Future GPU Architectures Panel

Scott Hemmert

Scalable Computer Architectures
Sandia National Laboratories

Sandia is a Multiprogram Laboratory Operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy Under Contract DE-ACO4-94AL85000.

First, the obvious…

- **Three groups of applications with respect to GPUs**
  - **Group 1: Those that map well and perform better on GPUs than on CPUs**
    - **Tend to have regular data access patterns and be very compute intensive**
  - **Group 2: Those that perform OK on GPUs**
    - **May be relatively compute intensive, but have some complex memory access patterns or control flow**
  - **Group 3: Those that don't perform well on GPUs**
    - **Tend to be memory intensive and/or have irregular memory access patterns or have complicated control flow**

The Hurdles

- **Where do our important applications fall into the three groups?**
  - **Are there enough group 1 applications to justify it?**

- **Building a large scale (production) system from today's GPUs is a tough sell**
  - **Data movement issues**
    - **GPUs sit on I/O bus and data must be moved from main memory to GPU memory**
    - **GPUs can't control the NIC making system interconnect less efficient**
  - **Some concern about system reliability**
    - **Do commodity GPUs provide important resiliency features**
  - **System software issues?**
  - **Programming issues?**

Looking Forward – Some Predictions

- GPU architectures will become more flexible
  - More applications will benefit
    - Many group 2 applications may move to group 1
  - but, “relative” FLOPS/watt will decrease (flexibility comes at the cost of efficiency)
- CPU/GPU memory spaces will merge
  - Data movement will no longer by mandatory (may still be beneficial)
- CPUs and GPUs will eventually be integrated onto same die
  - Two possible outcomes (both may happen simultaneously)
    - Heterogeneous multi-core (separate CPU and stream cores)
    - Homogeneous multi-core (CPU cores with attached stream/vector units)
  - Which is easier to program?
  - Which will have more success in base graphics market?
- GPU architectures will evolve enough that many codes will need to be re-optimized