TOWARD DARK SILICON IN SERVERS

SERVER CHIPS WILL NOT SCALE BEYOND A FEW TENS TO LOW HUNDREDS OF CORES, AND AN INCREASING FRACTION OF THE CHIP IN FUTURE TECHNOLOGIES WILL BE DARK SILICON THAT WE CANNOT AFFORD TO POWER. SPECIALIZED MULTICORE PROCESSORS, HOWEVER, CAN LEVERAGE THE UNDERUTILIZED DIE AREA TO OVERCOME THE INITIAL POWER BARRIER, DELIVERING SIGNIFICANTLY HIGHER PERFORMANCE FOR THE SAME BANDWIDTH AND POWER ENVELOPES.

Nikos Hardavellas
Northwestern University
Michael Ferdman
Carnegie Mellon University
Babak Falsafi
Anastasia Ailamaki
École Polytechnique Fédérale de Lausanne

Although 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. Contrary to popular belief, however, CMPs are not a panacea for server processor designs. Despite the inherent scalability in threaded server workloads, increasing core counts can't directly translate into performance improvements because chips are physically constrained in power and off-chip bandwidth.

Whereas transistor counts grow exponentially following Moore's law, the transistor threshold and supply voltages do not scale commensurately,[1] and the power consumption of the additional transistors can no longer be mitigated through circuit-level techniques. Although a trade-off exists between performance and power consumption, the transistors' switching speed and frequency cannot be reduced sufficiently to keep at bay the power consumption of exponentially more transistors and simultaneously deliver reasonable performance. The multiplying core counts constitute a substantial fraction of the chip's transistors, contributing to both dynamic and static power. Voltage-frequency scaling (VFS) might lower the dynamic power, but static power dissipation and performance requirements impose a limit. At the same time, the lethargic drop of supply voltages and the shrinking range of operational voltage (in the last decade, the difference between V_{th} and V_{dd} narrowed by 70 percent[1]) dampen the impact of VFS.

Even if the core power limitations could be eluded through highly efficient core designs or low-operational-power transistors, the rising core and thread counts drastically increase pressure on the limited and nonscalable off-chip bandwidth, encountering the bandwidth wall. Traditional approaches to alleviate off-chip bandwidth pressure call for larger on-chip caches, which further drive up the chip's power consumption, reducing the power available to the cores. Thus, without a technological miracle to overcome the power constraints imposed by thermal cooling and power delivery, we will soon inevitably enter an era of *dark silicon*, building dense and fast devices that we can't afford to power.

In this work, we show how we can use the abundant, power-constrained, and underutilized die real estate effectively to improve server performance and power efficiency by populating the die area with a large, diverse array of application-specific heterogeneous cores. These specialized CMPs can achieve peak performance and power efficiency by dynamically powering up only a small number of cores specifically designed for the given workload, with all but the most application-specific cores remaining dark (dynamically disabled) when not in use. Specialized CMPs

Published by the IEEE Computer Society

show promise in improving the aggregate performance and energy efficiency of server workloads. This is especially true if the specialized CMPs are coupled with emerging memory technologies, which mitigate the off-chip bandwidth wall and fully expose the processor to the power constraints.

Methodology

We consider aggregate server throughput as the performance metric. To design CMPs that attain peak performance while staying within the physical constraints, it is imperative to jointly optimize a large number of design parameters. Complexity and runtime requirements make it impractical to rely on simulation for a large-scale design-space-exploration study. Instead, we rely on first-order analytical models of the dominant components, relating the effects of technology-driven physical constraints to the performance of server workloads running on future CMPs.

We construct detailed parameterized models that conform to the *International Technology Roadmap for Semiconductors* (*ITRS*, www.itrs.net) projections of future manufacturing technologies. Using the analytical models as constraints, we derive peak-performance designs by jointly optimizing supply and threshold voltage, clock frequency, core count, manufacturing process, cache size, and memory technology.

Our models have been independently vetted and used in a recent study of heterogeneous computing.[2] Similar models were independently developed and validated against PARSEC benchmarks, demonstrating that multicore performance will not scale with technology.[3] We corroborate these results, propose specialized computing as a promising solution, and evaluate its potential.

Hardware model

To evaluate a range of core design choices, we model CMPs with cores built in one of three ways: general-purpose (GPP), embedded (EMB), or specialized (SP). GPP are scalar four-way multithreaded in-order cores modeled after Sun's UltraSPARC,[4] representing the class of modern CMPs targeted for the high-throughput server environment. A four-way multithreaded core achieves speedup of 1.7× over a single-threaded core.[5] EMB are advanced low-power embedded cores similar to the cores in ARM11 MPCore (www.arm.com/products/CPUs/ARM11MPCoreMultiprocessor.html), representing a power-conscious core design paradigm, without a fundamental change in processor performance when executing server workloads compared to GPP cores.

To estimate the performance and power efficiency of specialized cores, we evaluate hypothetical application-specific cores. Rather than representing a specific core design, we envision a CMP populated with specialized hardware such as GPUs, digital signal processors (DSPs), and field-programmable gate arrays (FPGAs) in addition to application-specific integrated circuits (ASICs) implementing common server operations. At any given time, only the subset of these hardware components that's best suited to execute efficiently the current workload would be powered up; we broadly use the label SP to represent a single powered-up hardware component within this design. Based on published data on general-purpose cores and ASICs running an optimized context-adaptive binary arithmetic coding (CABAC) segment of the H.264 video encoder,[6] we estimate that SP cores can deliver 20× the performance of a GPP core at 10× less power. This estimate is conservative because the CABAC segment is heavily control-intensive and hampers many hardware optimizations.

