Advances in Modern Computer Architecture

Parangat Mittal
University of California Los Angeles
parangat@ucla.edu

ABSTRACT

The field of Computer Architecture is undergoing a transformative shift, driven by an increasing demand for performance and efficiency in modern applications. Key milestones in this evolution include the transition from CISC to RISC architectures, advancements in GPU, and the emergence of AI accelerators. The development of full-scale custom chips showcases the potential for optimized performance and innovation across various fields. Despite the enduring relevance of x86 architecture in certain areas, the rise of custom silicon promises to define the future of computing, ushering in a "Golden Age" characterized by enhanced capabilities and diverse applications. This paper provides a comprehensive overview of these advancements, offering insights into the technological innovations that are reshaping the computing landscape.

KEYWORDS

Computer Architecture, Custom Silicon, RISC ISA, AI Accelerators, GPU

1 Introduction

In an increasingly data-driven and AI-centric world, delivering a seamless user experience has become paramount. This modern shift demands new levels of performance, efficiency, and power from the computing systems. While software has evolved rapidly to keep pace with these demands, it has become imperative for hardware to innovate as well. Computer Architecture has transitioned from early mainframes to personal computers to an era of specialized hardware with custom Silicon chips.

Conventional processors, while versatile for general-purpose computing, are struggling to keep pace with the exponential growth of data and the rising complexity of workloads, including the recent generative AI. These processors face significant challenges in delivering the necessary performance and efficiency required for modern applications. In response, the development of new and specialized hardware has given rise to chips tailored for specific domains. This shift marks a departure from general-purpose processors, towards more dedicated and optimized solutions that can meet the specific demands of complex computing tasks.

1.1 End of Scaling Laws

Traditional off-the-shelf processors, primarily dominated by the x86 architecture, have reached their limits in addressing the demands of modern use cases. The significant factors contributing to this limitation are the end of Moore's Law, the breakdown of Dennard Scaling and the implications of Amdahl's Law (Figure 1).

1.1.1 Moore's Law. Moore's Law, an observation made by Gordon Moore in 1965, postulated that the number of transistors on a chip would double approximately every two years, leading to exponential growth in computing power. However, this exponential growth has significantly slowed as transistors approach atomic scales. Physical and quantum limitations make it increasingly difficult to continue shrinking transistors, meaning that packing more transistors into chips no longer guarantees the same performance gains. This slowdown necessitates alternative approaches to advance computing capabilities.

1.1.2 Dennard Scaling. Dennard Scaling, proposed by Robert Dennard in 1974, predicted that as transistors become smaller and smaller, their power density remains constant, allowing for improvements in performance and energy efficiency. However, this scaling law began to break down owing to issues such as increased leakage currents and heat dissipation, and led to a plateau in power efficiency improvements. The failure of Dennard Scaling has created a significant challenge for maintaining the energy efficiency of modern processors.

1.1.3 Amdahl's Law. Amdahl's Law, formulated by Gene Amdahl in 1967, addresses the limitations of parallel processing. It states that the maximum speedup of a task achieved by parallel computing is limited by the sequential portion of the task. Even with an infinite number of parallel processors, the speedup is bounded by the tasks that must be performed sequentially. This law underscores the diminishing returns of adding more processing units to improve performance, becoming a bottleneck as multi-core processors are used to achieve desired compute performances.

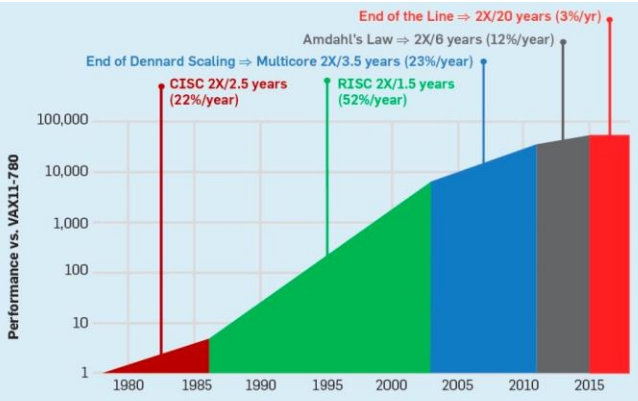

Figure 1: Performance gains guided by physical scaling laws

2 Timeline of Advancements in Silicon

The journey of computer architecture has been accentuated by the development of specialized processors designed to meet specific needs. The following timeline highlights some of key milestones in this evolution:

1980s-1990s: Early Domain-Specific Chips

1980s: Introduction of the Application-Specific Integrated Circuit (ASIC), which paved the way for the creation of custom chips tailored for specific applications.

