**Introduction**

General-Purpose computation on Graphics Processing Units (GPGPU) has made great strides in aiding scientists and engineers with problems requiring massive computational power. One bottleneck to performance is the latency required for data to migrate from the CPU to the GPU. Future architectures may include designs where the CPU and GPU are merged onto the same chip eliminating this bottleneck. We attempt to search this design space and explore the design tradeoffs with a total fixed area constraint involving a single out-of-order execution core and several smaller in-order SIMD execution cores connected by a 2D mesh and a shared L2 cache.

**Simulation Workflow**

Simulation concerns:
- The design space consists of over 1 million configurations.
- A simulation can take between a couple hours to several days.
- We simultaneously try two approaches to explore the design space.
- We randomly sample the design space and perform sensitivity tests on the best configurations to find more optimal configurations.
- We are in the process of using GPRSkit 0.2\(^1\) to generate a genetically programmed response surface (GPRS) in the form of an equation from a relatively small set of simulations.

**Framework**

We use the M5 Simulator\(^2\) with Jiayuan Meng’s patches\(^3\) for:
- SIMD
- Multithreading in SE mode
- Directory-based coherence
- 2D Mesh Interconnect
- Banked caches

Available benchmarks:
- Filter, FFT, Shortest Path, KMEANS, Mergesort, Needleman-Wunsch, Hotspot

We adapt and use a rough area model from Tarjan et al.\(^4\) which uses measurements from a publicly available Opteron die photo. We restrict total area to between 380 and 420 mm\(^2\).

**Preliminary Results**

For a small workload with the Filter benchmark, we simulated 482 configurations. Below, we plot execution time of all of the simulations in order, the number of IO cores, the SIMD width, and the number of SIMD groups for all 482 simulations sorted according to best performance, and list the top 10 configurations.

**Future Work**

Interesting directions for future work include simulation with larger workloads, exploration of thermal effects/constraints, comparison with a configuration in which we use a third level cache as a means for off-chip communication between the out-of-order core and set of many in-order cores.