Each core is supported by 64-Kbyte Level 1 (L1) instruction and data caches and a shared unified nonuniform cache architecture (NUCA) L2 cache. We model a 2D torus interconnect for the L2 cache and separately optimize the cache size for each technology node with CACTI 6.0.[7] We do not evaluate deeper cache hierarchies because prior research shows that a NUCA organization outperforms any multilevel cache design.[8]

Technology model

Projecting from current technologies to future trends, we model CMPs across four fabrication technologies: 65 nm, 45 nm, 32 nm (due in 2013), and 20 nm (due in 2017). We scale technology parameters across technologies in accordance to the *ITRS* projections and model bulk planar CMOS for the 65- and 45-nm nodes, ultra-thin-body fully-depleted metal-oxide-semiconductor

Table 1. Workload and miss-rate model parameters.

Workload	Description	α	β	γ	Mean error (%)	Maximum error (%)
Online transaction processing (OLTP)	TPC-C v3.0 on IBM DB2 v8 ESE, 100 warehouses (10 Gbytes), 64 clients, 2-Gbyte buffer pool	0.5785	0.4750	−0.589	1.3	8.2
Decision support systems (DSS)	TPC-H throughput test on IBM DB2 v8 ESE, Queries 2, 6, 13, 16, 480-Mbyte buffer pool, 1-Gbyte database, 16 clients	0.5925	0.5154	−0.327	0.5	6.5
Web server (Apache)	SPECweb99 on Apache HTTP Server v2.0, 16,000 connections, fast common gateway interface (fastCGI), worker threading model	1.0081	2.1104	−0.503	1.2	4.9

field-effect transistors (MOSFETs) at 32 nm, and double-gate FinFETs at 20 nm.

To mitigate the power wall, we evaluate lowering the leakage current by using high-V_{th} transistors. These low-operational-power (LOP) transistors experience orders of magnitude lower subthreshold leakage current, while achieving 54 to 68 percent of the switching speed of high-performance (HP) transistors. We explore CMPs that use high-performance transistors for the entire chip, high-performance transistors for the cores and LOP for the cache, and LOP transistors for the entire chip.

Area model

Based on the *ITRS* projections, we model a 310 mm^2 die. Our algorithms eliminate all candidate designs that exceed the 310-mm^2 die area. We proportionally allocate 72 percent of the die for cores and cache,[9] with the remaining area allocated to the interconnect and system-on-chip (SoC) components. We estimate the core area by scaling existing designs.

For GPP cores, we scale an UltraSPARC T1 core,[4] measuring 13.67 mm^2 at 65 nm. For EMB cores, we scale an ARM11 core, measuring 2.48 mm^2 at 65 nm. The SP core area is equal to the area of an EMB core in our model. We use the *ITRS* to estimate the area required for an ECC-protected L2 cache and scale the cores and caches across technologies.

Performance model

Our performance model is based on Amdahl's law, assuming 99-percent application parallelism unless otherwise noted. Because of space constraints, we present only a summary of the model in this text; the exact formulas and derivations used for the presented results are available elsewhere.[10,11]

We estimate the performance of a single core by calculating the aggregate number of user instructions committed per cycle (UIPC), which is proportional to overall server-system throughput.[12] We compute the UIPC by accounting for the memory access latency as a function of load/store frequency and cache miss rates, empirically measuring the fraction of load/store instructions and the L1 miss rate of each application using a 16-core full-system simulation in Flexus.[12] We use the L2 access latency reported by CACTI 6.0.[7] Memory-access latency is projected using 7-percent annual improvement in DRAM latency, starting with 53 ns in 2007 (PC-667 at 65 nm).

Table 1 presents our workloads. Unless noted otherwise, the results are averaged across all workloads.

L2 cache miss rate and data-set evolution models

The cache miss rate plays a dominant role in performance. To estimate a workload's cache miss rate, we curve-fit empirical measurements across L2 cache sizes between 256 Kbytes and 64 Mbytes. We find that the x-shifted power law, $\alpha(x + \beta)^{\gamma}$, offers the closest fit for our data, having only a 1.3-percent average error across all measurements. Table 1 lists the miss-rate model parameters for each workload.

We provide the full miss-rate scaling formulas in a previous work,[10] including details on the curve fitting of miss rates and data-set growth projections based on the TPC-A, -B, -C,

and -E workloads. When considering designs in a future fabrication technology, we use the year of the technology's introduction to project the application data-set size with 29 percent annual growth. We adjusted cache miss rates based on the projected data-set size for each technology.

Off-chip bandwidth model

We model the chip bandwidth requirements by estimating the off-chip activity rate, scaled from the simulation measurements. The off-chip bandwidth usage is proportional to the L2 miss rate, the number of cores, and the activity of the cores (such as clock frequency and core performance). The maximum available bandwidth is calculated on the basis of the number of pads and the maximum off-chip clock as provided by the *ITRS* for each technology. Based on the *ITRS*, two-thirds of the pads are used for power and ground, leaving at most one-third available for off-chip memory signaling. Our algorithms discard candidate CMP designs with a bandwidth utilization that exceeds available bandwidth.

