirable.

Who can take the course: PhD, Masters, and 4th year UG Students

Course Description

Objective: Usage of accelerators such as GPGPUs see a constant upsurge and is expected to grow in future. Especially, the applications in the domain of AI/ML make heavy use of parallel computing platforms such as GPUs. The objective of this course is to delve into the design and implementation subtleties of different layers with a system design lens. Further, this course will explore the interaction of different layers such as the underlying hardware architecture, compiler and run-time system, and the operating system.

Detailed breakdown of lecture hours for different topics is provided in the following table.

Module	**Topic**	**No. of lectures**
Introduction and Background	Multi-core CPUs and Co-processors, Special purpose Accelerators and General purpose accelerators. Discrete vs. Shared Memory Accelerators, Shared memory computing systems, Introduction to multi-threading and parallel computing.	**5**
Applications	Representative applications from the HPC and ML domain, Computing and memory access characteristics, Interaction with I/O and other sub-systems	**5**
Hardware Architecture	Background on super scalar architecture, SIMD architecture, SIMT architecture and GPUs, Warp scheduling, Memory hierarchy and Address translation	**10**
Compiler and Run-time	Introduction to CUDA and OpenCL, SIMD ISA, PTX assembly, Static and dynamic instrumentation techniques for CUDA kernels, GPU execution profilers, Program optimization techniques: programmer hints and transparent techniques	**10**
CPU-I/O interface and memory addressing	Overview of I/O subsystem, PCI interfacing, DMA, IOMMU, Peer-to-peer DMA, Interrupt handling, Virtual memory sub-system, Unified virtual memory and GPU drivers	**10**
Total lectures		**40**

Reference Books:

1. J. L. Hennessy and D. A. Patterson. Computer Architecture: A Quantitative Approach. Morgan Kaufmann/Elsevier-India.
2. D. E. Culler and J. P. Singh with A. Gupta. Parallel Computer Architecture: A Hardware/Software Approach. Morgan-Kaufmann publishers.
3. David B. Kirk and Wen-mei W. Hwu. Programming Massively Parallel Processors: A Hands-on Approach
4. Remzi H. Arpaci-Dusseau and Andrea C. Arpaci-Dusseau, *Operating Systems: Three Easy Pieces*. Online 2018.
5. Daniel P. Bovet, Marco Cesati, *Understanding the Linux Kernel - from I/O ports to process management* (Third ed.), O'Reilly 2005.

Proposer(s): Date:

DPGC Convener

Chairman SPGC

DOAA