Silicon Efficiency in Post-Moore Servers

Ali Ansari[‡], Shanqing Lin[‡], Ayan Chakraborty[‡], Bugra Eryilmaz[‡],
Mohammad Alian[◇], Babak Falsafi[‡], Michael Ferdman[§]

[‡]EPFL [◇]University of Kansas [§]Stony Brook University

***Abstract*—Server CPUs in the cloud have inherited their core microarchitecture from the desktop and mobile world, with performance primarily measured by single-core IPC. Furthermore, cores are integrated with large cache hierarchies within sockets and rely heavily on these caches to contain chip power envelopes, with little consideration given to utilization by workloads. Wasted silicon impacts both operational and embodied emissions in server platforms. In this work, we measure and compare silicon efficiency measured in performance per area and performance per watt of online and analytic services running on two x86 and an ARM server. We show that while x86 platforms offer higher single-core performance, the ARM server has the potential to achieve up to 2.5× higher socket-level performance per area and performance per watt than the x86 servers in the absence of system-level bottlenecks (e.g., memory or network bandwidth).**

***Index Terms*—datacenters, efficiency, post-moore**

I. Introduction

Datacenters are massive computing infrastructures built with cost-effective volume servers to provide global IT services. Datacenters have enjoyed exponential growth in recent decades, emerging as a pillar of modern society where nearly all daily activities are digitized, with projections for growth continuing well into the coming decades. Unfortunately, this growth coincides with the demise of Moore's Law and Dennard Scaling, resulting in an unprecedented increase in global datacenter electrical consumption, leading to higher emissions from both embodied carbon and sources of electricity [1].

To maximize return on investment, datacenters employ volume servers based on the basic computer organization and operating systems developed for desktops in the early 90s [12]. Accordingly, server CPUs have been traditionally designed to optimize single-core performance, measured in IPC (Instructions Per Cycle). While high single-core performance helps reduce latency, especially in online services with tight Service-Level Objectives (SLOs), it also results in disproportionately low levels of silicon utilization. This underutilization is due to the infrequent use of SIMD/vector units [20], overprovisioned cache capacities, idle CPU cycles caused by long latency memory stalls, and the inherently low instruction-level and memory-level parallelism of server workloads [16].

Recently, a few vendors have opted for sockets with leaner cores that allow for a higher core count in the same area and power envelope as sockets with wide high-performance cores to exploit the thread-level parallelism of server workloads [3], [6]. Lean-core designs forego the conventional wisdom of optimizing single-core performance in servers in favor of having a larger number of cores to improve the socket-level performance. Multiple new examples of widely deployed ARM-based server CPUs are available, ranging from Cavium's ThunderX to Amazon's Graviton (now in its fourth generation), HiSilicon's Kunpeng, and Ampere's Altra. Ampere has even recently announced a chip with 192 cores [10].

Considering these two different philosophies of designing server CPUs in the Post-Moore era, either with a smaller number of wide high-performance cores, or a larger number of lean less-powerful cores, a crucial question emerges: which of these two approaches is best for optimizing the capital expenditure (CAPEX) and operational expenditure (OPEX) associated with building and running these CPUs?

CAPEX encompasses both the material costs involved in building a chip as well as the emissions from facility-infrastructure construction and chip manufacturing. As chips become larger and more complex to continue delivering increasing performance in the absence of Moore's law, there has been a rapid increase in CAPEX, fueled by the increase in emissions not only from building larger chips but also from building all the additional hardware needed in modern chips to enable their complex functionality [15], [18].

OPEX comprises the electricity consumed to power on and cool down these CPUs in datacenters. Since the 1990s, the Thermal Design Power (TDP) of CPUs has risen from a single-digit value to around 100 W in 2000 and has then stabilized for about ten years thanks to Dennard Scaling. However, TDPs are increasing rapidly with the latest CPUs due to the end of Dennard Scaling and Moore's Law [7].

To estimate the impact of the two different design philosophies on CAPEX and OPEX, we examine the *silicon efficiency* of a chip, quantified by two metrics: performance per area and performance per watt. A chip that extracts higher performance out of a given area budget makes better use of the associated CAPEX. Likewise, a chip with higher performance per watt reduces the OPEX associated with the chip's life cycle.

