A Taxonomy of GPGPU Performance Scaling

Abhinandan Majumdar, Gene Wu, Kapil Dev
Joseph L. Greathouse, Indrani Paul, Wei Huang, Arjun Karthik Venugopal, Leonardo Piga, Chip Freitag, Sooraj Puthoor

**Goals**
- Observe how GPGPU performance scales at different hardware configurations
- Quantitatively determine principal performance scaling trends across 267 GPGPU kernels from 97 programs profiled on 891 GPU configurations
- Performance studied across 5x change in core frequency, 3.3x change in memory bandwidth, and 11x difference in compute units

**Platform**
- **AMD FirePro™ W9100 GPU**
  - 2,816 processing elements (44 CUs) at 930 MHz
  - 16 KB of L1 data cache per CU
  - 1 MB of L2 cache shared across all CUs
  - 16GB GDDR5 GPU memory at 1.25 GHz
  - 320 GB/s memory bandwidth

**Experimental Setup**
- Core frequency variation: 200 MHz to 1 GHz
- Memory frequency: 150 MHz to 1.25 GHz
- Variation in number of CUs: 4 to 44
- June 20, 2014 beta of AMD FirePro™ drivers
- AMD APP SDK version 2.9
- AMD CodeXL version 1.4

**Performance Scaling Trends**

**Conclusion and Future Work**
- The performance of many kernels scales as more CUs are added
- Around 40% of the kernel iterations do not scale to modern GPU sizes
- Future studies should examine whether existing benchmarks (and input sets) are representative GPGPU workloads
- Future work could also consider other hardware configurations, such as cache sizes or double precision rate, and could characterize power and energy