The Path Forward: Specialized Computing in the Datacenter

Nikos Hardavellas*, Michael Ferdman†‡, Anastasia Ailamaki‡, Babak Falsafi‡

*Department of Electrical Engineering and Computer Science, Northwestern University
†Department of Electrical and Computer Engineering, Carnegie Mellon University
‡School of Computer and Communication Sciences, École Polytechnique Fédérale de Lausanne
{nikos@northwestern.edu, mferdman@ece.cmu.edu, anastasia.ailamaki@epfl.ch, babak.falsafi@epfl.ch}

ABSTRACT

Popular belief holds that the cores on chip will grow at an exponential rate, following Moore's Law, with a commensurate increase in performance. However, by exploring the design space of multicore chips across technologies under a large array of design parameters, we observe that physical constraints in power and off-chip bandwidth prohibit such performance increase. This leads us to conclude that server chips will not scale beyond a few tens of cores, potentially leaving the die real-estate underutilized in future technology generations. We observe that heterogeneous multicores can leverage the die area to overcome the initial power barrier, delivering significantly higher performance for the same off-chip bandwidth and power envelopes. Thus, specialized computing, especially when coupled with emerging memory technologies, promises significant increases in performance and energy-efficiency compared to general-purpose computing in the datacenter.

INTRODUCTION

As Moore's Law continues for at least another decade, the number of cores on chip will continue to grow at an exponential rate. While workloads with limited parallelism pose performance challenges with chip multiprocessors (CMPs), server workloads with abundant parallelism are believed to be immune, capable of scaling to the parallelism available in the hardware. However, contrary to popular belief, despite the inherent scalability in threaded server workloads, increasing core counts cannot directly translate into performance improvements because chips are physically constrained in power and off-chip bandwidth.

Multicores are not a panacea for server processor designs. While Moore's Law enables more transistors on chip [4], the static power consumption of the additional transistors can no longer be mitigated through circuit-level techniques [1]. Although a trade-off exists between cache performance and leakage power, the cache latency cannot be sufficiently reduced to deliver reasonable performance and simultaneously limit the leakage power. Additionally, the multiplying core counts and thread contexts constitute a substantial fraction of the chip's transistors, steadily raising both static and dynamic core power consumption. While voltage-frequency scaling may lower the dynamic power of the cores and enable more cores on chip, static power dissipation and performance requirements impose a limit. Thus, despite the abundant parallelism present in server workloads, server multicore designs are rapidly approaching the power wall.

Considering a large array of design parameters, we construct detailed models which conform to ITRS projections of future manufacturing technologies. We jointly optimize supply and threshold voltage, clock frequency, core count, manufacturing process, cache size, and memory technology to conclude that, without a technological miracle, server CMPs will not scale beyond a few tens of cores due to physical power and off-chip bandwidth constraints, leaving the die real-estate underutilized. We observe that heterogeneous multicores, by reducing energy waste through specialization, can leverage the die area to overcome the initial power barrier, delivering significantly higher performance under the same physical constraints. Thus, specialized computing shows promise in improving the aggregate performance and energy efficiency of the datacenter. This is especially true when heterogeneous CMPs are coupled with emerging memory technologies, which mitigate the bandwidth wall and fully expose the CMP to the power wall.

METHODOLOGY

Complexity and run-time requirements make it impractical to rely on full-system simulation for a large-scale design-space exploration study. Instead, we rely on first-order analytical models of the dominant components, with parameters tuned through full-system simulation. Our algorithm uses the analytical models as constraints, always finding the core count and cache size of the peak-performing design.

We model CMPs across four fabrication technologies: 65nm, 45nm, 32nm (due in 2013) and 20nm (due in 2017). For each technology node, we utilize parameters and projections from the International Technology Roadmap for Semiconductors (ITRS) 2008 Edition [4]. In agreement with ITRS, we model bulk planar CMOS for the 65nm and 45nm nodes, ultra-thin-body fully-depleted MOSFETs for 32nm technology, and double-gate FinFETs for the 20nm node.

We model multicore processors running server workloads (i.e., TPC-C, TPC-H and SPECweb) with cores built in one of three ways: general purpose (GPP), embedded (EMB), or specialized (Ideal-P). GPP cores are similar to the cores in Sun UltraSPARC T1. We model 4-way multi-threaded scalar in-order cores, as similar cores have been shown to optimize performance for server workloads [2]. Because general-purpose cores consume an inordinate amount of power and area compared to embedded cores, we also evaluate cores similar to the ones in ARM11 MPCore. To evaluate the potential of heterogeneous multicores running server workloads, we also study cores with ASIC-like properties: Ideal-P cores deliver 20x the performance of a GPP core on 1/8th the power under control-intensive workloads [3]. A heterogeneous multicore processor will enable only the Ideal-P cores that most closely match the requirements of the available work, and use GPP cores for non-critical or complex/uncommon parts of the program, thereby exhibiting near-ASIC properties.

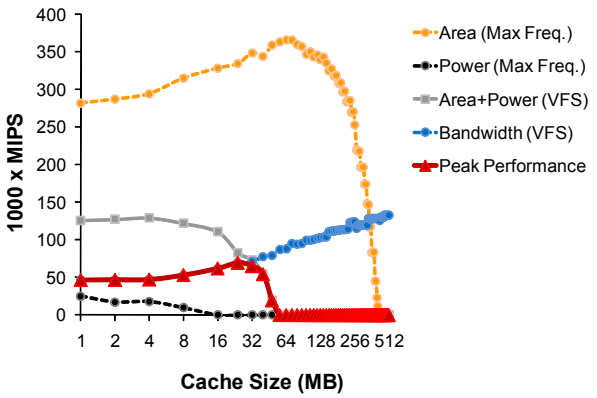

FIGURE 1: Performance of physically-constrained CMPs.