In this paper, we employ a suite of monolith server workloads and microservices to evaluate the single-core and socket-level silicon efficiency of three commodity server CPUs: two x86 servers, Ice Lake-SP from Intel and Zen 3 from AMD, and an ARM server, Altra from Ampere. We ensure reasonable operating conditions for all systems we study, including maintaining target SLOs for the online services. Our results lead to the following conclusions:

- Zen 3 and Altra achieve about 1.35× and 2.26× higher average single-core performance per area than Ice Lake-SP, despite achieving only 0.87× and 0.51× the single-core performance of Ice Lake-SP, on average. These results underscore the appealing opportunity to optimize

server CPU core microarchitectures for silicon efficiency rather than absolute single-core performance.
- Assuming ideal performance scalability with the given number of cores, Altra outperforms Ice Lake-SP in terms of socket-level performance per area and per watt by 1.43× and 1.48×, on average, respectively. Altra's silicon efficiency advantage suggests an opportunity for improving the CAPEX and OPEX of datacenters by utilizing the design philosophy of building CPUs with a large number of less powerful cores.
- In practice, modern server CPUs cannot fully leverage their intrinsic silicon efficiency potential due to scalability limitations arising primarily from system bottlenecks like limited memory and network bandwidth, inter-chiplet communication delays, along with synchronization and load imbalance issues in a few software stacks. Therefore, there is a need for hardware-software co-design to enhance the performance scalability of server workloads on modern platforms and to implement effective workload consolidation strategies to maximize these sockets' utilization.

II. Methodology

This section provides an overview of the workloads and platforms under examination, the key evaluation metrics, and the methodology employed for tuning workload-specific parameters and conducting measurements.

Workloads and Platforms. We study monolith server workloads from CloudSuite 4.0 [4] and Media Service from DeathStarBench [17], referred to as DSB Media Service in this paper, representing microservices. In Table II, we classify workloads into three groups. The first group, from Data Caching to DSB Media Service, consists of network-intensive online services with relatively high kernel-space activity, executing more than 10% of the instructions in the kernel space. The second group, Web Search and Web Serving, represents online services executing instructions mostly in the user space ($>$ 98%). The last group consists of analytics workloads. Table I summarizes the configuration of the platforms utilized in our study. All platforms run Ubuntu 22.04 with Linux kernel version `5.15`.

Metrics. We define *performance* as the server's handled client requests per second for online services and the reciprocal of the total execution time for analytics. Moreover, we quantify *silicon efficiency* by performance per area and performance per watt. We collect the core, chiplet, and socket area of the CPUs studied in this work from various public resources [2], [3], [5], [8], [9], [19]. To ensure a fair comparison, we fix the clock frequency at 2.45 GHz, the maximum sustainable frequency across our platforms. Fixing the frequency among the platforms isolates the impact of frequency scaling from our performance studies. Because 2.45 GHz is different than Ice Lake-SP and Altra's nominal frequency, we use `turbostat` and `impitool` on these platforms and measure a maximum power consumption of 185 W and 160 W when these sockets are fully utilized with our workloads. For Zen 3, our power measurement value matches the reported TDP value at 2.45 GHz. We use these power values for calculating the socket-level performance per watt results shown in Table II.

Measurements. We tune the workloads' parameters to reach the maximum throughput at a target SLO for online services and minimize the execution time for analytics. Our target SLO is 1 ms, 5 ms, 150 ms, 250 ms, and 100 ms 99^{th} percentile tail latency for Data Caching, Data Serving, Web Search, Web Serving, and DSB Media Service, respectively.

We choose eight cores (the number of cores per chiplet on Zen 3) as the smallest granularity for running DSB Media Service, because running a service including over 30 separate containers (i.e., microservices) on a single core is not representative. Therefore, single-core experiments in this paper will consider the results of eight cores for DSB Media Service.

We run workloads with and without SMT on Ice Lake-SP and Zen 3 platforms and report the results for whichever setup achieves a higher single-core or socket-level performance.

