Accepted for publication in IEEE Micro, Special Issue on Energy-Aware Computing, Sep/Oct 2012

Optimizing Datacenter TCO with Scale-Out Processors

Boris Grot[1] Damien Hardy[2] Pejman Lotfi-Kamran[1]
Chrysostomos Nicopoulos [2] Yanos Sazeides[2] Babak Falsafi[1]

[1]EPFL [2]University of Cyprus

Abstract

Large-scale datacenters use performance and total cost of ownership (TCO) as key optimization metrics. Conventional server processors, designed for a wide range of workloads, fall short in their ability to maximize these metrics due to inefficient utilization of their area and energy budgets. Emerging scale-out workloads are memory-intensive and require processors that can take full advantage of the available memory bandwidth and capacity to maximize the performance per TCO. While recently introduced processors based on low-power cores improve both throughput and energy-efficiency as compared to conventional server chips, we show that a specialized *Scale-Out Processor* architecture that maximizes on-chip computational density delivers the highest performace/TCO and performance/Watt at the datacenter level.

1 Motivation

Our world is in the midst of an information revolution, driven by ubiquitous access to vast data stores via a variety of richly-networked platforms. Datacenters are the workhorses powering this revolution. Companies leading the transformation to the digital universe, such as Google, Microsoft, and Facebook, rely on networks of mega-scale datacenters to provide search, social connectivity, media streaming, and a growing number of other offerings to large, distributed audiences. A scale-out datacenter powering cloud services houses tens of thousands of servers, necessary for high scalability, availability, and resilience [8].

The massive scale of such datacenters requires an enormous capital outlay for infrastructure and hardware, often exceeding $100 million per datacenter [14]. Similarly expansive are the power requirements, typically in the range of 5-15MW per datacenter, totaling millions of dollars in annual operating costs. With demand for information services skyrocketing around the globe, efficiency has become a paramount concern in the design and operation of large-scale datacenters.

In order to reduce infrastructure, hardware, and energy costs, datacenter operators target high compute density and power efficiency. TCO is an optimization metric that considers the costs of real-estate, power delivery and cooling infrastructure, hardware acquisition costs, and operating expenses. Because server acquisition and power costs constitute the two largest TCO components [5], servers present a prime optimization target in the quest for more efficient datacenters. In addition to cost, performance is also of paramount importance in scale-out datacenters designed to service thousands of concurrent requests with real-time constraints. The ratio of performance to TCO (performance per dollar of ownership expense) is thus an appropriate metric for evaluating different datacenter designs.

Table 1: Server chip characteristics.

Processor	Type	Cores/ Threads	LLC size (MB)	DDR3 interfaces	Freq (GHz)	Power (W)	Area (mm^2)	Cost ($)
Existing designs								
Big-core big-chip	Conventional	6/12	12	3	3	95	233	800
Small-core small-chip	Small-chip	4/4	4	1	1.5	6	62	95
Small-core big-chip	Tiled	36/36	9	2	1.5	28	132	300
Proposed designs								
Scale-Out in-order	SOP	48/48	4	3	1.5	34	132	320
Scale-Out out-of-order	SOP	16/16	4	2	2	33	132	320

Scale-out workloads prevalent in large-scale datacenters rely on in-memory processing and massive parallelism to guarantee low response latency and high throughput. While processors ultimately determine a server's performance characteristics, they contribute just a fraction of the overall purchase price and power burden in a server node. Memory, disk, networking equipment, power provisioning and cooling all contribute substantially to acquisition and operating costs. Moreover, these components are less energy-proportional than modern processors, meaning that their power requirements do not scale down well as the server load drops. Thus, maximizing the benefit from the TCO investment requires getting high utilization from the entire server, not just the processor.

To achieve high server utilization, datacenters need to employ processors that can fully leverage the available bandwidth to memory and I/O. Conventional server processors use powerful cores designed for a broad range of workloads, including scientific, gaming, and media processing; as a result, they deliver good performance across the workload range, but fail to maximize either performance or efficiency on memory-intensive scale-out applications. Emerging server processors, on the other hand, employ simpler core microarchitectures that improve efficiency, but fall short of maximizing performance. What the industry needs are server processors that jointly optimize for performance, energy, and TCO. The Scale-Out Processor, which improves datacenter efficiency through a many-core organization tuned to the demands of scale-out workloads, represents a step in that direction.

2 Today's Server Processors

Multi-core processors common today are well-suited for massively parallel scale-out workloads running in datacenters for two reasons: (a) they improve throughput per chip over single-core designs; and (b) they amortize on-chip and board-level resources among multiple hardware threads, thereby lowering both cost and power consumption per unit of work (i.e., thread).

Table 1 summarizes principal characteristics of today's server processors. Existing datacenters are built with server-class designs from Intel and AMD. A representative processor [13] is Intel's Xeon 5670, a mid-range design that integrates six powerful dual-threaded cores, a spacious 12MB last-level cache (LLC), and consumes 95W at the maximum frequency of 3GHz. The combination of powerful cores and relatively large chip size leads us to classify conventional server processors as *big-core big-chip* designs.

