THEME ARTICLE: THE PAST, PRESENT, AND FUTURE OF WAREHOUSE-SCALE COMPUTING

Server Architecture From Enterprise to Post-Moore

Babak Falsafi, *Ecole Polytechnique Federale de Lausanne, 1015, Lausanne, Switzerland*

Michael Ferdman, *Stony Brook University, Stony Brook, NY, 11794, USA*

Boris Grot, *University of Edinburgh, Edinburgh, EH8 9AB, U.K.*

Luiz Barroso started his career at Digital Equipment Corporation, investigating workload-optimized multiprocessor server architectures marketed to enterprises in the 1990s. These high-margin, low-volume products lost their market to more cost-effective enterprise servers built from high-volume desktop CPUs riding Moore's law. The enterprise market has slowly transitioned to the cloud, where desktop PCs have formed the backbone of computing in data centers since the early 2000s to minimize cost and maximize the return on investment. Moving forward, with the absence of Moore's law, future servers require a clean-slate, cross-stack design to scale in compute, communication, and storage capacity while reducing operational, capital, and environmental costs.

Today, data centers form the backbone of all digital services. Data center growth is fueled by the global availability of online services, digitalization of industries, sustained data deluge, and AI. At the heart of data centers are volume servers that trace their roots back to the desktop PCs of the 1990s, running Linux and other open source software. Scaling computing with inexpensive PCs as building blocks to maximize the return on investment is analogous to RAID storage in the 1980s replacing high-margin enterprise disks with an array of cost-effective PC disks.

For the past two decades, data centers have enjoyed riding Moore's law and improvements in chip density and efficiency to scale their services while minimizing construction growth and overall energy consumption. Modern data centers heavily rely on the consolidation of workloads, virtualization, compression, and emerging software construction paradigms (e.g., microservices and serverless) to maximize server utilization while maintaining service-level contracts.

Unfortunately, with a slowdown in Moore's law[a] and chip efficiency, power consumption in servers is witnessing an unprecedented growth. Figure 1 depicts the thermal design power (TDP), a measure for the maximum amount of heat a chip can dissipate, for several generations of CPUs and GPUs from 2008 through 2023. While the figure does not plot TDP ratings before 2008, CPU TDPs remained flat, at 100–150 W, for roughly two decades. Since 2019, CPUs have been experiencing a sharp rise in TDP, suggesting that future improvements in design will be accompanied by higher power consumption, with implications for data center growth. GPUs' TDP is rising even faster because of their much higher arithmetic density as compared to CPUs.

The insatiable appetite for computing continues to grow, driven by, on one hand, the expansion in data collection, storage, and analysis and, on the other, the proliferation of AI. In the absence of Moore's law, these trends demand building and operating more servers, with a commensurate impact on overall server energy consumption and environmental impact through operational and embodied emissions.

To maximize silicon efficiency, post-Moore servers require a clean-slate approach to the design of the computing stack, from algorithms down to hardware. Much like the enterprise multiprocessor servers of the 1990s, post-Moore server design requires proper resource provisioning for workloads as well as precise metrics and measurement methodologies to quantify a server design's output based on the provisioned resources.

[a] https://en.wikipedia.org/wiki/Transistor_count

0272-1732 © 2024 IEEE
Digital Object Identifier 10.1109/MM.2024.3418975
Date of publication 1 July 2024; date of current version 14 October 2024.

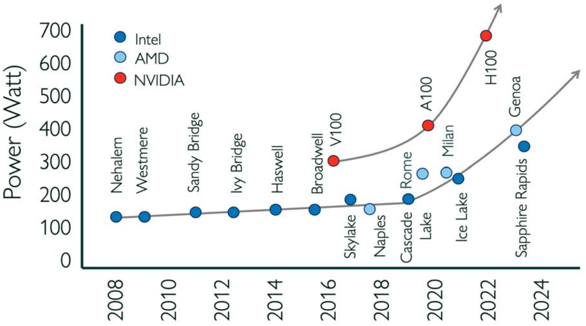

FIGURE 1. Increasing thermal design power in server CPUs and GPUs. I/O: input–output; Misc: miscellaneous.

In this article, we reflect on the convergence of server architecture from integrated enterprise-class multiprocessors of the 1990s to the modern-day cloud server architecture. We then present requirements for metrics to evaluate the performance and efficiency of post-Moore servers, emerging workloads, services, and computational paradigms in data centers as well as trends in the design of post-Moore servers.