1990s: Rise of Graphics Processing Units (GPUs), initially designed for rendering graphics but later adapted for general-purpose computing tasks.

2000s: Specialized Processors and Early AI Chips

2001: Introduction of the IBM Cell processor, designed for the PlayStation 3, which marked an early attempt at creating a processor for specific high-performance applications.

2006: Nvidia's G80 GPU architecture, which significantly improved the computational capabilities of GPUs, paving the way for their use in general-purpose computing.

2010s: Emergence of AI and Machine Learning Accelerators

2011: Launch of the Nvidia Tegra 3, one of the first mobile processors to integrate a GPU for enhanced mobile gaming and computational tasks.

2015: Google introduces the Tensor Processing Unit (TPU), a custom ASIC designed specifically for accelerating machine learning workloads in its data centers.

2016: Introduction of the Apple A10 Fusion chip, featuring a custom-designed CPU and GPU for enhanced performance.

2017-2020: Rise of Custom Silicon in Consumer and Enterprise Markets

2017: AWS introduces ARM-based Graviton chip for cloud services, shifting towards custom silicon in the data center.

2018: Apple A12 Bionic chip, featuring the Neural Engine, a custom-designed component for AI and machine learning tasks.

2019: Qualcomm launches Snapdragon 8cx, a custom ARM-based chip for Windows laptops, pushing Windows-on-ARM ecosystem.

2019: AWS releases Inferentia, a custom chip designed for high-performance inference in AI applications.

2020-2024: State-of-the-Art Custom Silicon

2020: Apple announces the M1 chip, its first custom ARM-based processor for Mac computers, marking a significant shift away from x86 architecture.

2021: Nvidia launches the A100 Tensor Core GPU, designed for AI, data analytics, and high-performance computing (HPC) applications.

2022: Microsoft announces Project Brainwave, a custom AI chip for accelerating deep learning models in its Azure cloud platform.

2023: Meta unveils the MTIA (Meta Training and Inference Accelerator), designed to optimize AI training and inference workloads.

2024: Qualcomm introduces the Snapdragon X Elite chip, a high-performance ARM-based processor for Windows laptops, showcasing significant advancements in power efficiency and performance.

3 Overview of Technological Advancements

3.1 Transition from CISC to RISC

The transition from Complex Instruction Set Computing (CISC) to Reduced Instruction Set Computing (RISC) has been a pivotal change in computer architecture. CISC, dominated by x86 architecture, focuses on complex instructions that can perform multiple low-level operations. However, RISC ISAs, such as ARM and RISC-V, have simpler instructions executed more efficiently, allowing improved power efficiency and performance. RISC architectures often excel in executing specific domain kernels in a simpler manner, making them well-suited for modern workloads.

3.2 Evolution of GPUs

GPUs have evolved from being purely graphics processors to becoming essential components for general-purpose computing. Initially integrated into motherboards as Integrated GPU, they are now used alongside CPUs as dedicated GPUs capable of handling parallel processing tasks, making them ideal for AI, ML, and scientific computing. While they have made significant progress in media and graphics rendering, they have now become indispensable for training complex neural networks and accelerating large-scale matrix operations.

3.3 Specialized Accelerators

Domain-specific accelerators are designed to offload specific tasks from the CPU, for better performance. These accelerators are tailored to specific computational patterns and data types, enabling them to achieve higher throughput and efficiency than general-purpose CPUs on the same tasks. Specific examples include Tensor Processing Units (TPUs) for AI and Field Programmable Gate Arrays (FPGAs) for reconfigurable hardware acceleration.

3.4 Piecing it all together: Custom Chips

Custom silicon chips combine the strengths of CPUs, GPUs, and accelerators onto a single hardware, tailored to meet specific domain requirements. These allow for unprecedented levels of optimization and performance. Custom silicon reduces latency, enhances data transfer rates, and optimizes power consumption, leading to significant performance gains over generic processors. Every major big tech company has now pivoted towards development of Custom Silicon chips, which reaps economic and strategic benefits apart from obvious technological benefits.