For workloads whose scalability issues originate from the software stack, we co-locate multiple instances of the workload, with each instance assigned to a disjoint set of cores. This approach allows us to improve the socket-level utilization for these workloads across all of our platforms.

Special Tuning for Data Caching. Data Caching, a high throughput workload, triggers frequent interrupts for sending and receiving network packets. As a result, the cores handling software interrupt handlers experience high utilization and must be isolated from those running the application logic. Following insights from prior work [13], we empirically find that devoting half of the cores to these handlers and the rest to the application logic offers a good balance for socket-level experiments. For Zen 3, we also apply this policy for the cores within a chiplet.

III. Experimental Results

In this section, we begin by comparing the silicon efficiency of all platforms under consideration at both the core and socket levels. Following this comparison, we briefly go over the scalability limitations of these workloads and platforms.

A. Silicon Efficiency

Table II showcases the single-core performance of the platforms for all workloads normalized to Ice Lake-SP. We find that Zen 3 offers competitive single-core performance to Ice Lake-SP for the second and third groups of workloads. For the first group of workloads, Zen 3 exhibits over 15% lower single-core performance. We attribute this performance drop to the smaller L2 capacity of Zen 3, which results in a higher L2 MPKI for the kernel instructions (for example, 9 on Zen 3 compared to 0.3 on Ice Lake-SP for Data Caching). Altra consistently demonstrates lower single-core performance than both x86 platforms for all workloads. These findings suggest that the microarchitectural advantages of x86 platforms effectively translate into higher single-core performance compared to the simpler ARM core utilized in Altra.

We also report the single-core performance per area normalized to Ice Lake-SP in Table II. Considering the cores'

TABLE I
PLATFORMS' CONFIGURATIONS.

		Core					Socket				Platform	
Machine	Name	Max IPC	L1 I/D (KB)	L2 (MB)	SMT	Area (mm2)	Phys. Cores	LLC (MB)	TDP @ nominal Freq.	Area (mm2)	DRAM (# DIMMs, size)	Network BW
Intel Xeon Gold 6338N	Ice Lake-SP	5	32/48	1.25	Yes	6.2	32	48	185 W @ 2.2 GHz	640	8, 32 GB	100 Gbps
AMD EPYC 7763	Zen 3	5	32/32	0.5	Yes	4	64	256	225 W @ 2.45 GHz	640+416 (cores+IO)	8, 32 GB	100 Gbps
Ampere Altra Q80-30	Altra	4	64/64	1	No	1.4	80	32	210 W @ 3 GHz	574	16, 16 GB	100 Gbps

TABLE II
PERFORMANCE AND SILICON EFFICIENCY OF THE PLATFORMS NORMALIZED TO ICE LAKE-SP.

	Performance		Single Core Perf/Area		Ideal Socket Perf/Area		Ideal Socket Perf/Watt		Actual Socket Perf/Area			Actual Socket Perf/Watt		
	Zen 3	Altra	Zen 3	Altra	Zen 3	Altra	Zen 3	Altra	Ice Lake SP	Zen 3	Altra	Ice Lake SP	Zen 3	Altra
Data Caching	0.74	0.42	1.15	1.85	0.90	1.16	1.22	1.21	0.90	0.47	0.40	0.90	0.63	0.42
Data Serving	0.84	0.39	1.30	1.72	1.01	1.08	1.38	1.12	0.91	0.91	0.62	0.91	1.24	0.64
Media Streaming	0.73	0.33	1.13	1.47	0.88	0.92	1.20	0.96	0.28	0.18	0.32	0.28	0.24	0.33
DSB Media Service	0.68	0.24	1.05	1.05	0.82	0.66	1.11	0.68	0.79	0.70	0.53	0.79	0.95	0.55
Web Search	0.95	0.60	1.47	2.67	1.15	1.68	1.56	1.75	0.91	0.98	1.63	0.91	1.33	1.69
Web Serving	0.92	0.73	1.43	3.25	1.12	2.05	1.51	2.12	0.89	1.13	1.40	0.89	1.53	1.46
Data Analytics	0.95	0.69	1.47	3.06	1.15	1.93	1.56	2.00	0.29	0.20	0.34	0.29	0.27	0.35
Graph Analytics	1.13	0.70	1.75	3.12	1.37	1.96	1.86	2.04	0.62	0.32	0.79	0.62	0.43	0.82
In-memory Analytics	0.98	0.87	1.52	3.86	1.19	2.43	1.61	2.52	0.48	0.32	0.57	0.48	0.43	0.59
Geometric Mean	0.87	0.51	1.35	2.26	1.05	1.43	1.43	1.48	0.62	0.47	0.62	0.62	0.64	0.65