> *TO MAXIMIZE SILICON EFFICIENCY, POST-MOORE SERVERS REQUIRE A CLEAN-SLATE APPROACH TO THE DESIGN OF THE COMPUTING STACK, FROM ALGORITHMS DOWN TO HARDWARE.*

FROM ENTERPRISE TO CLOUD

Server architecture has undergone a major transformation from the enterprise servers of the 1990s to the cloud servers of today. These transformations have been the result of exponential improvements in CPU performance and memory density, thanks to Moore's law, the demand for scalability from enterprise IT to the public cloud, and the focus on return on investment in data centers. In this section, we briefly go over these transformations, Luiz's contributions to these transformations, and our work during this period on introducing cloud-native server architecture.

Enterprise Servers

The primary driver for server architecture innovation in the 1990s was commercial server software, initially in the database management market and, eventually, the Web server market. With increasing data volumes, the growing gap between dynamic random-access memory (DRAM) and disk, and the exponential reductions in the cost of DRAM, these workloads primarily became in memory.

At the higher end of premium enterprise servers, vendors like Sun Microsystems, HP, Digital Equipment Corporation (DEC), and Silicon Graphics started building mid- (e.g., 4–16 CPUs) to large-scale (e.g., 128 CPUs) shared-memory multiprocessors to enhance memory capacity and ride Moore's law with DRAM and RISC CPUs. At the lower end of the premium market, x86 servers, whose CPU performance was competitive with their RISC counterparts, were riding the same Moore's law while benefiting from the economies of scale of desktop CPU volumes.

Many in the academic community were evaluating commercial server workloads on enterprise servers of the time and pointing out the mismatch between the workloads' requirements and the CPU microarchitecture, leading to underutilized silicon. These studies indicated that commercial server workloads benefited from more capacity in the cache hierarchies because of larger instruction and data working sets but showed limited instruction-level parallelism (ILP) due to dependent accesses to memory.

Luiz was among those who carefully evaluated this mismatch[2] and then argued and demonstrated, through DEC Piranha, the case for custom multicore processors for commercial workloads. These multicore processors could exploit thread-level (rather than instruction-level) parallelism with older generation (scalar) RISC pipelines and unusually large cache hierarchies required for commercial workloads.

While RISC CPUs had phenomenal penetration in the embedded and mobile platform market, they eventually lost ground to x86 CPUs in the enterprise server market. The RISC servers' high margins and low volumes rendered them economically unsustainable. Instead, Intel and AMD created a portfolio of CPUs that included enterprise server CPUs to help transition the market to x86.

Transition to the Cloud

The first decade of the 2000s witnessed the emergence of data centers as warehouse-scale computers, built with stripped-down volume servers interconnected with commodity network gear to implement massive Internet services, such as search, online retail, and social networks. The key goal was to build services with the cheapest building blocks available at the time (i.e., the desktop PCs) to maximize return on investment. The servers ran Linux and open source software and relied on distributed computing to exploit parallelism and provide fault tolerance.

A key mismatch between volume servers and the services they were running was rooted in the tight latency requirements in service-level agreements (SLAs) with customers.[4] The desktop PC and its operating system (OS) were either running in the nanosecond timescale computing and accessing memory or in the millisecond timescale, accessing a hard disk through the OS.[1] Meanwhile, the data that services were hosting in memory was made available through the networks, where switch-based fabrics operated in the microsecond timescale. Therefore, all OS overheads (e.g., thread context switches, interrupts, and network communication software stacks) in the microsecond timescale were directly on the critical path of the services.

IT companies scaled their services throughout this decade by building more and larger data centers at an unprecedented pace while maintaining low server utilization to guarantee SLAs. Volume servers are inherently not energy proportional, and, as such, low utilization implies high electricity consumption. More data centers also meant higher capital expenditure and overall lower return on investment.

In the decade of the 2010s, data center operators continued riding Moore's law with DRAM and CPUs, focusing on efficiency in infrastructure to maximize electricity delivery and optimizing software stacks to increase performance per dollar while maintaining SLAs. The two key technologies that improved server utilization were the consolidation of workloads and virtualization, allowing for multitenancy to sell unused resources to cloud customers.

Cloud-Native Server Architecture

With Dennard scaling ending and Moore's law slowing down, many academics continued evaluating opportunities to design servers that matched commercial server software requirements, including database and Web workloads. These included analyses of the core complexity, various forms of multithreading, and capacity provisioning in multicore cache hierarchies.

In 2011, Hardavellas et al.[8] revisited workload-centric server architecture. They demonstrated that, for the combination of stringent chip power constraints, emerging high-bandwidth and energy-efficient memory fabrics, and the abundance of request-level parallelism in server workloads, a custom many-core CPU would be optimal for throughput, power, and area in servers. Unlike conventional CPU designs, such custom CPUs would have minimal on-chip memory (i.e., megabytes) to hold the instruction working set of deep server software stacks and minimal complexity (i.e., power and area) cores to access off-chip data and exploit request-level parallelism across server threads.

