Homework Slides - 3/11/2013 (paper 2)

Jaylen VanOrden

S1 - GPUs and Parallel Computing

- Moore's law continues, but single core clock rate + performance scaling ceases, parallelised HW/SW becomes more mainstream
- GPUs bring 1000s of threads and high memory bandwidth to very many simple processing cores
- Systems built around GPUs are leaders in performance and efficiency
- Three issues affect future GPU scaling:
  - Power supply voltage scaling
  - Slowing memory bandwidth scaling
  - Parallel programming
- Goal power cost per flop is around 20 pJ
  - current CPUs run about 1.7 nJ
  - current GPUs run about 225 pJ

S2 - GPUs and Parallel Computing

- Linear power scaling by feature size can give only about 4x power savings, other savings must be architectural
- Processors evolved to get the most performance per chip area, resulting in high-overhead, fast single threads
- Energy costs for memory retrieval will become more important, with energy scaling similar to latency scaling
- Memory bandwidth scaling is slowing and transfer cost is high, so memory locality is becoming more important
- One possible solution is to place the GPU and DRAM on the same chip, or use chip stacking and/or vias (TSVs)
- Might improve DRAM bandwidth by reducing overfetching and increasing data density (compression?)
- Future systems will likely focus on a CPU and GPU on the same die and using the same memory architecture

S3 - GPUs and Parallel Computing

- Echelon - a GPU system built to address scaling concerns and efficiency of very parallel computation
  - Heterogeneous cores
    - Latency-optimized cores for single-thread perf.
    - Throughput-optimized cores for parallel perf.
  - Adding a cache to the register file
  - On-chip thread scheduling based on moving only PC
  - "Temporal SIMT", some threads share PC, registers
  - Programmer-specifiable memory system, allows for private or shared L2 or global L3 caches
  - Memory regions marked for consistency, allows programmers to selectively move data structures into less consistent but faster memory areas
  - Memory regions also marked for LOC vs TOC

?s - GPUs and Parallel Computing

- The paper assumes DRAM 'pin bandwidth' will increase from 4 Gbps in 2010 to 50 Gbps in 2017 but is vague on why or how this will happen. Do we have more information on this?
- How does branching/merging work in thread coalescing? What exactly is temporal SIMT?