4 RISC Advancements

4.1 Transition from CISC to RISC

The transition from Complex Instruction Set Computing (CISC) to Reduced Instruction Set Computing (RISC) architectures, like ARM and RISC-V, offers several advantages. RISC architectures feature simplified instruction sets, leading to more efficient hardware implementations, reduced chip complexity, lower power consumption, and potentially higher clock speeds. This simplicity also makes RISC architectures well-suited for pipelining and parallel execution of instructions, enabling higher performance through efficient instruction scheduling and management. Open ISAs like RISC-V offer extensions that enable the implementation of customized instructions tailored to specific applications, providing granular optimization for domain-specific hardware.

4.2 Porting Applications to ARM/RISC-V

Porting applications to ARM and RISC-V architectures involves addressing compatibility issues and optimizing code for the new instruction set architectures. While challenging, advancements in development tools and support ecosystems have made this transition more feasible. The Arm Total Design Ecosystem and open-source toolchains for RISC-V play a critical role in facilitating this process, providing developers with the resources needed to adapt their software to new architectures, and promote rapid adoption of RISC-based devices.

4.3 Off-the-shelf Solutions

Several vendors provide off-the-shelf ARM and RISC-V chips for general-purpose computing. ARM's ecosystem, with the ARM Cortex series, is widely used in mobile devices. Notable examples include Qualcomm Snapdragon processors, known for high performance and power efficiency; Ampere Altra processors, designed for cloud and edge computing; and Amazon Graviton processors in AWS data centers, offering cost-effective solutions for cloud workloads. SiFive, a key RISC-V provider, offers general-purpose chips such as the U-series for embedded and IoT devices, the E-series for edge computing, and the P-series for high-performance computing. The customizable nature of RISC-V allows developers to create tailored solutions, combining performance and efficiency in a versatile package.

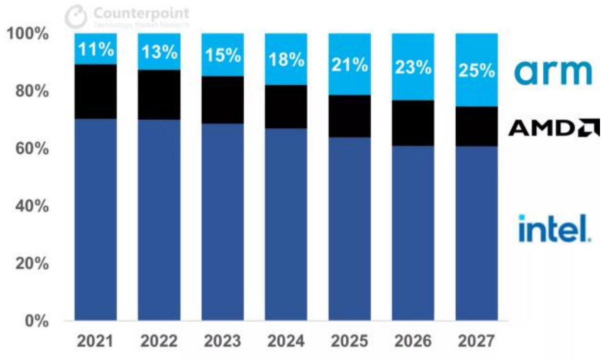

Figure 2: Market Share of Arm-based PCs

5 GPU Advancements

5.1 GPU Functionality

GPUs are specialized processors designed to handle media and graphical tasks. They excel at handling large amounts of data simultaneously, making them ideal for tasks like graphics rendering, scientific simulations, and machine learning. GPUs typically consist of thousands of smaller cores, each capable of performing simple calculations. These cores work together in parallel to execute a large number of instructions simultaneously.

GPUs interface with the CPUs through the system's bus, allowing for data transfer and coordination between the two processors. These have specialized ISAs which facilitate the parallel execution easier when certain tasks are offloaded. This offloading process involves transferring data to the GPU, instructing it to perform the calculations, and then retrieving the results.

5.2 NVIDIA's Contributions

NVIDIA has been at the forefront of GPU advancements, right from the early RIVA GPUs in 1997 to the latest generation of GeForce and RTX family of GPUs. The core contributions made by NVIDIA in this space:

5.2.1 CUDA Architecture. CUDA is a parallel computing platform and programming model developed by NVIDIA. It enables developers to harness the power of GPUs for general-purpose computing tasks. CUDA provides a set of libraries, tools, and APIs that simplify GPU programming and accelerate applications in various domains.

5.2.2 Tensor Core. Introduced in the Volta architecture, Tensor Cores are specialized processing units within NVIDIA GPUs designed to accelerate deep learning and AI workloads. They are optimized for matrix operations, which form the core of deep learning algorithms. These cores significantly improve the performance of training and inference tasks for neural networks.

6 Accelerators and Coprocessors

6.1 Functionality of the Accelerators

Accelerators are designed to handle specific tasks more efficiently than general-purpose CPUs. They work alongside CPUs, offloading intensive computations and enhancing overall system performance. They are similar to GPUs in the sense of co-processing, but unlike GPUs they are designed for a diverse range of tasks and workloads, not just graphics. They typically focus on a narrow set of operations, allowing them to achieve higher performance and efficiency for those specific tasks. Accelerators can be integrated into the CPU or implemented as separate chips.