Ferdman et al.[5] studied the mismatch between volume server architecture in the cloud and scale-out services based on open source software stacks. In particular, they listed several microarchitectural characteristics salient in scale-out workloads that were drastically different from desktop workloads: 1) the instruction supply bottleneck due to large instruction working sets in server software stacks; 2) low ILP and memory-level parallelism (MLP) in server software stacks; 3) secondary data working sets that are orders of magnitude larger than on-chip memory; and 4) low on- and off-chip per-thread memory traffic. The workloads were also found to make minimal use of floating-point arithmetic and vector operations. These characteristics were at odds with state-of-the-art volume servers at the time, which were being shipped with half a dozen ILP-centric desktop x86 cores padded with 12 MB of LLC per socket, with two sockets serving tens of gigabytes of DRAM. Figure 2 illustrates the silicon mismatch between the needs of scale-out services and a Westmere die.

These results laid the foundation for the first generation of cloud-native server CPUs,[9] called *scale-out processors*. Scale-out processors integrated multiple server "pods" sharing memory and input–output (I/O) ports on a single die. Figure 3 illustrates the anatomy of a 64-core scale-out processor with four pods. Each pod was optimized for throughput per area of silicon and uses three-way, out-of-order, 64-bit ARM cores modeled after Cortex A15, the flagship out-of-order ARM core at the time. The L2 was judiciously sized to capture the instruction working set and accelerate instruction supply while accommodating data reuse in the OS and application metadata. The cores were interconnected with a crossbar to statically interleaved L2 banks, memory controllers, and I/O bridges.

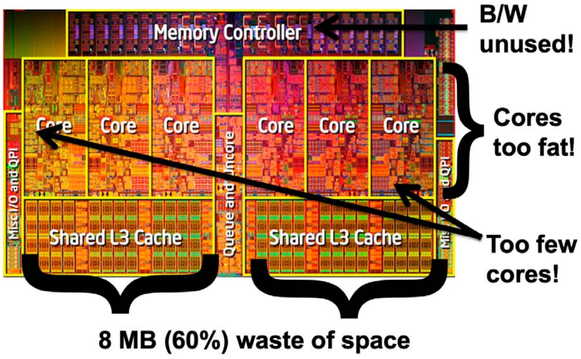

FIGURE 2. Server CPU microarchitecture/workload mismatch.

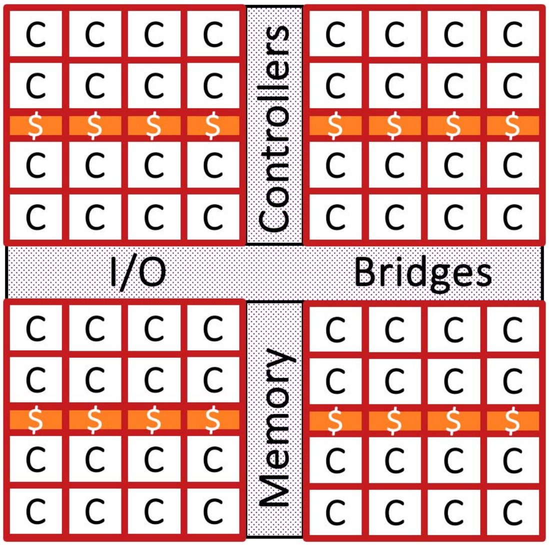

FIGURE 3. A scale-out processor with four 16-core pods. C: core.

Lotfi-Kamran et al.[9] further studied the on-chip network traffic for scale-out workloads and concluded that there is little (e.g., $<$ 2% on average) on-chip coherence traffic. The majority of traffic is either supplying instructions from the LLC or checking whether data are available on-chip and fetching it from off-chip because of the large disparity between on-chip and off-chip memory capacity. To maximize silicon efficiency and optimize interconnect topology, the authors proposed a custom on-chip interconnect for each pod to connect the cores in a column of request/reply relays to statically interleaved L2 banks. The L2 banks were fully connected with memory controllers and I/O bridges through a flattened butterfly network.

REGARDLESS OF THE GOAL OF A PROCESSOR DESIGN, THE PRIMARY TARGET IS ALWAYS TO MAXIMIZE PERFORMANCE WHILE STAYING WITHIN THE TARGET ENGINEERING COSTS AND TECHNOLOGY LIMITATIONS.

Each pod ran a full software stack and had its data sharded across physical memory partitions. Because pods operated as independent servers with no contention on microarchitectural resources, optimally sizing a pod and scaling the number of pods per chip would allow for a linear increase in throughput while maintaining the proper core-to-cache ratio and optimizing overall data center costs at the board and system level.