Recently, several companies have introduced processors featuring simpler core microarchitec-

tures that are specifically targeted at scale-out datacenters. Research has shown simple-core designs to be well-matched to the demands of many scale-out workloads that spend a high fraction of their time accessing memory and have moderate computational intensity [2]. Two design paradigms have emerged in this space: one type features a few small cores on a small chip (*small-core small-chip*), while the other integrates a large number of cores on a bigger chip (*small-core big-chip*).

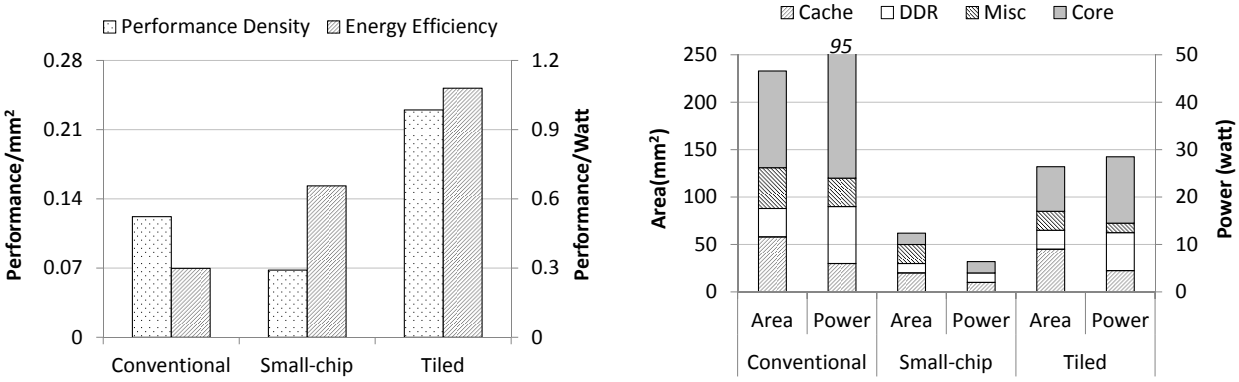
Companies including Calxeda, Marvell, and SeaMicro market small-core small-chip SoCs targeted at datacenters. Despite the differences in the core organization and even the ISA (Calxeda's and Marvell's designs are powered by ARM, while SeaMicro uses x86-based Atom processor), the chips are surprisingly similar in their feature set: all have four hardware contexts, dual-issue cores, clock speed in the range of 1.1 to 1.6GHz, and power consumption of 5-10W. We use the Calxeda design as a representative configuration, featuring four Cortex-A9 cores, 4MB LLC, and an on-die memory controller [3]. At 1.5GHz, our model estimates peak power consumption of 6W.

A processor representative of the small-core big-chip design philosophy is Tilera's Tile-Gx3036. This server-class processor features 36 simple cores and a 9MB LLC in a tiled organization [15]. Each tile integrates a core, a slice of the shared last-level cache, and a router. Accesses to the remote banks of the distributed LLC require a traversal of the on-chip interconnect, implemented as a two-dimensional mesh network with a single-cycle per-hop delay. Operating at 1.5GHz, we estimate the Tilera-like tiled design to draw 28W of power at peak load.

To understand efficiency implications of these diverse processor architectures, we use a combination of analytical models and simulation-based studies using a full-system server simulation infrastructure to estimate their performance, area, and power characteristics. Our workloads are taken from CloudSuite, a collection of representative scale-out applications that include web search, data serving, and MapReduce [1]. Details of the methodology can be found in Section 4.

Figure 1(a) compares the designs along two dimensions: performance density and energy-efficiency. Performance density, expressed as performance per mm^2, measures the processor's ability to effectively utilize the chip real-estate. Energy-efficiency, in units of performance per Watt, indicates the processor's ability to convert energy into useful work.

The small-core small-chip architecture offers a 2.2x improvement in energy-efficiency over a conventional big-core design thanks to the former's simpler core microarchitecture. However, small-chip has 45% lower performance-density than the conventional design. To better understand the



(a) Performance density and energy efficiency.

(b) Processor area and power breakdown.

Figure 1: Efficiency, area, and power of today's server processors.

trends, Figure 1(b) shows a breakdown of the respective processors' area and power budgets. The data in the figure reveals that while the cores in a conventional server processor take up 44% of the chip area, the small-core small-chip design commits just 20% of the chip to compute, with the remainder of the area going to the last-level cache, IO, and auxiliary circuitry. In terms of power, the six conventional cores consume 71W of the 95W power budget (75%), while the four simpler cores in the small-chip organization dissipate just 2.4W (38% of total chip power) under full load. As with the area, the relative energy cost of the cache and peripheral circuitry in the small-chip is greater than in the conventional design (62% and 25% of the respective chips' power budgets).

The most efficient design point is the small-core big-chip tiled processor, which surpasses both conventional and small-chip alternatives by over 88% on performance density and 65% on energy-efficiency. The cores in the tiled processor take up 36% of the chip real-estate, nearly doubling the fraction of the area dedicated to compute as compared to the small-chip design. The fraction of the power devoted to execution resources increases to 48% compared to 38% in small-chip.

