Design Space Exploration for Integrated CPU-GPU Chips

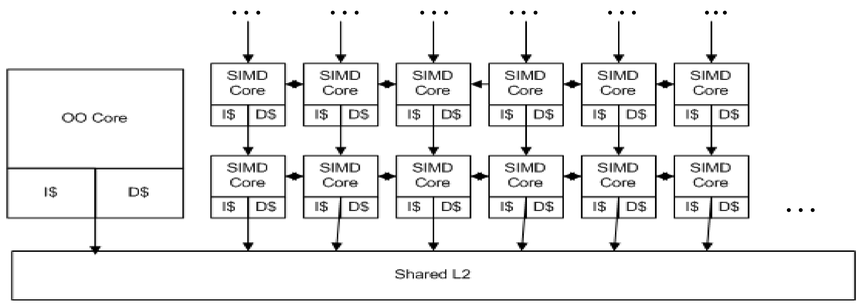
University of Virginia, Charlottesville, VA 22904

Paul Lee, Jiayuan Meng, Zhenyu Qi, Mircea Stan, Kevin Skadron

This work is supported by a grant from the Semiconductor Research Corporation under task 1607.

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.



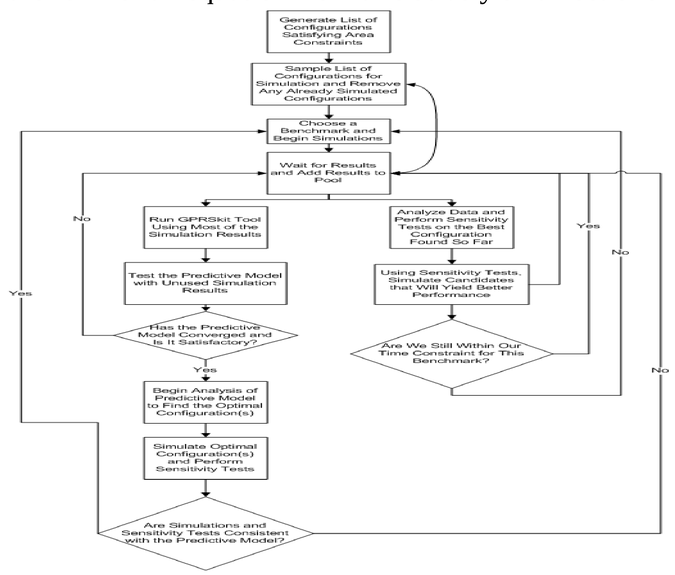
Simulation Workflow

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.

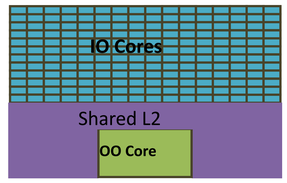


[1] Cook, H. and Skadron, K. 2008. Predictive design space exploration using genetically programmed response surfaces. DAC '08: 960-965

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

Parameters Varied for Simulation	
Param	Values
OO L1 Cache	32kB, 64kB, 128kB
OO Width	1, 2, 4, 8
OO IQ	32, 64, 96, 128
OO ROB	64, 128, 196, 256
OO Registers	128, 192, 256
OO LQ & SB	16, 32, 48, 64
OO Branch Predictor*	0, 1, 2
# of IOs	1, 2, 4, 8-128 in steps of 8
IO L1 Cache	16, 32, 64
SIMD Width	1, 2, 4, 8, 16, 32
# of SIMD Groups	1, 2, 4, 8, 16, 32, 64, 128
Shared L2	2048kB, 4096kB, 8192kB, 16384kB

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

Additional Fixed Parameters	
OO Clock	2.6 GHz
IO Clock	1.5 GHz
Physical Memory	4096MB
Coherence Protocol	MESI

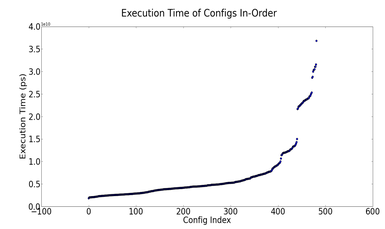
*

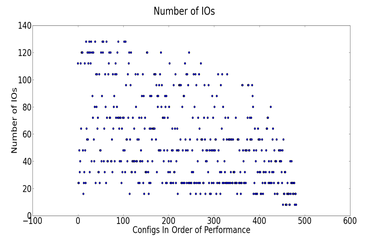
Branch Predictor Schemes					
Branch Predictor Scheme #	Choice Predictor PHT	Local Predictor PHT	Global Predictor PHT	Branch History Table	Total Bits
0	4k	1k	4k	1k 10-bit entries	28 Kbits
1	1k	512	2k	512 2-bit entries	8 Kbits
2	8k	4k	16k	1k 8-bit entries	64 Kbits

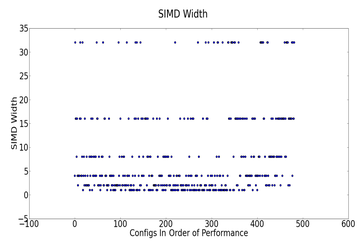
[2] http://www.m5sim.org/wiki/index.php/Main_Page
[3] http://www.cs.virginia.edu/~jm6dg/fractal/m5patches.htm
[4] Tarjan, D. and Boyer, M. and Skadron, K. 2008. Federation: repurposing scalar cores for out-of-order instruction issue. DAC 2008: 772-775

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.







Top 10 Configurations out of 482 Simulations												
OO Cache Size	OO Width	OO IQ	OO LQ & SB	OO Registers	OO ROB	OO Predictor	Num IOs	IO Cache Size	SIMD Width	SIMD Groups	L2 Size	Execution Time (ps)
64	8	128	32	256	256	2	112	16	4	2	4096	1800841735
32	8	128	64	256	128	2	24	32	32	1	8192	1960788830
32	4	64	64	256	128	0	24	64	16	2	8192	2009837445
64	4	128	48	192	196	1	48	16	16	1	8192	2015755665
64	4	32	48	256	256	0	32	64	8	1	8192	2024269555
64	4	32	48	128	64	2	40	32	16	1	8192	2025062270
64	8	96	16	128	128	1	112	16	4	1	4096	2044027755
64	8	32	64	128	128	1	64	32	2	2	8192	2057101200
64	4	128	16	256	128	2	120	16	4	1	4096	2058722050
32	2	96	64	256	196	0	120	16	4	4	4096	2061625335

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.