Compared to the contemporary designs, scale-out processors provided a much higher level of thread-level parallelism, more silicon area given to cores than caches, and a much faster instruction supply. The first generation of Cavium ThunderX followed this chip organization with a crossbar; in-order million-instructions-per-second cores (retrofitted with an ARM decoder); and larger L1 instruction caches for better instruction supply. Fast forward to now, with Intel announcing their Sierra Forest line of cloud-optimized processors featuring 144 single-threaded cores organized in four-core clusters that share an L2 and a shared LLC with a per-core capacity under 1 MB. The architecture of Sierra Forrest reflects the design philosophy of scale-out processors in terms of more cores, leaner cores, and less cache per core compared to conventional server CPUs.

METRICS

Regardless of the goal of a processor design, the primary target is always to maximize performance while staying within the target engineering costs and technology limitations. As such, a principled approach to designing next-generation cloud-native processors must first define and measure performance and then incorporate performance into metrics that will be used to optimize performance within the constraints of capital and operational cloud server expenditure.

Defining Performance

Luiz's work at DEC was among the first to consider servers as a separate class of systems with a unique set of characteristics,[2] demonstrating that server performance is dominated by memory stalls and limited ILP. On real systems, a barrage of simulated client requests can be used to assess the peak throughput the server can achieve, measured in queries or requests served per second. In a simulation, measuring multiple server requests end to end would be prohibitively slow. Moreover, the de facto CPU performance metric, instructions per cycle (IPC), allowing for fine-grain measurements of tens to hundreds of thousands of instructions at a time, would not be applicable in multiprocessor server workloads because synchronization in the software stack with threads spinning perturbs IPC.

While metrics such as MLP could help indicate memory bottlenecks in workloads, they fell short of offering tight estimates on overall performance. Wenisch et al.,[11] instead, showed that the number of user IPC (U-IPC) is proportional to server throughput for a wide range of server applications, such as running online transaction processing, decision support systems, and Web workloads generating dynamic content, because threads spin in the OS. As such, U-IPC could be used reliably as a microarchitectural metric for throughput with simulation.

With the transition to the cloud and the birth of global IT services, from search to social media, retail, and media, the customer experience in terms of response time became a negotiated metric in cloud SLAs. Strict quality of service (QoS) requirements were set on the response time of servers, setting bounds not just on the average latency but on request tail latency. Queuing theory dictates that tail effects become prominent at high server utilization. As a direct consequence, cloud operators must sacrifice throughput to meet tail latency constraints, intentionally running servers at reduced utilization to meet latency targets.

Today, the key metrics for cloud workloads are peak throughput at a given QoS target, where the target is defined as a maximum acceptable end-to-end latency for a specified fraction (e.g., the 95th percentile) of the requests. Luiz's work at Google pointed to the cumulative impacts of tail latencies at scale compounding into large end-to-end latencies. His work showed the critical impact of tail latencies in data centers,[4] where back-end services must meet much stricter QoS targets (e.g., the 99.9th percentile) to permit end-to-end latency targets (e.g., the 95th percentile) to be met. In practice, production servers rarely operate at their peak theoretical throughput, highlighting the need to design processors that specifically target the relevant metrics so as to optimize for efficiency and overall cost.

Silicon Efficiency

The widening gap between mainstream CPUs and the needs of cloud workloads inspired a critical look at how to quantify the performance and efficiency of cloud-native processors. With silicon area being a direct reflection of the processor cost, we proposed the *performance density* metric, which relates the QoS-conscious peak throughput that can be achieved by a square millimeter of silicon. Using performance density as a road map to understanding the cloud-native CPU landscape suggests processors with many lean cores and limited cache capacity.[9] The reluctance of the dominant server CPU vendors to shift toward designs with higher performance density and make more effective use of silicon led to a rapid emergence of a number of competing CPU designs targeting the cloud market, including the Cavium ThunderX, Amazon Graviton, link Altra, and Huawei TaiShan.

As chip power envelopes continue their exponential increase, and Moore's law crawls toward its ultimate demise, cloud-native processor design is approaching another inflection point. Integrating more silicon per server node increases the fabrication cost, but, with increasing clock frequencies and diminishing improvements in memory density, the design constraints shift away from processor silicon area toward overall power consumption and memory bandwidth.

