Power Aware Early Design Stage Hardware Software Co-Optimization

Souradip Sarkar[*,†,1], Wim Heirman[*,†],
Trevor E. Carlson[*,†], Lieven Eeckhout[*]

[*] *ELIS, Ghent University, Sint-Pietersnieuwstraat 41, 9000 Gent, Belgium*
[†] *Intel Exascience Lab, Kapeldreef 75, 3001 Leuven, Belgium*

ABSTRACT

Co-optimizing hardware and software can lead to substantial performance and energy benefits, and is becoming an increasingly important design paradigm. In scientific computing, power constraints increasingly necessitate the return to specialized chips such as Intel's MIC or IBM's BlueGene architectures.

To enable hardware/software co-design in early stages of the design cycle, we propose a simulation infrastructure methodology by combining high-abstraction performance simulation using Sniper with power modeling using McPAT and custom DRAM power models. Sniper/McPAT is fast — simulation speed is around 2 MIPS on an 8-core host machine — because it uses analytical modeling to abstract away core performance during multi-core simulation. We demonstrate Sniper/McPAT's accuracy through validation against real hardware; we report average performance and power prediction errors of 22.1% and 8.3%, respectively, for a set of SPEComp benchmarks.

KEYWORDS: Multicore; simulation; design-space exploration; power-modeling

1 Introduction

With limited increases in clock frequency because of power constraints, improving next-generation processor performance has become a real challenge. One increasingly attractive way to improve performance within a given power and energy budget is to optimize the system for a specific set of workloads. This avenue for optimizing performance is commonly used to evaluate designs from a range of different performance/power design points, from smartphones to tablets, game machines, data centers and supercomputers. Because computer systems are increasingly power and energy-constrained for numerous reasons including cooling, packaging, capital and operational costs, etc., it is to be expected that workload-optimized system design will become even more prevalent. A fundamental challenge regarding co-designing hardware and software is how to evaluate design decisions and make trade-offs early in the design cycle. A common approach in architecture design is to employ

[1]E-mail: {ssarkar,wheirman,tcarlson,leeckhou}@elis.UGent.be
[2]This work is funded by Intel and by the IWT. We used the compute infrastructure provided by the VSC Flemish Supercomputer Center, the ExaScience Lab, Leuven, Belgium and the Intel HPC Lab, Swindon, UK.

detailed cycle-accurate simulation. Unfortunately, cycle-accurate simulators are extremely slow, and are difficult to scale to large multi-core systems; further, developing such simulators is very time-consuming, To make things even worse, making a detailed cycle-accurate simulator power and energy-aware further increases development thus, they are inappropriate for the early design stages. and evaluation time. Clearly, driving hardware/software co-design through cycle-accurate simulation is particularly problematic.

We make the case for architectural simulation at a higher level of abstraction for driving early design stage hardware/software trade-off explorations (including 3D stacked memories), while considering both performance and power. Our simulation methodology leverages a mechanistic analytical performance model to abstract away core performance, i.e., core performance is estimated through an analytical model while simulating the uncore (memory hierarchy, interconnection network, etc.) at some level of detail in order to capture inter-core performance interactions. Coupling this high-abstraction performance simulation approach, called interval simulation as implemented in Sniper [CHE11], with high-level power modeling using McPAT [LAS+09] and custom DRAM power models, we achieve both good accuracy and speed. We demonstrate the power of Sniper/McPAT which is a hardware-validated, accurate (for both performance and power), parallel simulator that can run multi-threaded and multi-programmed workloads on multi-core hardware.

2 Sniper/McPAT Simulation Methodology

Sniper/McPAT combines Sniper for performance modeling with McPAT and custom DRAM models for power modeling. Sniper, in addition to generating an overall performance estimate, also generates a number of statistics that serve as input for estimating power consumption using McPAT.

3 Architectural exploration

We perform a design space exploration in which we compare four architectural alternatives. The main insight that we aspire to explore is with technological advancement by two technology nodes, from 45 nm to 22 nm, how can we best use the available improvements in transistor density and energy efficiency.

The first architecture considered is a conservative integration, in which we integrate the eight cores of the dual-socket quad-core Nehalem machine onto a single chip. Together with a slight increase in clock frequency (3.059 GHz) and cache sizes (512 KB L2 and 32 MB L3 caches), this forms our *8-core* design point.

In addition to this conservative scaling option, we also explore several more drastic modifications. The three alternate architecture design points that we consider in this trade-off study each have 16 cores or twice the number of cores compared to the conservative option, with each core having half the L2 cache size (256 KB versus 512 KB for the 8-core architecture). Other modifications are as follows:

- The *3D* design point does not integrate an L3 cache but uses 3D stacked memory instead, which has a higher memory bandwidth and slightly shorter memory access time compared to regular DDR3 memory. This architecture results in a slightly bigger chip and nearly twice the power budget.