Additionally, we evaluate CMPs that use 3D-stacked memory[13] as a high-capacity high-bandwidth L3 cache. We treat 3D-stacked memory as a large L3 cache because the memory it houses is insufficient for a large-scale server software installation. We model a 3D-stacked memory where each layer has a capacity of 8 Gbits at 45-nm technology.[13] The worst-case power consumption for each 8-Gbit layer is 3.7 W.[11] We model eight layers, for a total of 8 Gbytes, with an additional control/logic layer, increasing the average chip temperature by an estimated 10°C.[13] When we evaluate 3D-stacked CMPs, we account for the effects of the increased temperature on power dissipation. We estimate that memory access time improves by 32.5 percent using 3D stacking due to the more efficient communication between the cores and 3D memory.[11]

We compute L3 miss rates similar to L2. We present 3D-stacked DRAM as a case study; photonics and other emerging technologies for mitigating the off-chip bandwidth wall are expected to exhibit similar trends.

Power model

We use a first-order power model to compute the total chip power by summing the dynamic and static power of the individual components (cores, cache, interconnect, I/O, and SoC components). For each technology node we evaluate, we use the *ITRS* data to dictate the maximum allowable power for air-cooled chips with heat sink for that technology. The power limits are used as an input to the model to automatically discard all candidate CMP designs that exceed the power limits published by the *ITRS* for that technology.

Based on the *ITRS* projections, power delivery to the die will pose an additional constraint owing to poor signal integrity when large currents are delivered at low voltages. If alternative cooling technologies, such as liquid cooling, were employed, the power delivery would impose a CMP design constraint for which no mainstream solutions have been proposed to date. We therefore focus on air-cooled systems in this study and consider power limits based only on thermal cooling rather than on power delivery.

We use the Sun UltraSPARC,[4] ARM11 MPCore, and 10 percent of the UltraSPARC core as reference points for the dynamic power model of the GPP, EMB, and SP cores, respectively. We compute the dynamic power of N cores by scaling the reference core's dynamic power P_H proportionally to the gate capacitance of the target technology, the target frequency, and the supply voltage square:

$$P_{D,Ncores} = N \times P_H \times \frac{\text{Gate Capacitance } (R)}{\text{Gate Capacitance } (R_H)} \times \frac{\left(f_{V_{dd}} \times V_{dd}(R)\right)^2}{\left(V_{dd,H}\right)^2} \times \frac{F}{F_H}$$

$V_{dd,H}$ and F_H represent the supply voltage and frequency of the reference core H in technology node R_H, and $V_{dd}(R)$ is the nominal supply voltage of technology R. We use $f_{V_{dd}}$ to perform voltage-frequency scaling, trading off clock frequency for power. At temperature T°K, the supply voltage scaling is estimated by $2.3 \times V_{th}(R, T) \leq f_{V_{dd}} \times V_{dd}(R) \leq V_{dd}(R)$. We quantize the scaling factor $f_{V_{dd}}$ in steps of 10 percent. The frequency F is scaled with the supply voltage, such that $F \leq F_{max}(R) \times f_{Fmax}(f_{V_{dd}})$. We account for the nonlinear relationship between frequency and voltage by fitting published data to compute $f_{Fmax}(\,)$.[14]

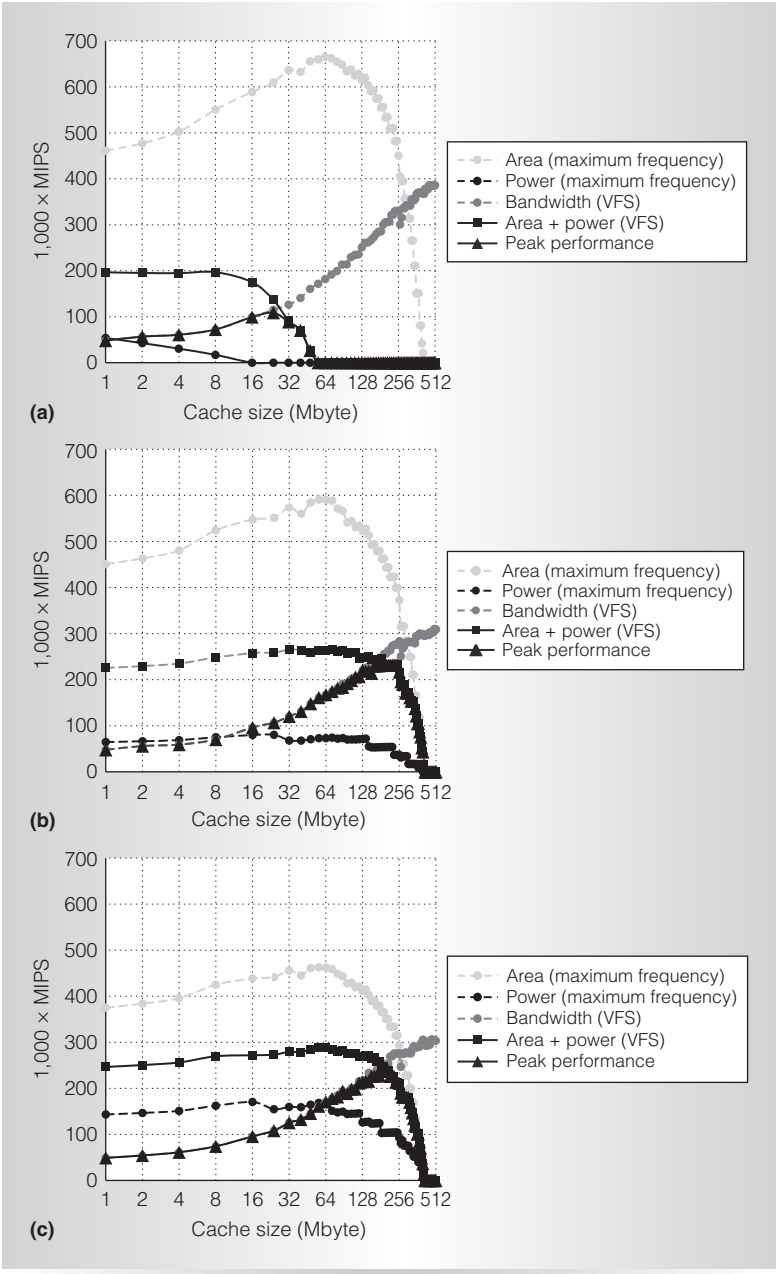

Figure 1. Performance of general-purpose (GPP) chip multiprocessors (CMPs) running Apache, using high-performance (HP) transistors for both cores and cache (a), HP transistors for cores and low-operational-power (LOP) for cache (b), and LOP transistors for both cores and cache (c) at 20 nm. The peak-performing CMPs are bandwidth-constrained for small caches and power-constrained for large caches, with the optimal design point at the intersection of the two constraints. (VFS: voltage-frequency scaling.)