Table 1 compares performance density and performance/power on CloudSuite 4.0 and DeathStarBench Media Service for a single core of AMD Zen 3 and Ampere Altra, both normalized to Intel Ice Lake. Ice Lake, Zen 3, and Altra are 6.2, 4.0, and 1.40 mm^2 in area per core in comparable technology nodes. The TDP of each chip is divided by the number of cores to get 5.78, 3.52, and 2 W, respectively, for each Ice Lake, Zen 3, and Altra core. The cores are all run at 2.4 5 GHz. Throughput is measured for a single query (i.e., for data, media, and Web services) or the entire workload execution time (i.e., for analytics).

> *LUIZ'S WORK AT GOOGLE POINTED TO THE CUMULATIVE IMPACTS OF TAIL LATENCIES AT SCALE COMPOUNDING INTO LARGE END-TO-END LATENCIES.*

The table indicates that, due to its more efficient x86 core design, Zen 3, on average, improves both metrics over Ice Lake. Altra, utilizing a narrower out-of-order ARM pipeline, improves performance density by up to 4× and performance/power by up to 2.65×. These numbers for Ampere are conservative because Altra uses

TABLE 1. Single-core performance density and performance/power for the AMD Zen 3 and Ampere Altra normalized to the Intel Ice Lake.

Workload	Throughput/mm^2		Throughput/W	
	Zen 3	Altra	Zen 3	Altra
Data analytics	1.38	3.87	1.46	2.53
Data caching	1.28	2.06	1.35	1.34
Data serving	1.22	1.94	1.29	1.27
DeathStarBench media service	0.89	1.53	0.95	1.00
Graph analytics	1.55	3.31	1.65	2.16
In-memory analytics	1.52	3.86	1.61	2.52
Media streaming	1.09	1.82	1.16	1.19
Web search	1.46	4.06	1.54	2.65
Web serving	1.37	4.04	1.46	2.64

stock ARM cores. AmpereOne with custom ARM cores for servers is slated to have higher silicon efficiencies.

EMERGING WORKLOADS AND PARADIGMS

At Google, Luiz helped architect the *scale-out* workload paradigm. Scale-out emerged from the need to cost-effectively deliver online services to a global base of users. By their very nature, these services were complex (e.g., Web search) yet needed to meet tight latency constraints and unprecedented throughput objectives. Luiz and colleagues demonstrated that scalability can be achieved by decomposing complex applications into stand-alone logical components, each of which can be deployed and scaled independently to offer the desired performance and reliability objectives in software.

While the scale-out paradigm enabled greater degrees of scalability and modularity compared to traditional enterprise *monoliths*, typical scale-out architectures featured just two or three tiers, with each tier responsible for a significant degree of the overall functionality. With increasing complexity and commoditization of online services, there emerged a need for greater modularity and elasticity at the software level, giving rise to microservices.

Microservices ushered in an era of truly disaggregated applications, each deployed as a graph of tens or even hundreds of nodes communicating via remote procedure calls (RPCs). The high degree of modularity endemic in microservices facilitated composability, whereby off-the-shelf pieces of software (potentially written in different programming languages) could be mixed and matched to rapidly engineer and deploy complex services. High modularity also enabled greater degrees of elasticity, empowering developers to scale different parts of their application separately as a function of the load placed on each microservice.

A performance analysis of microservices by Gan et al.[6] observed that the network stack accounted for a significant fraction of cycles due to frequent RPCs—the price of disaggregation. The trends at the microarchitectural level largely echoed those for scale-out and enterprise workloads, with a high fraction of stall cycles in the core front end and long-latency accesses to the main memory for data, both of which contribute to low IPC.[5] Microservices also amplify the tail latency problem because a single slow server in a long chain of RPCs can compromise the end-to-end latency.

In recent years, *serverless* computing has emerged as a popular way to deploy complex services in the cloud. Similar to microservices, a serverless application comprises a graph of loosely connected functions that communicate over RPCs. In serverless, however, functions are stateless, with any mutable state maintained in a separate service.

Serverless offers solutions to key drawbacks of microservices. First, traditional microservice instances are often housed within full-featured virtual machines (VMs) that can take considerable time to boot. This prolonged boot time can compromise latency targets, especially during sudden load spikes. Developers often provision additional VM instances as a precaution against load fluctuations, leading to overprovisioning and unnecessary costs as VMs are billed regardless of whether they are idle or active.

Serverless addresses these challenges by leveraging its stateless nature and lightweight *micro-VMs*, facilitating rapid on-demand scaling with instance boot times measured in seconds rather than minutes. Moreover, developers are billed solely for the actual runtime and memory usage of their functions, encouraging them to optimize instances' runtime and memory footprints. Consequently, serverless functions exhibit short execution times, often in milliseconds, and small memory footprints,[10] enhancing server utilization through efficient bin packing. However, as serverless adoption increases, it will necessitate adjustments in OS and hardware to efficiently handle the growing number of tasks and remote data access, ensuring optimal performance and resource utilization while managing potential inefficiencies in the software stack.

