A Comparative Power-Performance Analysis of Microarchitecture Effects on Heterogeneous CPU-GPU

Vijayalakshmi Saravanan, Ramalingam Sridhar
University at Buffalo, Department of Computer Science and Engineering
Email: {vsaravan, rsridhar}@buffalo.edu

University at Buffalo
The State University of New York

BACKGROUND

- Processor throughput was first improved by increasing clock frequency. Due to a variety of factors improving processor performance turned to adding multiple cores.
- Current advancement in HPC paves the way to integrate CPU and GPU in single chip. The main challenge in GPU processor architecture is rendering speed and more heat dissipation.
- The performance of the processors depends on many micro-architectural parameters such as issue-width, functional units and pipeline depth.
- This poster aims to evaluate the scalability of CPU-GPU per stream cores and micro-architectural behavior for parallel applications executing on each core type in heterogeneous CPU-GPU processor simulation.

INTRODUCTION

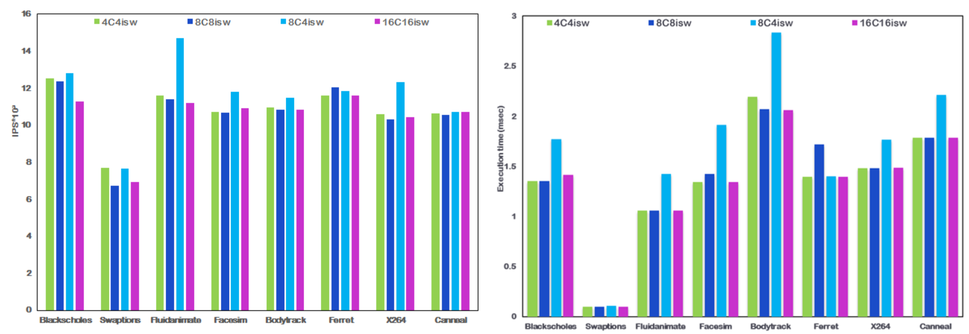
- Desktop, laptops, mobile and high-performance computing systems incorporating GPU's in order to increase the throughput. Within the last couple of years, heterogeneous computing systems evolved to provide fast computing.
- Modern processors widely employ large on-chip resources such as functional units, caches and pipeline which in turn increase the complexity. The performance of the processor is based on two important metrics such as architecture and application.
- The relationship between specific micro-architectural parameters of heterogeneous CPU-GPU computing are seldom analyzed in published studies.
- Similar study in IISWC 2014 compare the micro-architecture effects on CPU and GPU memory system behavior.
- This poster proposes to study the trade-offs between power-performance and thermal constraints of CPU-GPU processors.

METHODOLOGY

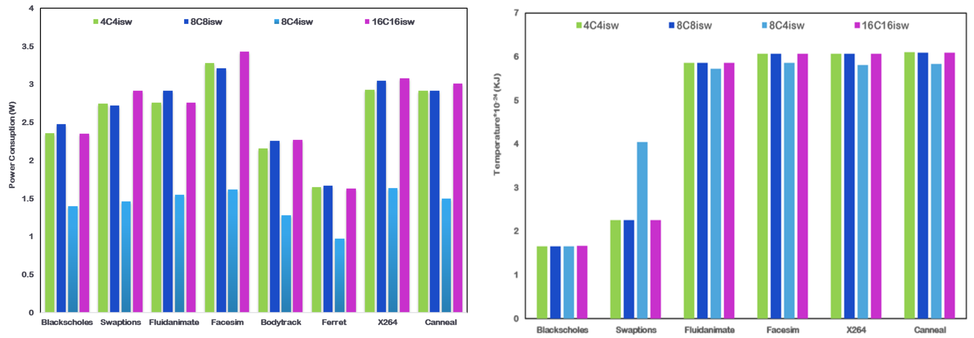
To analyze the micro-architectural behavior, we consider three micro-architecture design points that differ in issue-width, number of cores ranging from quad-core to 16 cores and their issue-width ranging from 4 to 16 and various pipe-line depth implementation and compared the power-performance.

a) Various PARSEC workloads such as Blackscholes, Swaptions, Facesim, Fluidanimate and Bodytrack are used to compute intensive workloads capable of parallel processing;
b) Blackscholes and Swaptions are number crunchers and hence need to have only higher computational processing whereas Facesim, Fluidanimate and Bodytrack require both higher computational and graphical processing.
c) Ferret and X264 are six-stage pipeline models with varying functional aspects. Ferret is computational intensive whereas X264 is both computational and graphical intensive workload.

RESULTS



Issue-width vs. Performance Results



Issue-width vs. Power, Thermal Results

EXPERIMENT

1. 64 bit, Intel core i7-2430M with a clock speed of 2.40GHz.
2. Simulator used: ESESC, Enhanced SuperESCalar and Samsung Exynos4 ARM processor architecture.
3. Benchmarks: CPU-GPU: PARSEC Benchmark suite.

Table. 1 Benchmarks and Workload size

Program	Application Domain	Workload Nature	Program size
Blackscholes	Financial Analysis	Data-parallel	65,536 options
Swaptions	Financial Analysis	Data-parallel	16 swap, 20,000 simulations
Facesim	Animation	Data-parallel	1 frame, 372,126
Fluidanimate	Animation	Data-parallel	10 frame, 300K particles
X264	Media processing	Pipeline	128 frames, 640x360 pixels
Ferret	Similarity search	Pipeline	256 queries, 34,973 images
Canneal	Engineering	Unstructured	2500000 elements
Bodytrack	Computer vision	Data-parallel	4 frames, 4,000 particles

4. Power-performance Metrics Used: Instructions per Second (IPS), Execution time and Energy and Temperature metrics obtained.

CONCLUSIONS

- ❖ Detailed results of a CPU-GPU heterogeneous simulation for various workloads are presented in this poster.
- ❖ Our experimental results show that for all the compute intensive workloads capable of parallel processing, eight-core with four issue-width is the optimal design which is high performing as well as energy and thermal efficient; except for Ferret which recorded a meagre 1.75% drop in IPS performance as compared to its best run with eight cores with eight issue-width. However, eight-core eight issue-width has the best execution time for X264 & four core with four issue-width for ferret when compared to others.
- ❖ If we consider pipeline modeled workloads such as ferret and X264 has the best throughput IPS and execution time for eight-four core with eight-four issue-width respectively.
- ❖ We can certainly see that high issue-widths/cores of 16 provide very little to no increases in performance in terms of IPS and the power dissipated at the same time is very high. Hence the ideal issue-width would be around eight-core with four issue-width (8C4isw) wherein power-performance optima is observed.

Acknowledgements: The authors thank "Schlumberger Foundation Faculty for the Future" (FFTF) and NSF. This work was supported in part by NSF grant CNS- 1422304.