area, as shown in Table I, single-core performance per area is calculated by the normalized performance multiplied by $6.2/4 = 1.55\times$ for Zen 3 and $6.2/1.4 = 4.43\times$ for Altra. These scaling factors reveal that Ice Lake-SP's higher single-core performance comes at the expense of a disproportionately larger core area, enabling Zen 3 and Altra to achieve an average of $1.35\times$ and $2.26\times$ higher single-core performance per area, respectively.

Table II also presents the platforms' ideal socket-level performance per area. By ideal scalability, we assume that the socket-level performance is the single-core performance multiplied by the number of cores available in a socket. Therefore, the ideal socket-level performance per area for Zen 3 and Altra can be simply calculated by multiplying their normalized performance by $(64/1056)/(32/640) = 1.21\times$ and $(80/574)/(32/640) = 2.79\times$. These scaling factors clearly highlight Altra's upper hand not only in single-core but also socket-level performance per area compared to the other platforms. Because these scaling factors are smaller than those of single-core performance per area, the socket-level performance per area opportunity is expected to be smaller for Zen 3 and Altra platforms. The reason is sockets have several components, such as the last-level cache and I/O interfaces, which add to the silicon requirements of a socket. The results indicate that socket-level performance per area of Zen 3 and Altra are $1.05\times$ and $1.43\times$ higher than Ice Lake-SP, on average. On the one hand, despite having twice as many cores in Zen 3 than Ice Lake-SP, Zen 3's socket-level performance per area is on par with Ice Lake-SP. The reason is twofold. First, Zen 3's immense last-level cache occupies a significant silicon area without providing a proportional single-core performance benefit. Second, Zen 3 implements the I/O die in a separate chiplet fabricated with a larger technology node. On the other hand, we note that Altra's silicon efficiency advantage stems mostly from workloads in the second and third groups, for which Altra offers $2\times$ higher performance per area than Ice Lake-SP, on average.

Following the same methodology, we measure the scaling factors for ideal socket-level performance per watt to be $(64/225)/(32/185) = 1.64\times$ and $(80/160)/(32/185) = 2.89\times$ for Zen 3 and Altra, respectively. Zen 3 has a higher scaling factor for socket-level performance per watt than per area. The reason is that dark silicon in its large last-level cache, while increasing the socket area, does not impose a proportional power consumption overhead on the design. Unlike Zen 3, we notice a similar scaling factor for both ideal socket-level performance per area and watt for Altra, indicating that Ice Lake-SP and Altra have similar provisioning of power per unit area (i.e., power density). The results suggest that Zen 3 offers a superior socket-level performance per watt than Ice Lake-SP by up to $1.86\times$, and $1.43\times$, on average. Altra also outperforms Ice Lake-SP on all workloads, except for DSB Media Service, providing up to $2.5\times$ higher performance per watt, and $1.48\times$ on average.

The last two columns in Table II present the actual socket-level performance per area and performance per watt we could achieve. The results are normalized to the ideal silicon efficiency of Ice Lake-SP. Comparing the ideal and actual silicon efficiency results suggests that, except for Web Search, other workloads reach a drastically lower silicon efficiency than their potential. The reason is the poor performance scalability of server workloads with the number of cores they utilize. Web Search has been known for its proper scalability because of the read-only nature of the workload and serving independent

requests with no inter-thread communications [11]. This workload showcases the superior silicon efficiency of Zen 3 and Altra platforms in practice, providing 1.33× and 1.69× higher performance per watt than Ice Lake-SP, respectively. However, the rest of the workloads experience various scalability limitations because of issues in both software and hardware, as will be briefly discussed in the next section. Accordingly, Zen 3 and Altra leave a significant opportunity behind, losing their silicon efficiency edge to Ice Lake-SP in several categories of benchmarks. Besides Web Search, only Web Serving exhibits a higher silicon efficiency on Altra than Ice Lake-SP, despite losing about 30% of its ideal silicon efficiency.

