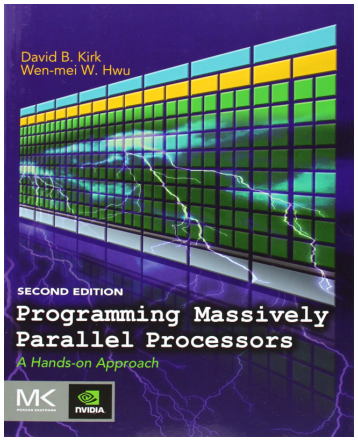
EE 8823: GPU Architectures

3-0-3 (2S,1D) Prerequisite: EE 6100, CS 6290 or equivalent



The last decade has seen the emergence of general-purpose graphics processing units (GPUs) as vehicles for accelerating general purpose scientific, enterprise, and embedded applications. This emergence has coincided with the explosive growth of data parallel applications and the ascendance of energy efficiency as a driver of performance scalability. The research community has evolved a body of compiler and microarchitecture knowledge to address important bottlenecks to harnessing the enormous throughput and memory bandwidth of modern GPUs. This course first provides an in-depth coverage of important microarchitecture concepts and performance optimizations that have now become accepted in this research and product community. This is followed by coverage of more recent research advances in the performance and power optimization of GPUs.

GPUs are now seeing increasing computation from other models such as Systolic and Dataflow. The course concludes with an exposition of the key elements of these models in contrast t

Class Materials:

- D. Kirk, and W. Hwu, "Programming Massively Parallel Processors: A Hands-on Approach," Morgan Kaufman (pubs), Second Edition, Print Book ISBN: 9780124159921 eBook ISBN: 9780123914187
- Conference and Journal Publications
- Class Notes

Topical Outline

- Introduction
  - Bulk Synchronous Parallel (BSP) models
  - CUDA vs. OpenCL
  - BSP Algorithms for common primitives
- Microarchitecture
  - Basic microarchitecture concepts and the SIMT execution model
    - Kernel launch, scheduling, and control flow management
  - Memory hierarchy operation
    - Memory coalescing and shared memory management
    - Cache management
  - Discrete vs. integrated GPUs
- Control Divergence
  - Introduction to control divergence and solutions
  - Optimizations for control divergence management

        - Dynamic Warp Formation and Thread Frontiers
        - Thread Block Compaction and Dynamic Warp Subdivision
        - Emerging techniques
- GPU Memory Hierarchy: Key concepts underscoring the operation of memory hierarchies in discrete and integrated GPUs
    - Uniform virtual memory (UVM)
    - CPU- GPU coherency issues
    - Introduction to memory divergence and latency hiding techniques
    - Dynamic vs. static techniques for mitigating memory divergence
- Scheduling: Scheduling optimizations
    - Warp scheduling algorithms
    - Thread block scheduling
    - Optimizations for throughput vs. energy
- Advanced Microarchitecture Concepts: Optimizations of the GPU microarchitecture and memory system
    - Organization of the register files and RF-Core interconnect
    - Cache and memory system optimizations
    - Optimizations for power and energy efficiency
- Competing (Accelerator) Models and Architectures
    - Systolic Model of Computing
        - Basic model elements and algorithmic primitives
    - Data Flow Execution
        - Static and dynamic data flow

Course Grading:

Mid-term: 15%
Assignments: Mini-projects (40%)
Final Project: 35%
Final: 10%

The bulk of the course content will be based on material from workshop, conference, and journal publications. The midterm will test understanding of fundamental concepts. The project will address a coherent research theme in modern GPU architectures. The final exam will be based on a project report in conference paper format.