We estimate the L2 dynamic power by scaling published data for the Sun UltraSPARC T1 cache.[4] We scale the cache dynamic power across technologies similarly to the core power and adjust it proportionally to the cache access rate. We compute cache activity from the measured L1 miss rates, the core count, and the relative core performance. We compute the network dynamic power based on network activity, scaled over the same reference design. The network activity conservatively equals the cache activity, adjusted by the average hop count on a 2D-torus interconnect. We estimate the I/O subsystem power proportionally to the reference GPP core design, the L2 access rate by all participating cores, and the L2 miss rate, scaled across technologies similarly to core power. We cap the bandwidth to the maximum allowed at each technology, assuming that worst-case power is expended when all I/O pins are fully utilized. Exact formulas and constants used for the calculation of the dynamic power components are available in previous works.[10,11]

To estimate static power, we model only the subthreshold leakage because it will dominate gate and junction leakage in future technologies.[15] The cache's static power is directly proportional to its size, the supply voltage, the gate width, and the leakage current at the corresponding temperature. We estimate an average ratio of gate length to width of 3.0 across technologies and obtain gate lengths from the *ITRS*. We account for core leakage by estimating the number of transistors in a core using the *ITRS* density projections and assuming a 50-percent switching rate. We scale the leakage current to a target temperature by fitting the subthreshold coefficient.[16] We calculate leakage at 66°C, a typical operating temperature of today's chips.[4]

Analysis

We begin by explaining the progression of our peak-performance search algorithm on the basis of the results plotted in Figure 1a. The plot shows the aggregate chip performance as a function of the L2 cache size. The area curve shows the performance of designs that have unlimited power and off-chip bandwidth and are constrained only by the die area. We can leverage parallelism in these designs to achieve high performance by populating the entire die area with cores.

Following the area curve to the right, a larger portion of the chip area is dedicated

to the L2 cache. Although larger caches mean fewer cores, each core's performance is higher due to greater cache capacity, leading to higher aggregate chip performance up to approximately 64 Mbytes of L2 cache. The performance benefit beyond 64 Mbytes is outweighed by the cost of further reducing the core count, leading to an aggregate performance drop at larger cache sizes.

The power curve shows designs populated with cores running at the maximum frequency, with power limited due to air-cooling constraints, but having unlimited area and off-chip bandwidth. The high power of each core restricts these designs to a handful of cores, severely limiting the aggregate chip performance. Increasing cache size takes more power away from the already limited number of cores, dropping performance even further.

The bandwidth curve shows designs that are limited only in off-chip bandwidth, permitting unlimited area and power use. The core count and core frequency are jointly optimized to find the peak-performing configuration in light of the bandwidth constraint. Larger caches reduce off-chip bandwidth pressure, allowing the bandwidth-limited designs to achieve improved performance.

Conversely, the area+power curve shows designs limited in area and power but permitted to consume unlimited off-chip bandwidth. However, unlike the power curve, the area+power curve jointly optimizes the core count, voltage, and frequency, selecting the peak throughput design combination for each evaluated L2 cache size.

Finally, the peak performance curve follows the strictest constraint, showing the feasible CMP designs. At small cache sizes, the off-chip bandwidth serves as the performance-limiting factor. Beyond 20 Mbytes, however, the power consumed by the L2 cache restricts the number of cores, penalizing performance. Therefore, we conclude that the peak-performing GPP design at 20 nm using HP transistors should use approximately 20 Mbytes of L2 cache with the remaining power budget utilized by cores. Moreover, the gap between the peak performance curve and the area curve at 20 Mbytes cache indicates that the best possible 20-nm GPP design with HP transistors cannot use the majority of the die area for more cores, because that area can't be powered up.

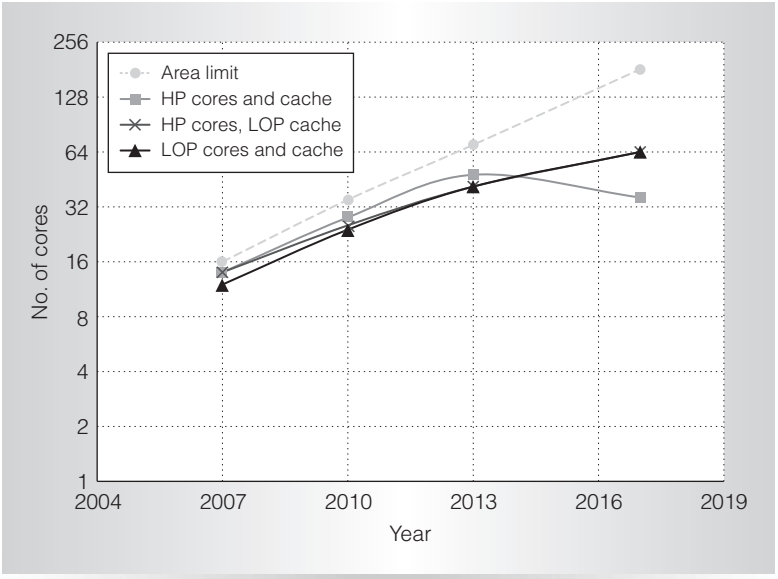

Figure 2. Core counts for peak-performance GPP CMPs using HP transistors for both cores and cache, HP transistors for cores and LOP for cache, and LOP transistors for both cores and cache. The gap between the LOP designs and the die-area limit suggests that an increasing fraction of the die area cannot be utilized by cores.

