General Purpose GPU

Aurojit Panda

apanda@cs.berkeley.edu
**Summary**

- SIMD helps increase performance while using less power
  - For some tasks (not everything can use data parallelism).
  - Can use less power since DLP allows use of many more, slower components for similar or better throughput.
- Vector processors were popular for HPC applications once..
- Popularized in a limited way for multimedia applications (SSE, etc).
- Popularized in a more full fledged version with GPUs.
- GPUs increasingly used for things that are unrelated to graphics.
Basic Problem

Maximize processing power while minimizing energy consumption
- NVIDIA: Compared to the latest quad-core CPUs, Tesla C2050 and C2070 Computing Processors deliver equivalent supercomputing performance at 1/10th the cost and 1/20th the power consumption.

- 3 of 5 the top supercomputers as rated by the Top500 project use GPUs for some of their computing.
Amazon began offering EC2 "Cluster GPU" instances in November.

For $2.00 an hour 2 NVIDIA Fermi M2050s rated at 512 GFlops each.

Amongst other things can run MapReduce jobs on GPUs using EMR.

Many instances used to mine bitcoins but...
EC2 Cluster Uses

- Also used by people in life sciences
  - Directly by a few places.
  - Through intermediaries like CycleComputing in some cases.
- Generally makes HPC application possible for people without much investment
The code looks very different from what is written for the CPU, it’s hard to reason about code where you need access to both.

- Lots of libraries and extensions try to fix this by providing better abstractions.
- Intel is trying to fix this (to an extent) with Larrabee

For some applications need a lot of traffic between main memory and GPU memory, no I/O support directly to the GPU

Poor debugging support

- NVIDIA and AMD are both trying to fix this in different ways
- Often end up emulating GPU on CPUs to get more information (WARP) or using traces

GPU virtualization wasn’t present until recently, not widespread
In the video Barroso talked about the problems with Wimpy cores, and GPUs do bring up similar questions about heterogeneity, and specialized cores.

Barroso’s point was that small wimpy cores (think a single SIMD lane on a GPU) aren’t useful unless programs using them are parallel.

Easier to do request level parallelism, don’t have to design for this parallelism.

*Not everything is easily vectorizable*

On the flip side, GPUs are much better for some tasks.
Both NVIDIA and Intel talk about the demise of the separation between CPUs and GPUs

- NVIDIA envisions a single package with a beefy ARM core and a GPU
- Beefier than what the Tegra 2 or Tegra 3 have right now

Lots of work on utilizing GPUs for:

- Data-mining, analytics, database queries
- Bio-informatics
- Automatic trading systems, options pricing, financial risk analysis
- Numerical analysis
- Other stuff...
Computing Power

![Graph showing the growth of computing power from 2002 to 2011 for CPU and GPU. The y-axis represents Gflops, and the x-axis represents years. The graph shows a significant increase in computing power for both CPU and GPU over the years, with the GPU showing a steeper increase.]
Predictions for the Cloud

- **High Probability**
  - Greater use for analytics, processing queries, other things
  - More GPGPUs in clusters (perhaps more per machine)

- **Somewhat Probable**
  - More HPC users move to the "cloud" (DE Shaw Research might not have to build its own supercomputer)
  - Separate clusters for GPU and CPU workloads(?)

- **Less Probable**
  - CPUs, as they are now, disappear (from clusters and computers)
  - GPUs, as they are now, disappear (from clusters and computers)