Our study in this section highlights the opportunities and challenges of modern server CPUs. First, the single-core performance metric seems not to be a correct optimization factor for server CPUs running workloads whose performance metric considers the SLO. We observe that the higher core count achieved with smaller cores compensates for the lower single-core performance and provides considerably higher performance per area and per watt potential. However, in practice, the workloads and platforms fall short of realizing this available potential, suggesting an interesting research area to bridge the gap.

B. Scalability

In this section, we briefly discuss the scalability limitations of the workloads on the platforms. We note that workloads may face separate scalability limitations on various platforms because of the workload's requirements and the underlying hardware organization.

Scalability Bottlenecks from Software. On the software front, we recognize three scalability bottlenecks. First, server workloads may need synchronization among the threads to manage access to shared data structures and objects. These synchronizations waste precious CPU cycles and place a scalability upper bound [14]. Data Caching, Data Serving, DSB Media Service, and analytics face this issue by executing an increased number of instructions per client query for the online services and the whole program execution for analytics when the number of cores given to the workload increases.

Second, we recognize a load imbalance among the threads handling the software interrupt handlers in the ARM Linux running on Altra. This issue prohibits Data Caching from scaling on Altra to the same extent it scales on Ice Lake-SP. A few cores become highly loaded while the rest of the cores are underutilized. Because of the SLO, even a single highly saturated core affects the tail latency and impedes Data Caching's scalability on Altra.

Third, analytics server workloads have serial phases during execution, like the reduce and aggregate phases in map-reduce application Data Analytics. These serial phases cannot benefit from the available cores, resulting in a disproportionate performance improvement with additional core count.

Scalability Bottlenecks from Hardware. The first hardware scalability issue is provisioning the right amount of hardware resources according to the workloads' requirements. Specifically, we focus on the network and memory requirements of the workloads.

Among our workloads, Data Caching and Media Streaming put significant pressure on the network bandwidth. Data Caching's network bandwidth utilization depends on the dataset's object size. In our case, the 100Gbps NICs available on our platforms could satisfy the network bandwidth requirements of Data Caching. Moreover, we could not achieve an ideal scalability curve for Data Caching, which translates into a lower network bandwidth requirement than the ideal case. For Media Streaming, all platforms saturate the network bandwidth when utilizing less than half of their available cores. Therefore, the scalability limitation stemming from a limited network bandwidth keeps more than half of the available cores idle.

The three analytics workloads put significant pressure on the memory bandwidth, hitting the maximum memory bandwidth available in our sockets during their execution. Therefore, correct memory bandwidth provisioning with additional DIMMs or technologies with higher bandwidth, like DDR5 memory modules, will help mitigate the fraction of time analytics spend waiting to fetch data from the main memory.

Besides hardware resource contention, inter-chiplet communication presents a scalability challenge for server workloads on the Zen 3 platform, particularly affecting those requiring frequent inter-thread communication and having stringent tail latency requirements. In our analysis, workloads such as Data Caching and DSB Media Service see diminished performance gains on Zen 3 as core counts increase from eight to 16, a point at which workloads extend beyond a single chiplet.

IV. Conclusion

This paper examines three distinct real server platforms from different vendors, aiming to evaluate their silicon efficiency, specifically their performance per area and per watt. Our findings challenge the conventional notion of improving single-core performance for server CPU design. We show that sockets containing a large number of lean cores have the potential to achieve higher silicon efficiency, which effectively translates to extracting higher performance out of a given CAPEX and OPEX budget in building and running a socket.

Despite Altra's potential to achieve over 2.5× higher performance per area and per watt compared to Ice Lake-SP, assuming ideal performance scaling with the core count, our evaluation uncovers challenges hindering the full exploitation of this opportunity. The reasons include insufficient memory and network bandwidth provisioning, software stack synchronization, inter-chiplet communication, and the imbalanced load of the interrupt handling cores. Each of these challenges opens up interesting opportunities for future work to improve the scalability of software stacks on modern hardware and also enhance the utilization of these sockets via effective workload collocation to realize the true silicon efficiency potential of sockets with hundreds of cores.