Figures 1b and 1c extend the analysis to designs that use slower low-leakage transistors. Because designs using only HP transistors are severely power limited, they can power only 20 percent of the cores that fit in the die at 20 nm. Using LOP transistors for the cache (see Figure 1b) enables larger caches that can support twice the number of cores and yield higher performance, with core leakage accounting for 20 percent of the chip power. Due to power constraints, the peak-performance designs must employ clocks at least 43 percent lower than the maximum frequency supported by the technology. Although LOP transistors are slower than HP transistors, they retain 54 to 68 percent of the maximum switching speed according to *ITRS*. As such, LOP devices are suitable to implement both the cores and the cache, yielding higher power efficiency (see Figure 1c). Ultimately, however, even after optimizing transistor types, peak-performance CMPs can power less than 35 percent of the cores that could fit on die.

Using an identical analysis, we find the highest performance design feasible for each technology generation. Figure 2 plots the core counts of the peak-performing designs

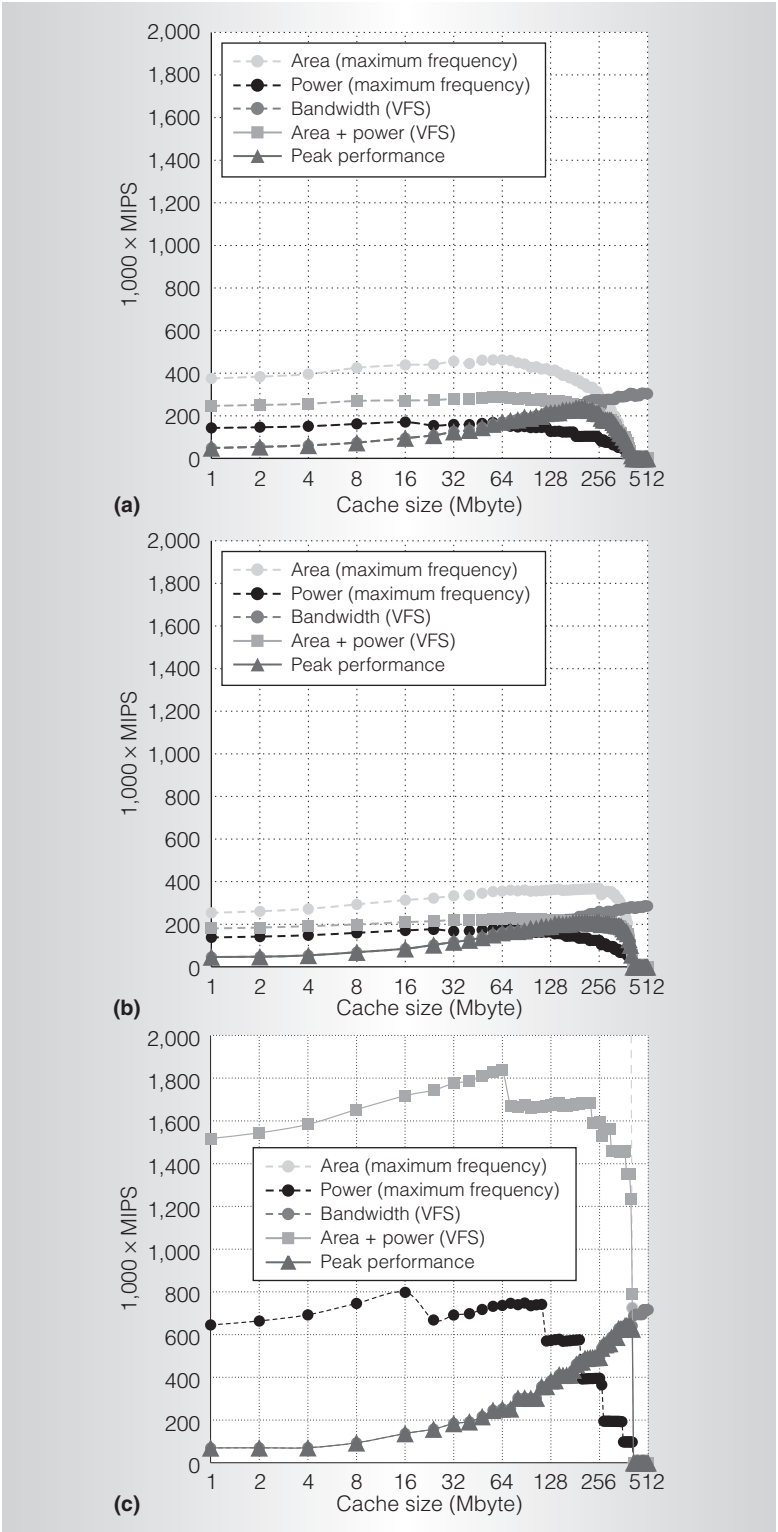

Figure 3. Performance of CMPs with GPP cores (a), embedded (EMB) cores (b), and specialized (SP) cores (c) using LOP transistors at 20 nm. SP designs significantly outperform GPP and EMB designs because SP cores are highly power-efficient.

as well as the theoretical number of cores that could fit into the die area at the corresponding technology. Beyond 2013 (32 nm), HP-based designs become impractical due to the chip power limits. Although LOP-based designs offer a way forward, the large gap that emerges between the LOP designs and the die-area limit suggests that either die sizes will shrink or a large portion of the chip silicon will have to remain dark (powered down).