Our results corroborate earlier studies that identify efficiency benefits stemming from the use of lower-complexity cores as compared to those used in conventional server processors [9, 11]. However, our findings also identify an important, yet unsurprising, trend; namely, the use of simpler cores by itself is insufficient to maximize processor efficiency and that the chip-level organization needs to be considered. More specifically, a larger chip that integrates many cores is necessary to amortize the area and power expense of un-core resources, such as cache and off-chip interfaces, by multiplexing them among the cores.

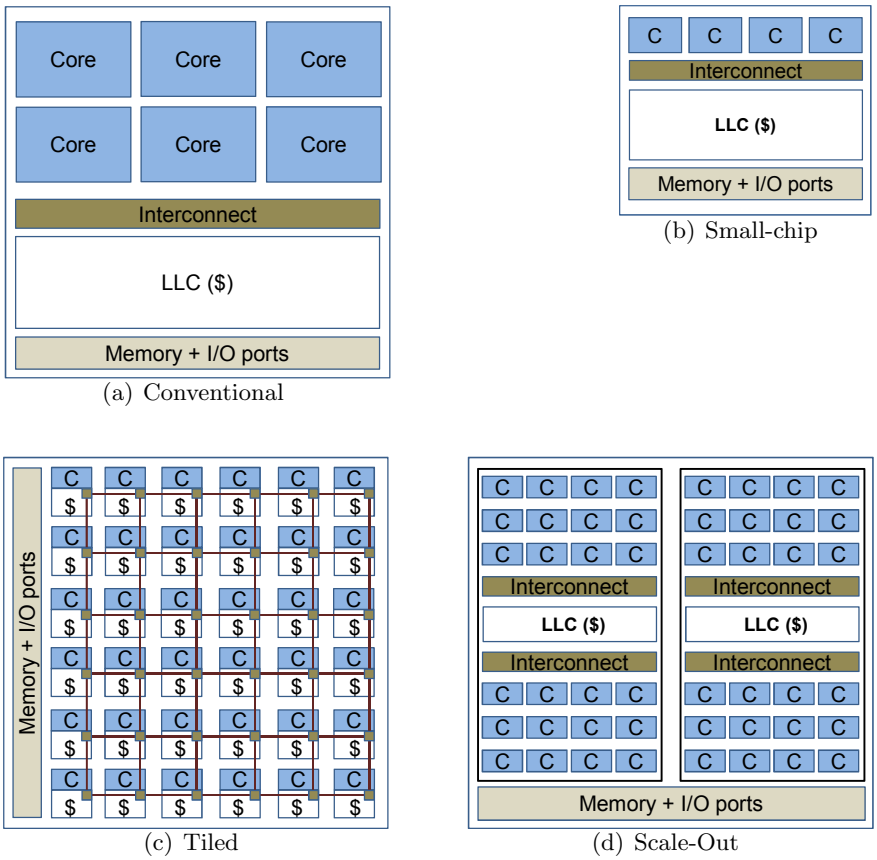
3 Scale-Out Processors

In an effort to maximize silicon efficiency on scale-out workloads, we have examined the characteristics of a suite of representative scale-out applications and the demands they place on processor resources. Our findings [6, 4] indicate that (a) large last-level caches are not beneficial for capturing the enormous data footprint of datacenter applications, (b) the active instruction footprint greatly exceeds the capacity of the L1 caches, but can be accommodated with a 2-4 MB secondary cache, and (c) scale-out workloads have virtually no thread-to-thread communication, requiring minimal on-chip coherence and communication infrastructure.

Driven by these observations, we developed a methodology for designing performance-density optimal server chips called *Scale-Out Processors* (SOPs) [12]. The SOP paradigm extends the small-core big-chip design space by optimizing the on-chip cache capacity, core count, and the number of off-chip memory interfaces in a way that maximizes computational density and throughput.

At the heart of a Scale-Out Processor is a coarse-grained building block called a *pod* – a stand-alone multi-core server. Each pod features a modestly sized 2-4 MB last-level cache for capturing the active instruction footprint and commonly accessed data structures. The small LLC size reduces the cache access time and leaves more chip area for compute. To further reduce the latency of performance-critical LLC accesses, SOPs use a high-bandwidth crossbar interconnect instead of a multi-hop point-to-point network. The number of cores in a pod is empirically chosen in a way that maximizes cache utilization without causing thrashing or penalizing interconnect area and delay.

Scalability in the SOP architecture is achieved by tiling at the pod granularity up to the available area, power, or memory bandwidth limit. The multiple pods share the off-chip interfaces to reduce cost and maximize bandwidth utilization. The pod-based tiling strategy reduces chip-level complexity and provides a technology-scalable architecture that preserves the optimality of each



(a) Conventional

(b) Small-chip

(c) Tiled

(d) Scale-Out

Figure 2: Comparison of server processor organizations.

pod across technology generations. Figure 2 compares the Scale-Out Processor chip architecture to conventional, small-chip, and tiled designs.

