EE 7700-2—GPU Microarchitecture

URL: `http://www.ece.lsu.edu/gp/`

Offered by:

David M. Koppelman

3191 P. F. Taylor Hall, 578-5482, koppel@ece.lsu.edu, http://www.ece.lsu.edu/koppel

Office Hours: Monday - Friday, 14:00-15:00.

Prerequisites By Topic:

- Computer architecture and digital logic.
- C++ and machine language programming.

Text

Papers, technical reports, etc. (Available online.)

00-2

Course Objectives

Understand low-level GPU organization.

Be able to fine-tune GPU codes based on this low-level knowledge.

Understand fundamental issues in the CPU v. Manycore v. GPU v. FPGA "debate."

Course Topics

NVIDIA GPU Organization

CUDA Programming

GPU Algorithms

Tuning based on machine instruction analysis.

Tuning using performance counters.

Graded Material

00-3

Midterm Exam, 35%

Fifty minutes in-class or take-home.

Final Exam, 35%

Two hours in-class or take-home.

Yes, it's cumulative.

Homework, 30%

Written and computer assignments.

Lowest grade or unsubmitted assignment dropped.

Course Usefulness

00-4

Material in Course Needed For:

Those designing the next generation of GPU chips.

Those writing high-performance scientific programs.

Those writing high-performance graphics programs.

Those interested in future computer architectures.

Compiler writers.

00-5

Course Resources

Slides and other material via `http://www.ece.lsu.edu/gp/`

Web site also has homework assignments, exams, grades, and other material.

Announcements are on course home page and available as a Web (RSS) Feed.

00-6

GPU Definition

GPU:

A processor designed to execute a class of programs that includes 3D graphics rendering and scientific computation. GPU organization is fundamentally different than CPU.

GPUs originally designed *only* for 3D graphics.

Large economies of scale made them cheap.

Resourceful scientific users disguised their work as 3D graphics.

GPU makers started supporting scientific and other non-graphical work.

GPU evolved into a second kind of processor, with 3D graphics just one application.

00-7

Organization of Systems Using GPUs

Desktop Computer

GPU and CPU each of their own memory.

Communicate over high-speed bus. (Often, PCI Express)

CPU in charge.

00-8

GPU v. CPU

GPU can execute more floating-point operations per second.

GPU can read data at a higher rate.

CPUs are easier to program.

CPUs can run certain programs faster.

00-9

Why GPUs are important.

They can do certain things much better than a CPU.

They have succeeded where many have failed.

May be only viable way of overcoming fundamental barriers to faster computation.

00-10

GPUs succeeded where many failed: Establishing a new computer architecture.

CPUs today share similar architecture.

Are designed for multiple uses: business, scientific.

Exceptions are minor: embedded.

In the past specialized architectures could be successful.

Cray supercomputers.

In most cases, different or specialized architectures failed:

Database machines.

LISP machines.

00-11

Alternatives failed because:

They were not that much faster.

Were very expensive to develop.

Market was small and so engineering big part of cost.

00-12

Reason GPUs succeeded:

First, their birth.

3D computer graphics is compute intensive. (Think frame rate.)

Need.

Graphics computation is well structured.

Facilitates development of specialized processor.

Amount of code relatively small.

Can be isolated in libraries.

Large market: Home computers, game consoles.

Can amortize development costs.

00-13

Their evolution to non-graphical use.

Computational characteristics of 3D graphics code shared by other types of programs.

GPUs could easily be modified to support non-graphical use.

00-14

CPU GPU Similarities

Both run programs.

Machine languages are similar.

Both have instructions for integer and FP operations.

High-end chips are roughly same area, and draw same power.

00-15

GPU Incidental (and diminishing) Differences

Special hardware for *texture fetch and filtering* operations.

Special hardware for *interpolation*.

Trend is less specialized hardware.

GPU Important Performance Spec Differences

GPUs can do more FP operations per second.

GPUs can transfer more data per second.

GPU Programmability Differences

Extensive tuning required to achieve performance.

GPUs perform poorly on certain problems, regardless of tuning.