Heterogeneous Multi-core Architectures: Optimizing Power and Performance

Naman Jain, Prashanth Suresh
Department of Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh PA 15213
{namanj, psuresh}@andrew.cmu.edu

***Abstract*—In order to satisfy the high-performance and low-power requirements for advanced embedded systems with greater flexibility, it is necessary to develop chip multiprocessors. However, homogeneous multi-core systems do not achieve higher potential per watt when compared to heterogeneous architectures as serialized code sections can be accelerated without programmer effort. Therefore, this project aims to evaluate single ISA heterogeneous multi-core architectures as a way to reduce processor power dissipation by switching cores to achieve target performance level. The project involves testing different sizes and configuration of cores on a single chip which achieves maximum power reduction. gem5 full-system simulator or Sniper Multi-core simulator will be used for performance modelling and McPAT will help for power modelling.**

I. Introduction

The increase in need for machines with higher performance, computational power and increase in complexity in the design of uniprocessor has been the driving force for increase in interest in design of multi-core architecture. A multi-core processor has multiple cores integrated on a single chip. A multi-core architecture where every core is just an image of the other is called homogeneous multicore. Heterogeneous is a set of cores which may differ in area, performance, power dissipated etc. The various design issues in multi-core architecture include resource sharing, power consumption, performance, area of the cache, cache coherence etc. In order to harness the resources provided by a multicore architecture the application must show a certain level of parallelism.

With additional cores, power consumption and heat dissipation becomes a concern and must be simulated before layout to determine the best floorplan which distributes heat across the chip, while being careful not to form any hot spots [8]. Distributed and shared caches on the chip must adhere to coherence protocols to make sure that when a core reads from memory it is reading the current piece of data and not a value that has been updated by a different core. This adds to the total power consumption as well.

To satisfy the high-performance and low-power requirements for advanced embedded systems with greater flexibility, it is necessary to develop parallel processing on chips by taking advantage of the advances being made in semiconductor integration [12]. Heterogeneous cores provide different power/performance tradeoff [3]. In order to benefit from heterogeneous multicore architectures, the scheduler needs to consider the power/performance asymmetry of heterogeneous multicore architectures when making a scheduling decision.

II. Related Work

Considerable amount of work has been done on power-related optimizations for processor design. These can be broadly classified into two categories: (1) work that uses gating for power management, and (2) work that uses voltage and frequency scaling of the processor core to reduce power [2].

Gating-based power optimizations [13], [14] provide the option to turn off (gate) portions of the processor core that are not useful to a workload. However, for all these techniques, gating benefits are limited by the granularity of structures that can be gated, the inability to change the overall size and complexity of the processor. Also, these designs are still susceptible to static leakage inefficiencies.

Chip-wide voltage and frequency scaling reduces the parameters of the entire core [15]. While this reduces power, the power reductions are uniform across both the portions of the core that are performance critical for this workload as well as the portions of the core that are not. Furthermore, voltage and frequency scaling is fundamentally limited by the process technology in which the processor is built. Heterogeneous multi-core designs address both these deficiencies.

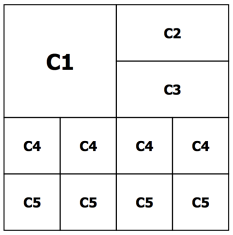

Fig. 1: Example Heterogeneous Design

Fine-grained voltage/frequency scaling techniques using multiple clock domains have been proposed recently which obviate some of the disadvantages of conventional scaling-based techniques. However, similar to gating-based approaches, the benefits are likely to be limited by static leakage inefficiencies as well as the number of voltage domains that can be supported on a chip.

Core switching to reduce power was introduced previously in [14]. Recently, it has also been used for reducing power density in a homogeneous multiple-core architecture, through the use of frequent core switches to idle processors.

Overall, having heterogeneous processor cores provides potentially greater power savings compared to previous approaches [1], [2] and greater flexibility and scalability of architecture design.

III. Technical Details

To attain high-performance and low-power requirements for advanced embedded systems it is necessary to develop parallel processing on chips by taking advantage of the advances being made in semiconductor integration. The subsequent sections explain why heterogeneous processors are chosen, proposes a design to get power/performance optimum value and explains the caveats related to this implementation.

A. Why Heterogenous?

In an ideal homogeneous architecture, processing capability and power consumption are evenly distributed among its cores. Thus the decrease of computation time and the increase of system power are linear with the number of active cores. So the total energy cost (the Power-Delay product) of solving a particular problem is constant. However, these assumptions are not valid. First, as per Amdahl's law, performance improvement suffers from the law of diminishing return such that speedup may not scale linearly with the number of cores. Second, not all energy consumed by the system contributes to useful computation. Power consumption overhead is not linear with the number of active cores. Third, the achievable performance is limited by architectural factors such as I/O bandwidth and cache size.

In a heterogeneous multicore architecture a core may differ from the other cores in many ways. First, we focus on the performance aspect of multicores. Fig. 1 shows an example of a heterogeneous multi-core architecture on a single chip. The parallelizable portion of the code can be executed on smaller cores C2, C4, C5, whereas large critical sections can be executed on the large core C1. Building a performance heterogeneous core is desired because many simple cores together provide high parallel performance while complex cores help in providing high serial performance. Amdahl's Law is still relevant as we enter a heterogeneous multi-core computing era [9]. Amdahl's Law is a simple analytical model that helps developers to evaluate the actual speedup that can be achieved using a parallel program. If p is the number of processors and α is the parallelizable fraction of a program, then Amdahl's law states that speedup is given by

$$Speedup = \frac{1}{\frac{\alpha}{p} + (1 - \alpha)}$$

B. Design

A single processor will contain many small simple cores, and a few larger complex cores. Simple cores will be scalar, in-order and might have a smaller cache and lower clock frequency. Complex cores will have superscalar design, and may be equipped with high-performance features such as out-of-order instruction scheduling, aggressive prefetching or a vector unit. However, they will be larger and will consume significantly more power. Heterogeneous architectures are motivated by their potential to achieve a higher performance per watt as compared to homogeneous systems because each application can run on a core that best suits its state. For example, a gaming app can run on a simple core when user is setting configurations or have paused a game, but the application should switch to complex core once the actual intensive gaming starts.