Compared to a tiled design, a Scale-Out Processor increases the number of cores integrated on chip by 33%, from 36 to 48, within the same area budget. This improvement in compute density is achieved by reducing the LLC capacity from 9 to 4 MB, freeing up chip area for the cores. The resulting SOP devotes 47% of the chip area to the cores, up from 36% in the tiled processor. Our evaluation shows that the per-core performance in the SOP design is comparable to that in a tiled one; while the smaller LLC capacity in the Scale-Out Processor has a dampening effect on single-threaded performance, the lower access delay in the crossbar-based SOP compensates for the higher miss ratio by accelerating fetches of performance-critical instructions. The bottom line is that the SOP design improves chip-level performance (i.e., throughput) by 33% over the tiled processor. Finally, the peak power consumption of the SOP design is higher than that of the tiled processor due to the former's greater on-chip compute capacity. However, as our results in Section 5 demonstrate, greater chip-level processing capability of the Scale-Out Processor is beneficial from a TCO perspective despite the increased power draw at the chip level.

Table 2: TCO parameters.

Parameter	Value
Rack dimensions (42U): width x depth x inter-rack space	0.6m x 1.2m x 1.2m
Infrastructure cost	\$3000/m^2
Cooling and power provisioning equipment cost	\$12.5/W
Cooling and power provisioning equipment space overhead	20%
SPUE (fan and power supply efficiency factor)	1.3
PUE	1.3
Personnel cost	\$200 per rack/month
Networking gear	360W, \$10,000 per rack
Motherboard	25W, \$330 per 1U
Disk	10W, \$180, 100-year MTTF
DRAM	1W, \$25, 800-year MTTF per GB
Processor	30-year MTTF

4 Methodology

We now describe the cost models, server hardware features, workloads, and the simulation infrastructure used in evaluating the various chip organizations at the datacenter level.

4.1 TCO Model

Large-scale datacenters employ high-density server racks to reduce the space footprint and improve cost efficiency. A standard rack can accommodate up to 42 *1U* servers, where each server integrates one or more processors, multiple DRAM DIMMs, disk- or flash-based storage nodes, and a network interface. Servers in a rack share the power distribution infrastructure and network interfaces to the rest of the datacenter. The number of racks in a large-scale datacenter is commonly constrained by the available power budget.

Our TCO analysis, derived using EETCO [7], considers four major expense categories summarized below. Table 2 further details key parameters.

Datacenter infrastructure: Includes land, building, power provisioning and cooling equipment with a 15-year depreciation schedule. Datacenter area is primarily determined by the IT (rack) area, with cooling and power provisioning equipment factored in. The cost of this equipment is estimated per Watt of critical power.

Server and networking hardware: Server hardware includes processors, memory, disks, and motherboards. We also account for the networking gear at the edge, aggregation, and core layers of the datacenter and assume that the cost scales with the number of racks. The amortization schedule is 3 years for server hardware and 4 years for networking equipment.

Power: Predominantly determined by the servers, including fans and power supply, networking gear, and cooling equipment. The electricity cost is \$0.07/KWh.

Maintenance: Includes costs for repairing faulty equipment, determined by its mean-time-to-

failure (MTTF), and the salaries of the personnel.

4.2 Server Processors

We evaluate a number of datacenter server processor designs, as summarized in Table 1. We use publicly available data from the open web and Microprocessor Report to estimate core and chip area, power, and cost. We supplement this information with CACTI 6 for cache area and power profiles, and use measurements of commercial server processors' die micrographs to estimate the area of on-chip memory and I/O interfaces. Many aspects of the methodology are detailed in our ISCA 2012 publication [12].

Baseline parameters: To make the analysis tractable and reduce variability due to differences in the instruction set architecture, we model ARM-based cores for all but the conventional processor designs. Tiled and SOP-inorder configurations are based on the Cortex-A8, a dual-issue in-order core clocked at 1.5GHz. The small-chip design uses a more aggressive dual-issue out-of-order core similar to an ARM Cortex-A9, which is representative of existing products in the small-chip processor design space. For the conventional design, we model a custom 4-wide large-window core running at 3GHz.

At the chip level, we model the associated cache configurations, interconnect topologies, and memory interfaces of the target processors. Our simulations reflect important runtime artifacts of these structures, such as interconnect delays, bank contention, and off-die bandwidth limitations.

Effect of higher-complexity cores: While the low-complexity out-of-order and in-order cores are attractive from an area- and energy-efficiency perspective, their lower performance may be unacceptable for applications that demand fast response times and have non-trivial computational components (e.g., web search [9]). To study the effect of higher-performance cores on datacenter efficiency, we also evaluate a Scale-Out Processor organization based on ARM Cortex-A15, a triple-issue core clocked at 2GHz. Compared to the baseline dual-issue in-order core, a three-way out-of-order core delivers 81% higher single-threaded performance, on average, across our workload suite, while requiring 3.5x the area and over 3x the power per core.

Effect of chip size: The results captured in Figure 1 suggest that small-core designs on a larger chip improve performance-density and energy-efficiency as compared to small-chip organizations. To better understand the effect of chip size on datacenter efficiency, we extend the evaluation space of tiled and SOP processors with additional designs featuring twice the core, cache, and memory interfaces. These "2x" designs approximately match the area of the Xeon-based conventional processor considered in this study.