AI workloads represent a substantial portion of data centers' computational resources and energy consumption, with anticipated growth driven by the rise of generative AI technologies. To address these requirements, major cloud providers, such as Google, Microsoft, Amazon, and Huawei, have introduced AI accelerators, offering higher silicon efficiency compared to CPUs. AI workloads also align with trends favoring disintegration for the scalability of services and resource optimization, leading to their deployment as microservices and serverless instances. However, this trend also underscores the need for solutions capable of managing diverse requirements for computational resources while also supporting the frequent use of OS (e.g., memory management and scheduling) and network services (e.g., RPC).

POST-MOORE SERVERS

There are many opportunities to optimize silicon efficiency in the post-Moore era. These opportunities can bridge the gap between the workloads and emerging silicon technologies to minimize both capital expenditure (i.e., the amount of silicon used) and operational expenditure (i.e., the energy used) in data centers.[3]

Today's servers inherit their hardware and OSs from the desktop PCs of the 1990s, where silicon is fragmented across the hardware and OS boundary lines. This architecture fundamentally suffers from the same mismatch between the nanosecond timescale of CPU and memory with the millisecond timescale of OS services handling I/O devices (mentioned in the "From Enterprise to Cloud" section). Because data centers' primary function is processing, managing, and storing distributed data, accessing these data hosted in the server's memory requires CPU and OS intervention. Moreover, discrete I/O devices require their own memory (and CPU in the case of network interface cards), fragmenting silicon across the server and creating multiple copies of data. Finally, because the flow of data among services is not exposed to the OS within the server, the OS is oblivious to how CPU resources are scheduled and is unable to minimize OS scheduling overhead and maximize instruction and data affinity.

To reduce data's overall required silicon footprint, there are three post-Moore design pillars for servers: 1) *integration*, tightly connecting functionality with data in memory to reduce movement; 2) *specialization*, tailoring silicon resources for common functionality for data; and 3) *approximation*, optimizing silicon resources based on the expected output quality for computation. In the following, we present examples of opportunities to pursue with these pillars to optimize silicon efficiency in post-Moore servers.

Compute

While silicon density continues to improve slowly with new fabrication technologies, these technologies are becoming increasingly expensive. Chiplets have emerged as a promising approach to reduce the cost of fabrication both through enhancing yield (because yield is a function of die area) and the use of older technology nodes for the less latency-critical functionality. Chiplets also present an opportunity to build a single server node with disaggregated specialized functionality and tighter integration to reduce data movement across services. Services with large instruction supply bottlenecks, large off-chip datasets, and little ILP and MLP can use a scale-out pod[9] chiplet architecture with more logic-to-memory silicon provisioning. To accommodate a larger number of cores per chiplet, chiplet design can be tailored with a minimal frequency to achieve the desired tail latency in a disaggregated service. Microservices and serverless workloads would benefit from a chiplet architecture with hardware support for core-to-core connectivity, RPC operations, and a tighter integration with the network interface.

Moreover, the energy required to fetch a word of off-chip memory is four or more orders of magnitude larger than a fixed-point add operation. Popular analytic operators (e.g., in Apache Spark) filter, map, aggregate, join, or sort data in memory, which are distributed over multiple servers. These operators are embarrassingly parallel, require only fixed-point arithmetic, can easily consume the available memory bandwidth, and do not exhibit much reuse in the cache hierarchy. Custom chiplets for analytic operators can be properly integrated both near the memory to utilize the available memory bandwidth and near the network interface card to stream the results out in case of a request from a remote server.

Finally, numerical encoding is a fundamental enabler in reducing data's overall computational, communication, and storage requirements. There is a difference of two orders of magnitude in the energy and area footprint for 2-bit versus 32-bit arithmetic operations. Much computation centered around data operates on data that are statistical in nature (e.g., originating from sensors or simulation). Similarly, computational outcome is often statistical in nature, gauged by a given quality metric (e.g., training, inference, statistical ranking, or media). Identifying algorithmic techniques for approximation together with innovation in numerical encoding and microarchitecture can help maximize computational density and minimize the required memory capacity and bandwidth. Modern GPUs are increasingly supporting both multiple scaled numerical formats and sparsity to allow algorithms to minimize the computational, memory capacity, and bandwidth requirements of large language models.

Memory

Memory is roughly 50% of the cost of the server today and has been highly cost-sensitive since the inception of data centers. In recent years, the cost per gigabyte has plateaued due to the slowdown in Moore's law, further pushing up the overall memory price per server. There are many opportunities to explore technologies that enable both a more efficient use of memory and its capacity, directly improving the return on investment in servers.