This project will focus primarily upon optimizing power and performance with respect to size and number of cores for a given power and performance requirement in a system.

C. Architecture Modelling

gem5 [4] is a modular platform for computer system architecture research, encompassing system-level architecture as well as processor microarchitecture. It provides state-of-the-art detailed microarchitectural design. Homogeneous single ISA have been modelled before [17]. However, heterogenous design modelling is not yet confirmed to be accurate.

On the other hand, Sniper Multi-core [10] has easier interface for multiple heterogenous core modelling but the results might not be as accurate/detailed as gem5 because it does not use full-system simulation.

During our initial stage the more appropriate performance simulator will be chosen among them.

McPAT [11] will be used for power modelling as it has integrated power, area, and timing modeling framework for multicore architectures.

The architectural design will be tested using PARSEC/ SPLASH-2 benchmark suites. These benchmark suites focuses on emerging workloads and has been designed to be representative of next-generation shared-memory programs for chip-multiprocessors.

D. Caveats

To take full advantage of heterogeneous CMPs, the system software must use the execution characteristics of each application to predict its future processing needs and then schedule it to a core that matches those needs if one is available [1], [2]. The predictions can minimize the performance loss to the system as a whole rather than that of a single application. Recent work has shown that effective schedulers [6] for heterogeneous architectures can be implemented and integrated with current commercial operating systems. However, effective kernel scheduling will not be used for this project, but it can be assumed that the results i.e. power and performance gains will be much higher when integrated with heterogeneity-aware scheduling kernels.

IV. Schedule

Stage	Date	Progress
Initial	Oct 6	Setup gem5/Snipersim and McPAT
Midway	Oct 23	Complete heterogeneous design
75%	Nov 14	Measure power and performance for single heterogeneous configuration
Final	Dec 2	Obtain optimal power/performance sweet-spot
Sky	Dec 2	Obtain approximate heterogeneous system configuration for user specified requirements

REFERENCES

[1] Kumar, Rakesh, Dean M. Tullsen, and Norman P. Jouppi. "Core architecture optimization for heterogeneous chip multiprocessors." Proceedings of the 15th international conference on Parallel architectures and compilation techniques. ACM, 2006.

[2] Kumar, Rakesh, et al. "Single-ISA heterogeneous multi-core architectures: The potential for processor power reduction." Microarchitecture, 2003. MICRO-36. Proceedings. 36th Annual IEEE/ACM International Symposium on. IEEE, 2003.

[3] Lukefahr, Andrew, et al. "Composite cores: Pushing heterogeneity into a core." Proceedings of the 2012 45th Annual IEEE/ACM International Symposium on Microarchitecture. IEEE Computer Society, 2012.

[4] Binkert, Nathan, et al. "The gem5 simulator." ACM SIGARCH Computer Architecture News 39.2 (2011): 1-7.

[5] Mutlu, Onur, "Asymmetry" 18-740 Lecture 2.1 Fall 2013

[6] Ghiasi, Soraya, Tom Keller, and Freeman Rawson. "Scheduling for heterogeneous processors in server systems." Proceedings of the 2nd conference on Computing frontiers. ACM, 2005.

[7] Tsoi, Kuen Hung, and Wayne Luk. "Power profiling and optimization for heterogeneous multi-core systems." ACM SIGARCH Computer Architecture News 39.4 (2011): 8-13.

[8] Hyari, Abeer. "A comparative study on heterogeneous and homogeneous multiprocessors." University of Jordan (2009).

[9] Marowka, Ami. "Extending Amdahl's Law for Heterogeneous Computing." Parallel and Distributed Processing with Applications (ISPA), 2012 IEEE 10th International Symposium on. IEEE, 2012.

[10] Carlson, Trevor E., Wim Heirman, and Lieven Eeckhout. "Sniper: exploring the level of abstraction for scalable and accurate parallel multi-core simulation." Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis. ACM, 2011.

[11] Li, Sheng, et al. "McPAT: an integrated power, area, and timing modeling framework for multicore and manycore architectures." Microarchitecture, 2009. MICRO-42. 42nd Annual IEEE/ACM International Symposium on. IEEE, 2009.

[12] Uchiyama, Kunio, et al. Heterogeneous Multicore Processor Technologies for Embedded Systems. Springer, 2012.

[13] Folegnani, Daniele, and Antonio Gonzlez. "Reducing power consumption of the issue logic." In Workshop on Complexity-Effective Design. 2000.

[14] Ghiasi, Soraya et al. Using IPC Variation in Workloads with Externally Specified Rates to Reduce Power Consumption. 2000.

[15] Govil, Kinshuk, Edwin Chan, and Hal Wasserman. "Comparing algorithm for dynamic speed-setting of a low-power CPU." Proceedings of the 1st annual international conference on Mobile computing and networking. ACM, 1995.

[16] Heo, Seongmoo, Kenneth Barr, and Krste Asanovic. "Reducing power density through activity migration." Low Power Electronics and Design, 2003. ISLPED'03. Proceedings of the 2003 International Symposium on. IEEE, 2003.

[17] Spiliopoulos, Vasileios, et al. "Introducing DVFS-management in a full-system simulator." Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS). IEEE, 2013.