Processor price estimation: The price for the conventional processor is estimated by picking the lowest price ($800) among online vendors for the target Xeon 5670 processor. Prices for tiled ($300) and small-chip ($95) designs are sourced from Microprocessor Report (Nov. 2011 issue) and correspond to Tilera Tile-Gx3036 and Calxeda ECX-1000, respectively.

To compute the cost for the various SOP and "2x" tiled designs, we use Cadence InCyte Chip Estimation tool. Using the tool, we estimate the production volume of the Tilera Gx-3036 processor to be 200,000 units, given a selling price of $300 and a 50% margin. We use this production volume to estimate the selling price for each processor type, taking into account non-recurring engineering (NRE) costs, mask and production costs, yield, other expense categories, and a 50% profit margin. While the above estimates are used for the majority of the studies, we also consider the sensitivity of different designs to processor price in Section 5.3.

4.3 Workloads and Simulation Infrastructure

Our workloads, which include Data Serving, MapReduce, SAT Solver, Web Frontend, and Web Search, are taken from CloudSuite [1]. For the Web Frontend workload, we use the e-banking option from SPECweb2009 in place of its open-source counterpart from CloudSuite, as SPECweb2009 exhibits better performance scalability at high core counts. Functionally, all these applications have similar characteristics, namely (a) they operate on huge data sets that are split across a large number of nodes into memory-resident shards; (b) the nodes service a large number of completely independent requests that do not share state; and (c) the inter-node connectivity is used only for high-level task management and coordination [4]. Two of the workloads – SAT Solver and MapReduce – are batch, while the rest are latency-sensitive and are tuned to meet the response time objectives.

We estimate the performance of the various processor designs using Flexus full-system simulation [16]. Our metric of performance is the product of UIPC and processor frequency. UIPC, a ratio of committed user instructions over the sum of both user and system cycles, has been shown to be a more accurate performance metric in full-system evaluation than total IPC due to the contribution of I/O and spin locks in the OS to the execution time [16].

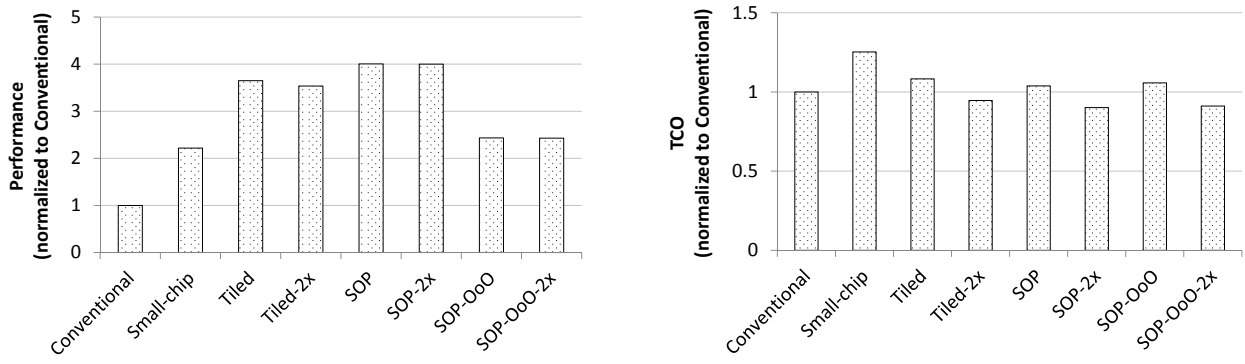
Due to space constraints, we only present aggregate results across all workloads. Performance data is averaged using a harmonic mean.

4.4 Experimental Setup

For all experiments, we assume a fixed datacenter power budget of 20MW and a power limit of 17kW per rack. We evaluated lower-density racks rated at 6.6kW, but found the trends to be identical across the two rack configurations and present one set of results for space considerations.

To compare the performance and TCO of different server architectures, we start with a rack power budget and subtract all power costs at both the rack and board level, excluding the processors. The per-rack costs include network gear, cooling (fans), and power conversion. At the 1U server level, we account for motherboard, disks (2), and memory (model parameter) power. The remaining power budget is divided by the peak power of each evaluated processor chip to determine the number of processors per server. Datacenter performance is then estimated based on the number of processors in each 1U server (using the per-processor performance data collected in simulation), the number of servers in a rack, and the number of racks in the datacenter.

Finally, we make no assumptions on what the optimal amount of memory per server is, which in practice varies for different workloads, and model servers with 32, 64, and 128 GB of memory per 1U. One simplifying assumption we do make is that the amount of memory per 1U is independent of the chip design. Underlying this assumption are the observations that (a) the data is predominantly read-only, and (b) the data is partitioned for high parallelism, allowing performance to scale with more cores and sockets until bottlenecked by the bandwidth of the memory interfaces. Bandwidth limitations are accounted for in our studies.



(a) Datacenter performance (higher is better)

(b) Datacenter TCO (lower is better)

Figure 3: Datacenter performance and TCO for various server processors normalized to a design based on a conventional processor.

5 Evaluation

5.1 Performance and TCO

We first compare datacenter performance and TCO for various processor designs assuming 64GB of memory per 1U server. The results are presented in Figure 3.