Parameter	Nehalem	8-core	3D	low-frequency	dual-issue
Sockets per system	2	1	1	1	1
Cores per socket	4	8	16	16	16
Core frequency	2.66 GHz	3.059 GHz	3.059 GHz	1.8 GHz	3.059 GHz
Core voltage	1.2 V	1.2 V	1.2 V	1.025 V	1.2 V
Issue width	4	4	4	4	2
ROB size	128	128	128	128	32
L2 cache size (per core)	256 KB	512 KB	256 KB	256 KB	256 KB
L3 cache size	8 MB per chip	32 MB	—	8 MB per 8 cores	8 MB per 8 cores
Memory bandwidth	8 GB/s	8 GB/s	128 GB/s	8 GB/s	8 GB/s
Memory latency	65 ns	65 ns	50 ns	65 ns	65 ns
Technology node	45 nm	22 nm	22 nm	22 nm	22 nm
Chip area	2× 243 mm^2	151 mm^2	181 mm^2	208 mm^2	187 mm^2
Maximum observed power	2× 99 W	80 W	130 W	58 W	102 W

Table 1: Simulated system characteristics used in the architectural exploration study.

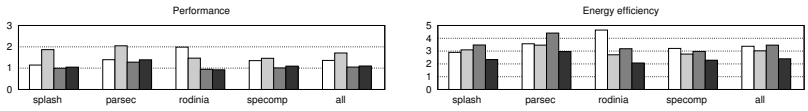

Figure 1: Average improvements per benchmark suite for the four 22 nm architecture design points over the 45 nm Nehalem baseline machine in terms of performance, energy efficiency and energy-delay product.

- The *low-frequency* design point reduces clock frequency and operating voltage which enables integrating 16 cores in a smaller power envelope. We assume 16 MB (2 times 8 MB) in total for the L3 cache in order to reduce off-chip memory bandwidth pressure.

- The *dual-issue* design point replaces the 4-wide out-of-order cores with 16 less aggressive dual-issue cores. Reducing cache sizes compared to the 8-core architecture allows for integrating twice the number of cores at a slight increase in chip area.

4 Results

Figure 1 summarizes the average improvements per benchmark suite for the four 22 nm architecture design points over the 45 nm baseline architecture in terms of performance, energy efficiency and energy-delay product (EDP). (Energy consumption in these results includes both dynamic and static energy consumption as reported by Sniper/McPAT.) Whereas the 3D design point yields the highest absolute improvement in performance, its power consumption is rather high so it does not lead to the best architecture when energy consumption is taken into account. Instead, when optimizing for energy, the low-frequency design point is the optimum configuration for this set of benchmarks.

The high performance of the 3D architecture is especially apparent for benchmarks that are DRAM bandwidth bound. One such example is `S-ocean.cont`, see Figure 2 (left) for cycle and energy stacks. In other applications with a moderate working set size, the 3D architecture suffers from the absence of an L3 cache.

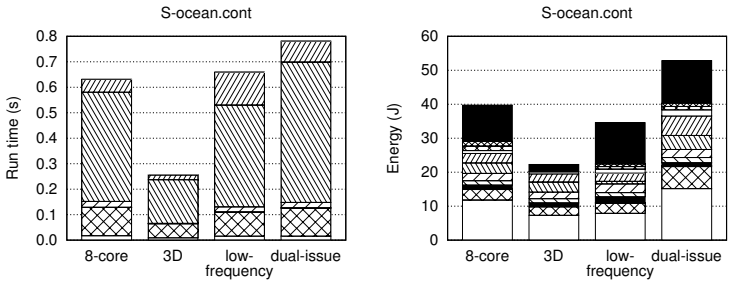

Figure 2: Time and energy stacks for `ocean.cont` from SPLASH-2.

Out of the four benchmark suites, Rodinia is the one that stands out by having poor performance on all of the 16-core architectures. Even though Rodinia is written with GPUs in mind, which have many small cores, the Rodinia benchmarks do not seem to parallelize very well on a multi-core CPU environment. One problem is that the data sets are not very large, which makes them fit in the caches — removing the benefit the 3D design point had.

We conclude that the 3D architecture has the highest performance on our selection of multi-threaded workloads. The high bandwidth that the 3D stacked memory architecture can provide reduces the need for extremely large caches, and allows a larger fraction of chip area to be used for cores. This is clearly beneficial for compute-intensive applications, which shows that 3D stacked memory can be an interesting alternative for more than just memory-bound applications. The conservative 8-core architecture combines good all-round performance with reasonable power consumption, as many of the applications have synchronization or data sharing problems that prevent them from properly making use of any of the 16-core architectures. When considering energy efficiency, the low-frequency architecture usually performs better than the 8-core design point — although the low-frequency architecture's absolute performance is lower making it less appropriate when considering derived metrics such as EDP. The dual-issue architecure, for this collection of benchmarks, does not seem to be an interesting choice at the 22 nm technology node.

References

[CHE11] Trevor E. Carlson, Wim Heirman, and Lieven Eeckhout. Sniper: Exploring the level of abstraction for scalable and accurate parallel multi-core simulations. In *International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, November 2011.

[LAS+09] Sheng Li, Jung Ho Ahn, Richard D. Strong, Jay B. Brockman, Dean M. Tullsen, and Norman P. Jouppi. McPAT: An integrated power, area, and timing modeling framework for multicore and manycore architectures. In *Proceedings of the 42nd Annual IEEE/ACM International Symposium on Microarchitecture*, MICRO 42, pages 469–480. ACM, December 2009.