Multicore processors with milliwatt cores

Lean cores deliver high performance at low power; for example, the UltraSPARC T1 consumes 2 W per thread.[4] However, embedded cores can deliver reasonable performance at orders of magnitude lower power; for example, 279 mW for ARM1176JZ(F)-S with an eight-stage out-of-order pipeline and dynamic branch prediction. Because of their small size and power efficiency, CMPs employing EMB cores can fit many more cores within the physical constraints. We find that EMB-based multicore CMPs generally exhibit trends similar to GPP-based multicore CMPs (see Figure 3).

Both GPP and EMB designs require similar-sized caches to remain within the bandwidth envelope. But to reach peak performance, EMB multicores require double the cores compared to GPP. However, the high core count provides only a marginal performance benefit because of Amdahl's law and the increased power consumption of the larger on-chip interconnect. As a result, the best EMB design trails GPP by 13 percent in absolute performance with a 99-percent parallel workload, achieving a speedup over GPP with 99.6 percent or higher workload parallelism.

We also evaluated multithreaded EMB cores and found they behave similarly to single-threaded cores due to the increased power and bandwidth requirements. Ultimately, peak-performance designs with EMB cores can power only 12 percent of the cores that fit in the die (see Figure 4), potentially leaving a large chip area underutilized.

Specialized multicore processors

Amdahl's law prohibits large core counts from delivering high aggregate performance

(except for embarrassingly-parallel applications). An alternative design is to deliver higher performance with fewer but more powerful cores. We evaluated an extreme application of this approach by considering specialized computing, where a multicore chip might contain hundreds of diverse application-specific cores, activating only those cores that are most useful to the running application while leaving the vast majority of the on-chip cores powered down. The few cores that are simultaneously powered up in this design reduce the impact of Amdahl's law on aggregate performance, whereas matching specialized cores to an application's requirements enables high performance at high power efficiency.

Figure 4 compares the number of cores for the peak-performing designs across the studied core types and process generations. We found that the peak-performance SP designs employ only 16 to 32 cores, with a large fraction of the chip die area occupied by a cache. For our workloads, the observation of low-core-count SP designs outperforming alternative designs holds up to 99.9 percent parallelism.

The superior power and performance characteristics of SP cores push the power envelope much further than is possible with other core designs. As a result, SP multicores attain 2× to 12× speedup over the GPP and EMB designs and are ultimately constrained by the limited off-chip bandwidth (see Figure 5). The performance improvement achieved by SP multicores on server workloads is in line with prior research on mobile applications.[17]

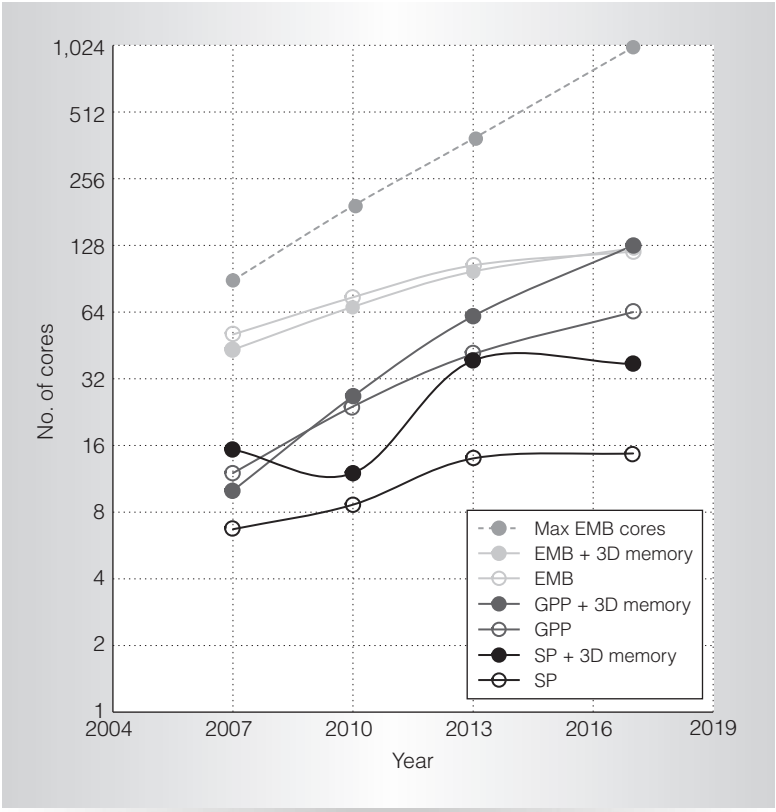

Figure 4. Core counts for peak-performance GPP, EMB, and SP CMPs, with conventional or 3D-stacked memory. GPP and EMB designs scale up to only a few tens to low hundreds of cores, although an order of magnitude more cores fit in the die. SP designs require only a handful of cores to attain peak performance.

Effect of bandwidth-mitigating technologies

Bandwidth considerations push cache sizes up, reducing the power available to employ more or faster cores for all core types. We evaluated 3D-stacked DRAM caches to observe the trends of future processor designs in light of technologies that might alleviate off-chip bandwidth pressure for future processors. We expect that other bandwidth-mitigating technologies (such as photonics) will exhibit similar trends.