In general, we observe significant disparity in datacenter performance across the processor range stemming from the different capabilities and energy profiles of the various processor architectures. Highly integrated processors based on small cores, namely tiled and SOP, deliver the highest performance at the datacenter level. The tiled small-core big-chip architecture improves aggregate performance by a factor of 3.6 over conventional and 1.6 over small-chip designs. The tiled design is superior due to a combination of efficient core microarchitectures and high chip-level integration – attributes which help amortize the power of both on-chip and server-level resources among many cores, affording more power for compute resources.

The highest performance is delivered by the SOP with in-order cores, a small-core big-chip design with the highest performance-density that improves datacenter performance by an additional 10% over the tiled processor. The SOP design effectively translates its performance-density advantage into a performance advantage by better amortizing fixed power overheads at the chip level among its many cores, ultimately affording more power for the execution resources at the rack level.

The Scale-Out Processor design based on out-of-order cores sacrifices 39% of the throughput at the datacenter level as compared to the in-order design. However, higher core complexity might be justified for workloads that demand tight latency guarantees and have a non-trivial computational component. Even with higher-complexity cores, the SOP architecture attains better datacenter performance than either the conventional or small-chip alternatives.

The differences in TCO among the different designs are not as pronounced as differences in performance, owing to the fact that processors contribute only a fraction (19-39%) to the overall datacenter acquisition and power budget. Nonetheless, one important trend worth highlighting is that while small-chip designs are significantly less expensive and more energy-efficient (by a factor of 8 and 2.2, respectively) than conventional processors on a per-unit basis, at the datacenter level, a small-chip design has a 25% higher TCO. The reason for the apparent paradox is that the limited

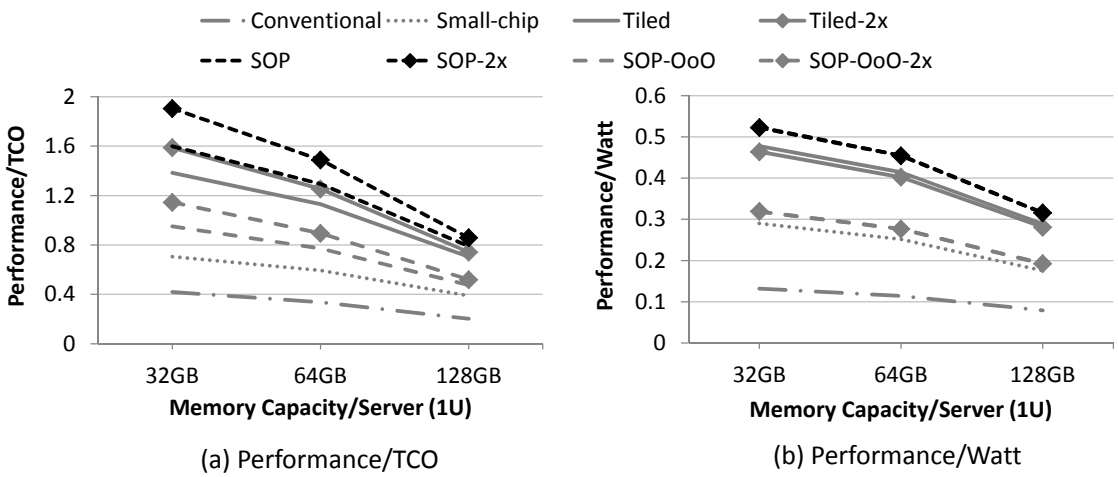


Figure 4: Datacenter efficiency for different server chip designs. Data not normalized.

computational capabilities of the small-chip design necessitate as many as 32 sockets (versus 2 for conventional) per 1U server in order to saturate the available power budget. The acquisition costs of such a large number of chips negate the differences in unit price and energy-efficiency, emphasizing the need to consider total cost of ownership in assessing datacenter efficiency.

5.2 Relative Efficiency

We next examine the combined effects of performance, energy-efficiency, and TCO by assessing the various designs on datacenter performance/TCO and performance/W. Figure 4 presents the results as memory capacity is varied from 32 to 128 GB per 1U server.

With 64GB of memory per 1U server, the following trends can be observed:

- The small-chip design improves performance/W by 2.2x over a conventional processor, but its performance/TCO advantage is just 1.8x due to the high processor acquisition costs, as noted earlier.
- A datacenter based on the tiled design improves performance per TCO by a factor of 3.4 over conventional and 1.9 over small-chip. Energy-efficiency is improved by 3.6x and 1.6x, respectively, underscoring the combined benefits of aggressive integration and the use of an efficient core microarchitecture.
- SOP designs with an in-order core further improve performance/TCO by 14% and performance/Watt by 10% over tiled processors through a more efficient use of the chip real-estate.
- Tiled and SOP designs with twice the resources (Tiled-2x and SOP-2x) improve TCO by 11-15% over their baseline counterparts by reducing the number of processor chips, thereby lowering acquisition costs.
- The SOP design with out-of-order cores achieves 40% lower performance/TCO than the design based on in-order cores. The more aggressive out-of-order microarchitecture is responsible for the lower energy- and area-efficiency of each core, resulting in lower throughput at the chip level. When the TCO premium is justified, which may be the case for profit-generating latency-sensitive applications (e.g., web search), the out-of-order SOP design offers a 2.3x per-

formance/TCO advantage (2.4x in performance/Watt) over a conventional server processor.