6.2 Google Tensor Processing Units (TPU)

TPUs are custom-designed to accelerate machine learning workloads. They are systolic arrays designed to accelerate the matrix operations, integrate tightly with Google's software stack, providing optimized performance for AI applications. TPUs offer high scalability, and energy efficiency for training and deploying large-scale machine learning models.

TPUs are used extensively in Google's data centers for training and deploying machine learning models, offering significant performance improvements over traditional hardware (Figure 3). The specialized design of TPUs enables faster training times and more efficient inference, supporting the rapid development and deployment of AI technologies.

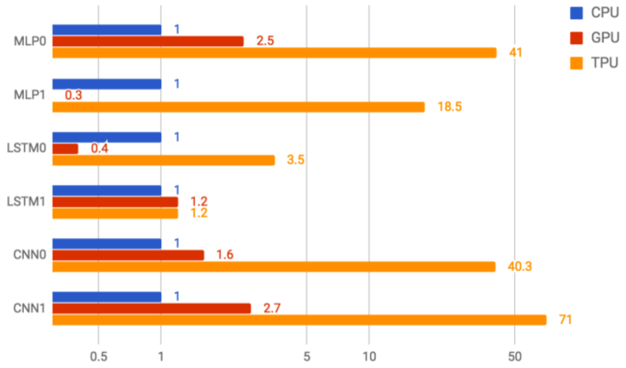

Figure 3: Performance comparison of CPU, GPU and TPU

6.3 AWS Inferentia and Trainium

6.3.1 Inferentia. AWS Inferentia is designed to accelerate inference tasks, providing high performance and low latency for deploying machine learning models. Inferentia delivers 2.3x higher throughput and up to 70% lower cost per inference, compared to other off-the-shelf processors. It offers 50% better performance/watt, making it a sustainable hardware for enterprises.

6.3.2 Trainium. AWS Trainium focuses on AI training workloads, offering substantial cost and performance benefits compared to traditional GPU-based solutions. Training is the costliest step in deploying any AI application, both in terms of time and money. This hardware solves the problem of limited budgets by offering 50% cost-to-train savings over comparable instances. Trainium adaptively supports a range of data types like FP32, TF32, BF16, FP16, UINT8, and the new configurable FP8, making it highly accurate with minimal to no loss of accuracy.

6.4 Meta's Recommender Hardware

Meta Training and Inference Accelerator (MTIA) chip is a cutting-edge AI accelerator specifically designed to enhance the performance of AI tasks. The chip features an 8x8 grid of Processing Elements (PEs), connected by a high-speed mesh network. These PEs are optimized for AI functions pertinent to Meta's Deep Learning Recommendation Models (DLRMs), including matrix multiplication, non-linear functions, and data movement. MTIA integrates RISC-V CPU cores for efficient control and coordination of these tasks. This innovative design offers 3x speed improvement over the previous generation while operating at 90W, significantly increasing compute density and performance efficiency.

A notable innovation is the integration of open-source software optimized for Meta's PyTorch AI framework, utilizing the Triton Compiler to generate high-quality kernels for the MTIA hardware. This integration enhances performance and fosters broader adoption within the AI community. By positioning itself with the MTIA chip, Meta gains a competitive advantage in the AI accelerator market, challenging industry leaders like NVIDIA and driving further innovation within the AI ecosystem.

7 Custom Full-Scale Chips

7.1 Functionality of Custom Chips

Custom silicon chips integrate CPUs, GPUs, and accelerators onto a single chip to deliver optimized performance for specific applications. These chips operate within a specialized tech stack which provides enhanced efficiency and flexibility. They are typically developed using a combination of custom-designed and licensed intellectual property (IP) blocks.

7.2 Apple M-Series Chips

Apple M1 chip featured an 8-core CPU with high-performance and efficiency cores, an 8-core GPU, and a 16-core Neural Engine for machine learning tasks (Figure 4). This shift from Intel-based Macs delivered up to 3.5 times faster CPU performance and double the battery life. One of the notable features include the Unified Memory Architecture, which allows the CPU, GPU, and Neural Engine to share the same memory, thereby reducing latency. Apple has optimized this piece of hardware for macOS, which delivers a seamless end-user experience.