Memory pooling is one such paradigm that has great potential. Many studies in the past have examined rack-level memory pooling to mitigate a load imbalance in data serving software stacks due to a skew in partitioning or object popularity. Recent research suggests that, in the cloud environment, memory is wasted in rented VMs because customers don't fully use the memory capacity and in rented physical servers where all CPUs are rented and memory can be stranded. Tighter integration of the network interface with the CPU would enable the OS to implement memory pooling and optimize capacity management. CXL also offers a promising scalability path for memory pooling with transaction-oriented serial links in future generation PCIe protocols, as traditional parallel interfaces in DDR are prohibitively pin intensive.

Virtualization is a foundational technology built on top of virtual memory to provide isolation for both conventional cloud containers and emerging serverless/microservices paradigms. Confidential computing also fundamentally relies on virtualization for protection. Modern page-based virtual memory dates back to the 1960s, and its current incarnations not only incur significant silicon footprint for TLBs (now in thousands of entries per core) but also operate on a microsecond timescale for page-based memory management operations (e.g., changing permissions).

Intermediate address spaces are a promising approach to enabling nanosecond-scale memory management operations[7] while eliminating the silicon footprint of page-based virtual memory. Emerging workloads, such as microservices, serverless functions, and virtualized network functions, can forego virtualization and specialize hardware/software for nanosecond-scale compartments using an intermediate address space. The ability to grant and revoke access to memory on the nanosecond timescale means that data can be supplied from across multiple domains from the same physical copy, thereby eliminating data copy operations and providing a tighter integration between the application and the system.

Data Center Tax

A common set of operations in data centers accounts for 30% of all CPU cycles. These operations, referred to as *data center tax*, are centered around the management and communication of data. These include CPU cycles in the data center spent moving and operating data among the CPU, memory, and network. Both DDR and CXL (with PCIe) bandwidth are slated to grow by more than 20% per year by 2030. Network bandwidth is also slated to grow by 20% in the next decade thanks to emerging fabrics. With these bandwidth trends, the sequential software running on the CPU to perform the data center tax will be a growing bottleneck. Mitigating the data center tax requires proper integration of specialized logic among the CPU, memory, and network to accelerate these operations and minimize data movement.

Many common data center tax operations are related to memory management. These include memory allocation, copying, compression/decompression, encryption/decryption, and scatter/gather operations. Because RPC is the most common mechanism of communication, the data center tax also includes data serialization/deserialization to allow microservices to translate their object formats in memory as they communicate. Recent products, such as Intel Sapphire Rapids, have started incorporating examples of such accelerators.

There are promising avenues to co-design custom processors for RPC.[12] These processors, when integrated with the network interface and the CPU in the same coherence domain, can minimize data movement. Figure 4 depicts the anatomy of the Cerebros RPC processor tightly integrated with the network interface logic and CPU. The processor executes the RPC layer (e.g., Apache Thrift or Google Protobuf) and acts as an intermediary stage between the network interface and the microservice running on the CPU. The processor not only processes RPC requests at line rate (e.g., 100 Gb/s), it also optimizes service throughput by monitoring and balancing load on cores and using affinity scheduling to improve locality.

Beyond memory allocation and management, common OS operations, such as thread scheduling and interrupt delivery, also consume a significant fraction of CPU cycles. Modern servers consolidate thousands of threads, which the OS has to manage within a coherence domain. Most importantly, the OS cannot maximize affinity and minimize data movement and interrupt delivery overhead because it has little information about the application's communication patterns. The combined effect of increasing core counts and emerging paradigms (i.e., microservices and serverless computing) exacerbates the overhead of thread management, scheduling, and interrupt delivery. Innovation in OS

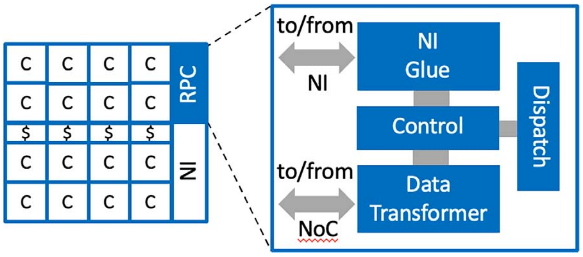

FIGURE 4. The anatomy of the Cerebros RPC processor. NI: network interface; NoC: network on chip.

together with microarchitecture will be instrumental in mitigating the OS-centric data center tax.