REFERENCES

[1] "Achieving Sustainable Data Center Growth," https://t.ly/L8W7V.

[2] "AMD EPYC 7763 Specs," https://www.techpowerup.com/cpu-specs/epyc-7763.c2373.

[3] "Ampere's Altra Max 80 Core Arm CPU Gets Benchmarked, Delidded, Measured," https://www.tomshardware.com/news/ampere-altra-max-80-ccore-arm-delidded.

[4] "CloudSuite 4.0," https://www.cloudsuite.ch/.

[5] "Golden Cove vs Zen 3 core size comparisons," https://www.reddit.com/r/intel/comments/vgc9hc/golden_cove_vs_zen_3_core_size_comparisons_by/.

[6] "Hot Chips 2023: Architecting for Flexibility and Value with Next Gen Intel® Xeon® Processors," https://t.ly/4fbuG.

[7] "How efficient are data centers really?" https://www.bulletin.ch/de/news-detail/wie-effizient-sind-rechenzentren-wirklich.html.

[8] "Ice Lake (client) - Microarchitectures - Intel," https://en.wikichip.org/wiki/intel/microarchitectures/ice_lake_(client).

[9] "Neoverse N1 - Microarchitectures - ARM," https://en.wikichip.org/wiki/arm_holdings/microarchitectures/neoverse_n1.

[10] "Sound The Siryn: AmpereOne 192-Core CPU," https://t.ly/2H0Ji.

[11] G. Ayers, J. H. Ahn, C. Kozyrakis, and P. Ranganathan, "Memory hierarchy for web search," in *2018 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 2018, pp. 643–656.

[12] L. Barroso, J. Dean, and U. Holzle, "Web search for a planet: The google cluster architecture," *IEEE Micro*, vol. 23, no. 2, pp. 22–28, 2003.

[13] S. Chen, S. GalOn, C. Delimitrou, S. Manne, and J. F. Martinez, "Workload characterization of interactive cloud services on big and small server platforms," in *2017 IEEE International Symposium on Workload Characterization (IISWC)*. IEEE, 2017, pp. 125–134.

[14] S. Eyerman and L. Eeckhout, "Modeling critical sections in amdahl's law and its implications for multicore design," in *Proceedings of the 37th annual international symposium on Computer architecture*, 2010, pp. 362–370.

[15] Y. Feng and K. Ma, "Chiplet actuary: A quantitative cost model and multi-chiplet architecture exploration," in *Proceedings of the 59th ACM/IEEE Design Automation Conference*, 2022, pp. 121–126.

[16] 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," *Acm sigplan notices*, vol. 47, no. 4, pp. 37–48, 2012.

[17] Y. Gan, Y. Zhang, D. Cheng, A. Shetty, P. Rathi, N. Katarki, A. Bruno, J. Hu, B. Ritchken, B. Jackson *et al.*, "An open-source benchmark suite for microservices and their hardware-software implications for cloud & edge systems," in *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems*, 2019, pp. 3–18.

[18] U. Gupta, Y. G. Kim, S. Lee, J. Tse, H.-H. S. Lee, G.-Y. Wei, D. Brooks, and C.-J. Wu, "Chasing carbon: The elusive environmental footprint of computing," *IEEE Micro*, vol. 42, no. 4, p. 37–47, jul 2022. [Online]. Available: https://doi.org/10.1109/MM.2022.3163226

[19] D. Patel, "Intel icelake server die size floorplan inefficiencies revealed," https://web.archive.org/web/20220823124607/https://semianalysis.com/intel-icelake-server-die-size-floorplan-inefficiencies-revealed/, 2020, archived: 2022-09-28.

[20] A. Sriraman, A. Dhanotia, and T. F. Wenisch, "Softsku: Optimizing server architectures for microservice diversity@ scale," in *Proceedings of the 46th International Symposium on Computer Architecture*, 2019, pp. 513–526.