A 3D-stacked memory cache pushes the bandwidth constraint beyond the power limits, leading to designs that are only power-constrained and achieve higher performance (see Figure 5). Eliminating the off-chip bandwidth bottleneck pushes designs back to the power-limited regime where the die area is underutilized due to an inability to power up all cores (see Figure 4). GPP and EMB CMPs attain only a modest performance improvement (less than 35 percent).

However, the reduction in off-chip bandwidth requirements when combining 3D memory with specialized cores results in significant speedup (3× at 20 nm) and relieves the pressure on the on-chip cache size. As a result, peak-performance designs with SP cores can be realized in an increasingly smaller silicon area, with the otherwise dark silicon used to implement a large collection of specialized cores to increase the likelihood of finding a core suitable for the current computation (see Figure 6).

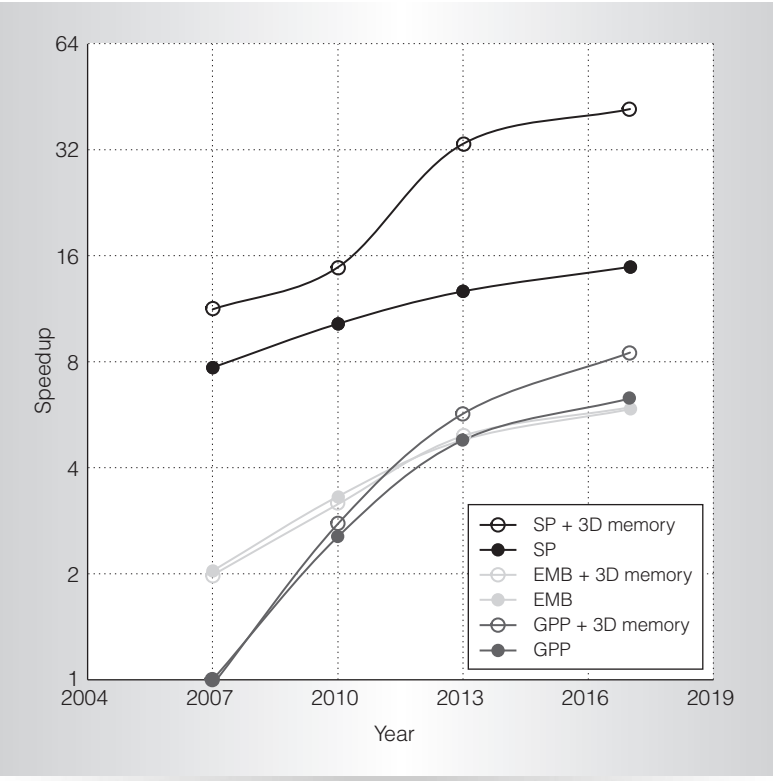

Figure 5. Speedup of peak-performance GPP, EMB, and SP CMPs using LOP transistors, with conventional and 3D-stacked memory. SP outperforms GPP and EMB designs by 2× to 12× across technologies.

As technology scaling continues, power constraints will prevent conventional multicore designs from scaling beyond a few tens to low hundreds of cores, leaving an increasing fraction of the die unused. Specialized multicores repurpose the unused dark silicon to implement a large number of workload-specific cores and power up only the few cores that most closely match the requirements of the executing workload. Although specialized multicores are an appealing design, further research is needed to realize them. We must characterize modern workloads to identify computational segments that are candidates for off-loading to specialized cores and devise core architectures suitable to execute them. Moreover, we must develop the software infrastructure and runtime environment that will facilitate code migration at the appropriate granularity. We plan to continue tackling these important issues and make specialized computing a reality. MICRO

References

1. Y. Watanabe, J.D. Davis, and D.A. Wood, ''Widget: Wisconsin Decoupled Grid Execution Tiles,'' *Proc. 37th Int'l Symp. Computer Architecture,* IEEE CS Press, 2010, pp. 2-13.
2. E.S. Chung et al., ''Single-Chip Heterogeneous Computing: Does the Future Include Custom Logic, FPGAs, and GPGPUs?'' *Proc. 43rd IEEE/ACM Int'l Symp. Microarchitecture,* IEEE CS Press, 2010, pp. 225-236.
3. H. Esmaeilzadeh et al., ''Dark Silicon and the End of Multicore Scaling,'' *Proc. 38th Int'l Symp. Computer Architecture,* ACM Press, 2011.
4. A.S. Leon et al., ''A Power-Efficient High-Throughput 32-Thread SPARC Processor,'' *IEEE J. Solid-State Circuits,* vol. 42, no. 1, 2007, pp. 7-16.
5. N. Hardavellas et al., ''Database Servers on Chip Multiprocessors: Limitations and Opportunities,'' *Proc. 3rd Biennial Conf. Innovative Data Systems Research,* 2007, pp. 79-87; www.cidrdb.org/cidr2007.
6. R. Hameed et al., ''Understanding Sources of Inefficiency in General-Purpose Chips,'' *Proc. 37th Int'l Symp. Computer Architecture,* IEEE CS Press, 2010, pp. 37-47.
7. N. Muralimanohar, R. Balasubramonian, and N. Jouppi, ''Optimizing NUCA Organizations and Wiring Alternatives for Large Caches with CACTI 6.0,'' *Proc. 40th IEEE/ACM Int'l Symp. Microarchitecture,* IEEE CS Press, 2007, pp. 3-14.
8. C. Kim, D. Burger, and S.W. Keckler, ''An Adaptive, Non-uniform Cache Structure for Wire-Delay Dominated On-Chip Caches,'' *ACM SIGPLAN Notices,* vol. 37, no. 10, 2002, pp. 211-222.
9. J.D. Davis, J. Laudon, and K. Olukotun, ''Maximizing CMP Throughput with Mediocre Cores,'' *Proc. 13th Int'l Conf. Parallel Architectures and Compilation Techniques,* IEEE CS Press, 2005, pp. 51-62.
10. N. Hardavellas, ''Chip Multiprocessors for Server Workloads,'' doctoral dissertation, Dept. of Computer Science, Carnegie Mellon Univ., 2009.
11. N. Hardavellas et al., ''Power Scaling: The Ultimate Obstacle to 1K-Core Chips,'' tech. report NWU-EECS-10-05, Northwestern Univ., 2010.