Figure 1 shows an example result of our model for GPP cores. The "Area" curve shows the performance of area-constrained designs operating at maximum frequency, assuming unlimited power and bandwidth. The "Power" curve shows power-constrained designs at maximum frequency, assuming unlimited area and bandwidth. The "Area+Power" curve shows voltage-frequency scaling (VFS), assuming unlimited bandwidth. Finally, the "Bandwidth" curve shows VFS designs subject only to bandwidth constraints. The "Peak Performance" design is initially bounded by bandwidth, eventually reaching the "Area+Power" VFS constraint at 32MB of cache.

RESULT HIGHLIGHTS

We find that EMB multicores exhibit trends similar to GPP multicores. The peak-performing designs are bandwidth-constrained at small cache sizes, becoming power-constrained for larger caches, with the highest performing designs at the intersection of the constraints. Both GPP and EMB designs require similar-sized caches to remain within the bandwidth envelope.

For peak performance, EMB multicores require twice the core count of GPP. Although additional cores deliver higher performance in 65nm technology, the higher core counts at smaller technologies provide a marginal performance benefit due to Amdahl's Law. While the best 20nm EMB design allows for 176 cores compared to 88 GPP cores, the EMB design trails 13% in absolute performance with a 99% parallel workload, achieving a speedup over GPP designs only with

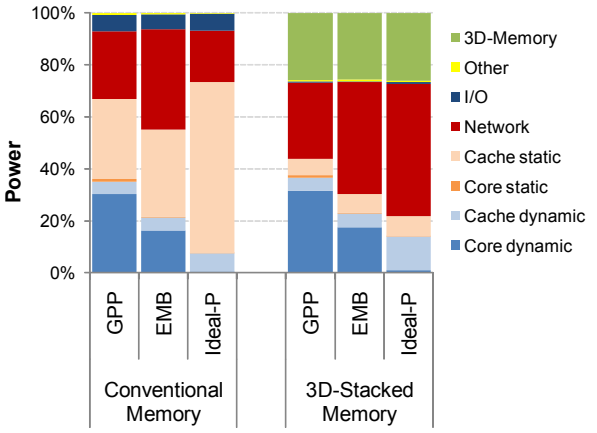

FIGURE 2: CMP power breakdown for TPC-C (20nm).

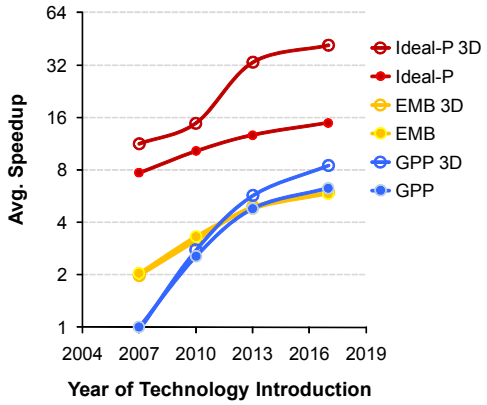

FIGURE 3: Speedup of GPP, EMB and Ideal-P CMPs.

99.6% or higher workload parallelism. Furthermore, higher core counts stress the interconnect, dissipating 68% more power than the interconnect of the GPP design (Figure 2). The EMB performance/watt is therefore similar to GPP designs, with the power efficiency of EMB cores outweighed by the power consumption of the interconnect.

While GPP and EMB designs are ultimately power-limited, the superior performance and power characteristics of Ideal-P cores results in Ideal-P CMPs achieving 3-7x higher performance with only 1/3rd to 1/4th of the cores required by their GPP and EMB counterparts (Figure 3). The remaining die real-estate can be utilized to implement a diverse collection of specialized cores, to increase the likelihood that a core matches the requirements of the available work. Moreover, the low running core count of Ideal-P designs increases performance even for workloads with relatively low parallelism, providing a much-needed respite from Amdahl's Law.

With all core designs we study, the use of a large 3D-stacked memory alleviates the off-chip bandwidth wall for most memory accesses, even when accounting for the exponentially growing application datasets. This allows for 2-3x more cores on average compared to bandwidth-limited designs, with a corresponding increase in performance (Figure 3).

In conclusion, specialized computing can leverage the die area to overcome the initial power barrier, reducing energy waste through specialization and delivering significantly higher performance and energy-efficiency in the datacenter.

REFERENCES

[1] S. Borkar. Microarchitecture and design challenges for gigascale integration. In Proceedings of the 37th Annual International Symposium on Microarchitecture, 2004.

[2] J. D. Davis, J. Laudon, and K. Olukotun. Maximizing CMP throughput with mediocre cores. In Proceedings of the 13th International Conference on Parallel Architectures and Compilation Techniques, 2005.

[3] R. Hameed, W. Qadeer, M. Wachs, O. Azizi, A. Solomatnikov, B.C. Lee, S. Richardson, C. Kozyrakis, and M. Horowitz. Understanding sources of inefficiency in general-purpose chips. In Proceedings of the 37th International Symposium on Computer Architecture, 2010.

[4] Semiconductor Industry Association. The international technology roadmap for semiconductors (ITRS), 2008.