CSE 4373/5373 - General Purpose GPU Programming

Profiling CUDA Applications

Alex Dillhoff

University of Texas at Arlington

Analysis Driven Optimization

Analysis Driven Optimization

- Global memory accesses are one of the largest bottlenecks in GPU applications.
- DRAM has high latency based on its design.
- Each cell has a transistor and a capacitor.
- If the capacitor is charged, it represents a 1.
- The process to detect the charges in these cells is on the order of 10s of nanoseconds.

Performance Limiters

NSight Compute

NSight Compute

- NVIDIA NSight Compute is a profiling tool for CUDA kernels.
- It features an expert system that can help you identify performance bottlenecks in your code.
- It is useful in tandem with methodically analyzing your code.
- It is NOT a debugger.

Profiling with NSight Compute