While the discussion above focuses on servers with 64GB of memory, the trends are similar with other memory configurations. In general, servers with more memory lower the performance-to-TCO ratio, as memory adds to server cost while diminishing the processor power budget. The opposite is true for servers with less memory, in which the choice of processor has a greater effect on both cost and performance.

The key result of our study that highly integrated server processors are beneficial from a TCO perspective are philosophically similar to the observations made by Karidis et al., who noted that high-capacity servers are effective in lowering the cost of computation [10].

5.3 Sensitivity to Processor Price

Figure 5 shows the effect of varying the processor price on the relative efficiency (performance/TCO) of the different designs assuming 64GB of memory per 1U server. For each processor type, we assume an ASIC design in 40nm technology and compute the price as a function of market size, ranging from 40K to 1M units, as described in Section 4.

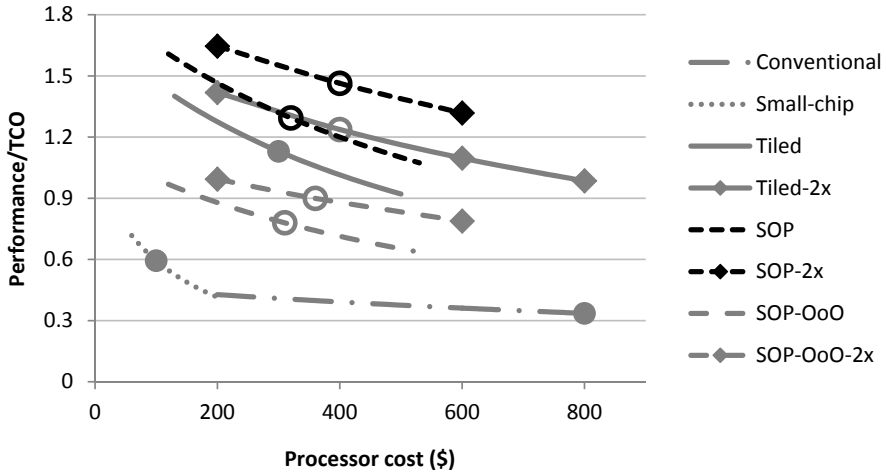


Figure 5: Relationship between the price per processor and TCO. Solid circles indicate known market prices; unfilled circles show estimated prices based on a production volume of 200K units.

In general, we observe that the price of larger chips has less impact on datacenter TCO as compared to that of smaller chips, since it takes fewer large chips to populate a server due to power constraints. In contrast, the small-chip design is highly sensitive to unit price due to the sheer volume of chips required per 1U server. For instance, there is a factor of 16 difference in the number of chips per server between conventional and small-chip designs.

A consistent trend in our study is that, from a TCO perspective, larger chips are preferred to smaller ones, as seen in the curves for the various tiled and SOP designs. While the additional chip area adds expense, the price difference is modest (around 16% or $50 per chip), as NRE and design costs dominate production costs. Furthermore, the increased cost is offset by the reduction in the number of required chips.

6 Conclusion

Our society is becoming increasingly dependent on real-time access to vast quantities of data for business, entertainment, and social connectivity. Scale-out datacenters with multi-megawatt power budgets are the enabling technology behind this transition to a digital universe. As the semiconductor industry reaches the limits of voltage scaling, effective datacenter design calls for unprecedented emphasis on energy-efficiency, performance, and TCO.

One way to improve both performance and TCO is through specialization by taking advantage of the workload characteristics. Scale-Out Processors are highly integrated chips specifically designed for the demands of scale-out workloads that maximize the throughput per chip via a combination of rich compute resources and a tightly coupled memory subsystem optimized for area and delay. Scale-Out Processors extend the advantages of emerging *small-core big-chip* architectures that (a) provide good energy-efficiency through the use of simple core microarchitectures, and (b) maximize the TCO investment by fully utilizing server hardware via abundant execution resources. In the near term, Scale-Out Processors can deliver performance and TCO gains in a non-disruptive manner by fully leveraging existing software stacks. Further out, demands for greater performance and energy-efficiency may necessitate even greater degrees of specialization, requiring a disruptive change to the programming model.

Acknowledgement

We thank Sachin Idgunji and Emre Ozer, both affiliated with ARM at the time of the writing, for their contributions to the paper. We thank the EuroCloud project partners for inspiring the Scale-Out Processors. This work was partially supported by EuroCloud, Project No 247779 of the European Commission 7th RTD Framework Programme - Information and Communication Technologies: Computing Systems.

References

[1] CloudSuite 1.0. `http://parsa.epfl.ch/cloudsuite`, 2012.

[2] M. Berezecki, E. Frachtenberg, M. Paleczny, and K. Steele. Many-Core Key-Value Store. In *International Green Computing Conference*, July 2011.

[3] Calxeda. Calxeda EnergyCore ECX-1000 Series. `http://www.calxeda.com/wp-content/uploads/2012/06/ECX1000-Product-Brief-612.pdf`, 2012.