CONCLUSION

Server architecture has undergone major transformations from the enterprise servers of the 1990s to the cloud servers of today. These transformations have occurred thanks to exponential improvements in CPU performance and memory density with Moore's law, the demand for scalability from enterprise IT to the public cloud, and the focus on return on investment in data centers. Luiz played an indispensable role in creating and realizing these transformations. Continuing Luiz's legacy in the post-Moore era requires a clean-slate approach to the design of the computing stack, from algorithms all the way down to hardware, to maximize silicon efficiency and address the needs of emerging workloads with proper design metrics.

REFERENCES

1. L. Barroso, M. Marty, D. Patterson, and P. Ranganathan, "Attack of the killer microseconds," *Commun. ACM*, vol. 60, no. 4, pp. 48–54, Mar. 2017, doi: 10.1145/3015146.
2. L. A. Barroso, K. Gharachorloo, and E. Bugnion, "Memory system characterization of commercial workloads," in *Proc. 25th Int. Symp. Comput. Archit.*, 1998, pp. 3–14, doi: 10.1145/279358.27936.
3. L. A. Barroso, U. Hölzle, and P. Ranganathan, *The Datacenter as a Computer: Designing Warehouse-Scale Machines* (Synthesis Lectures on Computer Architecture), 3rd ed. San Rafael, CA, USA: Morgan & Claypool, 2018.
4. J. Dean and L. A. Barroso, "The tail at scale," *Commun. ACM*, vol. 56, no. 2, pp. 74–80, 2013, doi: 10.1145/2408776.2408794.
5. M. Ferdman et al., "Quantifying the mismatch between emerging scale-out applications and modern processors," *ACM Trans. Comput. Syst.*, vol. 30, no. 4, pp. 1–24, 2012, doi: 10.1145/2382553.2382557.
6. Y. Gan et al., "An open-source benchmark suite for microservices and their hardware-software implications for cloud & edge systems," in *Proc. 24th Int. Conf. Archit. Support Program. Lang. Operating Syst.*, 2019, pp. 3–18, doi: 10.1145/3297858.3304013.
7. S. Gupta, A. Bhattacharyya, Y. Oh, A. Bhattacharjee, B. Falsafi, and M. Payer, "Rebooting virtual memory with Midgard," in *Proc. 48th Int. Symp. Comput. Archit. (ISCA)*, Valencia, Spain, 2021, pp. 512–525, doi: 10.1109/ISCA52012.2021.00047.
8. N. Hardavellas, M. Ferdman, B. Falsafi, and A. Ailamaki, "Toward dark silicon in servers," *IEEE Micro*, vol. 31, no. 4, pp. 6–15, Jul./Aug. 2011, doi: 10.1109/MM.2011.77.
9. P. Lotfi-Kamran et al., "Scale-out processors," in *Proc. 39th Int. Symp. Comput. Archit.*, 2012, pp. 500–511, doi: 10.1145/2366231.233721.
10. M. Shahrad et al., "Serverless in the wild: Characterizing and optimizing the serverless workload at a large cloud provider," in *Proc. USENIX Tech. Conf.*, 2020, pp. 205–218.
11. T. F. Wenisch, R. E. Wunderlich, M. Ferdman, A. Ailamaki, B. Falsafi, and J. C. Hoe, "SimFlex: Statistical sampling of computer system simulation," *IEEE Micro*, vol. 26, no. 4, pp. 18–31, Jul./Aug. 2006, doi: 10.1109/MM.2006.79.
12. A. P. Zarandi, M. Sutherland, A. Daglis, and B. Falsafi, "Cerebros: Evading the RPC tax in datacenters," in *Proc. 54th IEEE/ACM Int. Symp. Microarchit.*, 2021, pp. 407–420.

BABAK FALSAFI is a professor in the School of Computer and Communication Sciences, Ecole Polytechnique Federale de Lausanne, 1015, Lausanne, Switzerland and the founder of EcoCloud, an industrial and academic consortium investigating post-Moore datacenter technologies. Falsafi received his Ph.D. degree in computer science from the University of Wisconsin-Madison. He is a Fellow of IEEE and the Association for Computing Machinery. Contact him at babak.falsafi@epfl.ch.

MICHAEL FERDMAN is an associate professor of computer science at Stony Brook University, Stony Brook, NY, 11794, USA. His research focuses on the design of server systems and FPGA hardware accelerators. Ferdman received his Ph.D. degree in electrical and computer engineering from Carnegie Mellon University. He is a Senior Member of IEEE and a member of the Association for Computing Machinery. Contact him at mferdman@cs.stonybrook.edu.

BORIS GROT is a professor in the School of Informatics at the University of Edinburgh, Edinburgh, EH8 9AB, U.K. His research interests include server hardware and software stacks, and networking and datacenter-scale computing. Grot received his Ph.D. degree from the University of Texas at Austin. He is a Senior Member of IEEE. Contact them at boris.grot@ed.ac.uk.