Brief Introduction to CUDA in 2 days: Computer Organization

David Bunde

Knox College

Unit goals

- Idea of parallelism
- Benefits and costs of system heterogeneity
- Data movement and NUMA
- Generally, the effect of architecture on program performance

Constraints

- Brief time: Course has lots of other goals
  - One 70-minute lab and parts of 2 lectures
- Relatively inexperienced students
  - Some just out of CS 2
  - Many didn't know C or Unix programming

Approach taken

- Introductory lecture
  - GPUs: massively parallel, outside CPU, kernels, SIMD
- Lab illustrating features of CUDA architecture
  - Data transfer time
  - Thread divergence
  - Memory types (next time)
- “Lessons learned” lecture
  - Reiterate architecture
  - Demonstrate speedup with Game of Life
  - Talk about use in Top 500 systems

Survey results: Good news

- Asked to describe CPU/GPU interaction:
  - 9 of 11 mention both data movement and invoking kernel
  - Another just mentions invoking the kernel
- Asked to explain experiment illustrating data movement cost:
  - 9 of 12 say comparing computation and communication cost
  - 2 more talk about comparing different operations

Survey results: Not so good news

- Asked to explain experiment illustrating thread divergence:
  - 2 of 9 were correct
  - 2 more seemed to understand, but misused terminology
  - 3 more remembered performance effect, but said nothing about the cause

Conclusions and future plans

- Unit was mostly successful, but thread divergence is a harder concept
- Students interested in CUDA and about half the class requested more of it
  - Will be added constant memory and a small assignment to next offering (Winter 2013)
- Bottom line: A brief introduction is possible even to students with limited background