12. T.F. Wenisch et al., ''SimFlex: Statistical Sampling of Computer System Simulation,'' *IEEE Micro,* vol. 26, no. 4, 2006, pp. 18-31.
13. G.H. Loh, ''3D-Stacked Memory Architectures for Multi-core Processors,'' *Proc. 35th Int'l Symp. Computer Architecture,* IEEE CS Press, 2008, pp. 453-464.
14. T. Burd et al., ''A Dynamic Voltage Scaled Microprocessor System,'' *Proc. IEEE Int'l Solid-State Circuits Conf.,* IEEE CS Press, 2000, pp. 294-295.
15. S. Rodriguez and B. Jacob, ''Energy/Power Breakdown of Pipelined Nanometer Caches (90nm/65nm/45nm/32nm),'' *Proc. Int'l Symp. Low-Power Electronics and Design,* ACM Press, 2006, pp. 25-30.
16. H. Hua et al., ''Performance Trend in Three-Dimensional Integrated Circuits,'' *Proc. Int'l Interconnect Technology Conf.,* 2006, pp. 45-47.
17. N. Goulding-Hotta et al., ''The GreenDroid Mobile Application Processor: An Architecture for Silicon's Dark Future,'' *IEEE Micro,* vol. 31, no. 2, 2011, pp. 86-95.

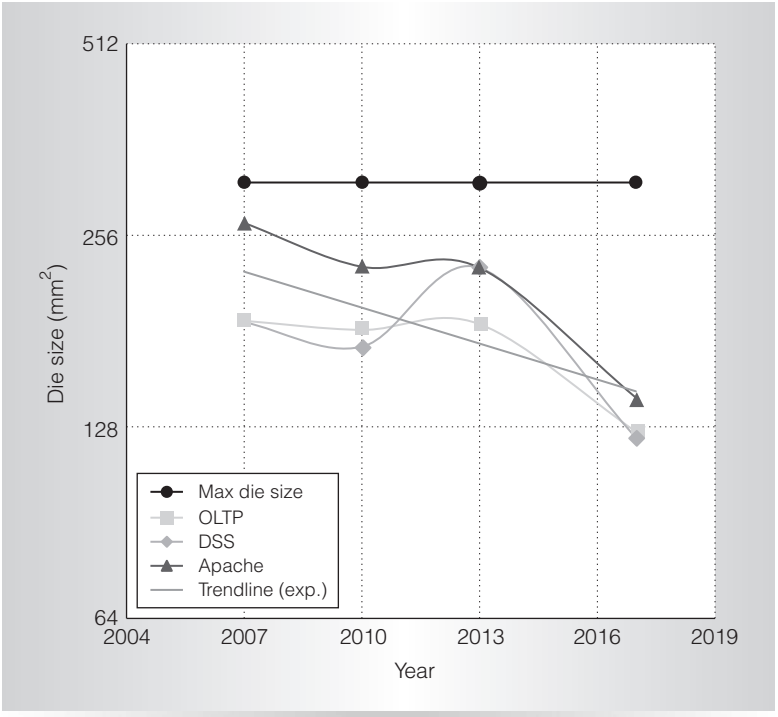

Figure 6. Die size of peak-performance SP CMPs with 3D-stacked memory. The gap between the trendline and the maximum die size indicates that an increasing fraction of the silicon area is left unutilized (dark). Instead of wasting it, specialized multicores repurpose it to implement application-specific cores.

Nikos Hardavellas is the June and Donald Brewer Assistant Professor of Electrical Engineering and Computer Science at Northwestern University, and the director of the Parallel Architecture Group at Northwestern (PARAG@N). His research interests are in hardware and software design for energy-efficient scalable parallel architectures, memory systems, and on-chip interconnects. Hardavellas has a PhD in computer science from Carnegie Mellon University. He's a member of the ACM and IEEE.

Michael Ferdman is a PhD candidate in electrical and computer engineering at Carnegie Mellon University. His research interests include computer architecture with an emphasis on proactive memory system design. Ferdman has an MS in electrical and computer engineering from Carnegie Mellon University. He's a student member of the ACM and IEEE.

Babak Falsafi is a professor of computer and communication sciences at École Polytechnique Fédérale de Lausanne, where he directs the EcoCloud center targeting robust, economic, and environmentally friendly cloud technologies. Falsafi has a PhD in computer science from the University of Wisconsin-Madison. He is a senior member of the ACM and IEEE.

Anastasia Ailamaki is a professor of computer science at École Polytechnique Fédérale de Lausanne. Her research interests are in optimizing database workloads for modern hardware and disks and in managing large data sets for scientific applications. Ailamaki has a PhD in computer science from the University of Wisconsin-Madison.

Direct questions or comments about this article to Nikos Hardavellas, Northwestern University, Technological Institute—EECS, 2145 Sheridan Rd., Evanston, IL 60208; nikos@northwestern.edu.