[4] M. Ferdman, A. Adileh, O. Kocberber, S. Volos, M. Alisafaee, D. Jevdjic, C. Kaynak, A. D. Popescu, A. Ailamaki, and B. Falsafi. Clearing the Clouds: A Study of Emerging Scale-Out Workloads on Modern Hardware. In *International Conference on Architectural Support for Programming Languages and Operating Systems*, March 2012.

[5] J. Hamilton. Overall data center costs. `http://perspectives.mvdirona.com/2010/09/18/OverallDataCenterCosts.aspx`.

[6] N. Hardavellas, M. Ferdman, B. Falsafi, and A. Ailamaki. Toward Dark Silicon in Servers. *IEEE Micro*, 31:6 –15, July/August 2011.

[7] D. Hardy, I. Sideris, A. Saidi, and Y. Sazeides. EETCO: a Tool to Estimate and Explore the Implications of Datacenter Design Choices on the TCO and the Environmental Impact. In *1st Workshop on Energy-efficient Computing for a Sustainable World*, Dec 2011.

[8] U. Hoelzle and L. A. Barroso. *The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines*. Morgan and Claypool Publishers, 1st edition, 2009.

[9] V. Janapa Reddi, B. C. Lee, T. Chilimbi, and K. Vaid. Web Search Using Mobile Cores: Quantifying and Mitigating the Price of Efficiency. In *International Symposium on Computer Architecture*, pages 314–325, 2010.

[10] J. Karidis, J. E. Moreira, and J. Moreno. True Value: Assessing and Optimizing the Cost of Computing at the Data Center Level. In *Conference on Computing Frontiers*, pages 185–192, 2009.

[11] K. Lim, P. Ranganathan, J. Chang, C. Patel, T. Mudge, and S. Reinhardt. Understanding and Designing New Server Architectures for Emerging Warehouse-Computing Environments. In *International Symposium on Computer Architecture*, pages 315–326, 2008.

[12] P. Lotfi-Kamran, B. Grot, M. Ferdman, S. Volos, O. Kocberber, J. Picorel, A. Adileh, D. Jevdjic, S. Idgunji, E. Ozer, and B. Falsafi. Scale-Out Processors. In *International Symposium on Computer Architecture*, 2012.

[13] Morgan, Timothy P. Facebook's open hardware: Does it compute? `http://www.theregister.co.uk/2011/04/08/open_compute_server_comment`, 2011.

[14] SAVVIS. SAVVIS Sells Assets Related to Two Data Centers for $200 Million. `http://www.savvis.com/en-US/Company/News/Press/Pages/SAVVIS%20Sells%20Assets%20Related%20to%20Two%20Data%20Centers%20for%20200%20Million.aspx`, 2007.

[15] Tilera. TILE-Gx 3036 Specifications. `http://www.tilera.com/sites/default/files/productbriefs/Tile-Gx%203036%20SB012-01.pdf`, 2011.

[16] T. Wenisch, R. Wunderlich, M. Ferdman, A. Ailamaki, B. Falsafi, and J. Hoe. SimFlex: Statistical Sampling of Computer System Simulation. *IEEE Micro*, 26:18 –31, July-Aug 2006.

Boris Grot is a postdoctoral researcher at EPFL. His research focuses on processor architectures, memory systems, and interconnection networks for high-throughput, energy-aware computing. Grot has a PhD in computer science from the University of Texas at Austin. He is a member of the ACM and IEEE.

Damien Hardy received his PhD degree in computer science from the University of Rennes, France, in 2010. He is currently working at the University of Cyprus as a postdoctoral researcher. His research interests include computer architecture, reliability, embedded and real-time systems, and data center modeling.

Pejman Lotfi-Kamran is a fifth-year PhD candidate at EPFL. His research interests include processor architecture and interconnection networks for high-throughput and energy-efficient datacenters. He is a student member of the ACM and IEEE.

Chrysostomos Nicopoulos is a Lecturer in the Department of Electrical and Computer Engineering at the University of Cyprus. His research interests are in the areas of Networks-on-Chip, computer architecture, multi-/many-core microprocessor and system design, and architectural support for massively parallel computing. Nicopoulos has a PhD in Electrical Engineering (specialization in Computer Engineering) from the Pennsylvania State University, USA. He is a member of the ACM and IEEE.

Yiannakis Sazeides is an Associate Professor at the University of Cyprus. He was awarded a PhD from the University of Wisconsin-Madison in 1999. He worked at Compaq and Intel toward the development and design of high performance processors. He is a member of the IEEE and the HiPEAC Network of Excellence where he contributes in efforts promoting reliability awareness and research. His interests lie in the area of Computer Architecture with particular emphasis on reliability, data center modeling, memory hierarchy, temperature, and analysis of dynamic program behavior.

Babak Falsafi is a professor of computer and communication sciences at EPFL, and the founding director of EcoCloud 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 a fellow of IEEE.

Direct questions and comments about this article to Boris Grot, EPFL IC ISIM PARSA, INJ 238 (Batiment INJ), Station 14, CH-1015, Lausanne, Switzerland; boris.grot@epfl.ch.