The M1 chip has significantly impacted the laptop market, and boosted Apple's market share with its performance and battery life. This success strengthened Apple's hardware and software ecosystem integration, pushing competitors to innovate and potentially lower prices. The M1's ARM architecture indicates a shift from x86 dominance, diversifying the tech landscape. Since then, Apple has evolved from M1 to M4 family of processors, with new innovations at each generation.

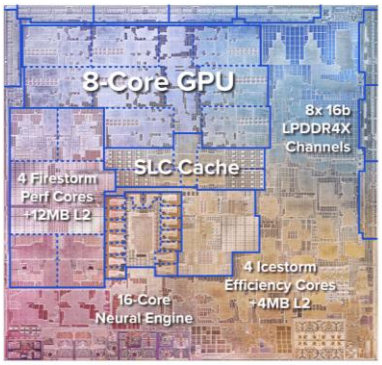
8-Core GPU
8x 16b LPDDR4X Channels
SLC Cache
4 Firestorm Perf Cores +12MB L2
4 Icestorm Efficiency Cores +4MB L2
16-Core Neural Engine

Figure 4: Architecture of Apple M1 chip

7.3 Qualcomm’s Custom Chips

Qualcomm Snapdragon chips integrate advanced features like 5G connectivity, AI processing, and extended battery life onto a single piece of hardware. The Snapdragon X Elite, featuring custom Nuvia cores, combines an 8-core CPU with high-performance and efficiency cores, an Adreno GPU, and a Hexagon DSP for advanced AI tasks. Compared to top Intel offerings, it delivers 51% faster performance and extended battery life, making it a strong contender in the high-performance laptop market.

The advanced Hexagon AI processor enhances on-device machine learning, particularly generative AI, and is first in the segment to support this. This also provides a larger platform for the Windows-on-ARM ecosystem. This chip's capabilities position Qualcomm as a significant player in the laptop market, driving competition, innovation, and ARM-based application adoption while reducing reliance on Intel and AMD.

7.4 Google Axion Chip

The Google Axion chip is set to revolutionize Google's extensive array of cloud services like BigTable, Spanner, and BigQuery, boosting performance and operational efficiency. The chip’s design is built on the ARM v9 architecture, featuring the ARM Neoverse V2 CPU, ensuring seamless compatibility with existing applications. It is tailored for high-performance tasks such as web servers, databases, data analytics, and AI, providing up to 30% better performance compared to the fastest ARM-based instances. It promises up to 50% better performance and up to 60% better energy efficiency than comparable current-generation x86-based instances. Axion’s enhanced efficiency, with titanium offloads improving networking, security, and storage operations, frees up capacity for customer tasks, leading to notable gains in both performance and energy efficiency.

7.5 Microsoft Custom Silicon

Microsoft's Azure Cobalt CPU is a 128-core chip built on an ARM Neoverse CSS design, tailored specifically for Microsoft. It is engineered to power general cloud services on Azure, with a keen focus on performance and power management. According to Microsoft, the design of the Cobalt CPU includes the ability to control performance and power consumption per core and on every single virtual machine, ensuring both high performance and efficient power usage. This level of control allows for optimized operation across diverse workloads, enhancing the flexibility and efficiency of Azure’s cloud services.

Currently, Microsoft is testing the Cobalt CPU on various workloads, including Microsoft Teams and SQL Server, and plans to offer virtual machines with this CPU to customers next year on its Azure cloud. It is preliminarily tested to be up to 40% better performance compared to the ARM-based servers currently in use within Microsoft's data centers. While detailed system specifications and benchmarks are not yet available, the early performance metrics suggest significant improvements.

8 Current Status of x86 Architecture

8.1 Intel’s Advancements in AI

Intel has made significant strides in AI processing with the introduction of CPU-based training cores, such as the Intel Habana Gaudi cluster, which are comparable to GPUs in performance. The Intel Gaudi AI accelerator delivers up to 40% better price/performance for training compared to Nvidia GPU-based instances (Figure 5). Additionally, the efficient architecture of the Intel Gaudi AI accelerator allows Supermicro to offer a significant price-performance advantage over GPU-based servers with the Supermicro X12 Server featuring Intel Gaudi AI accelerators.

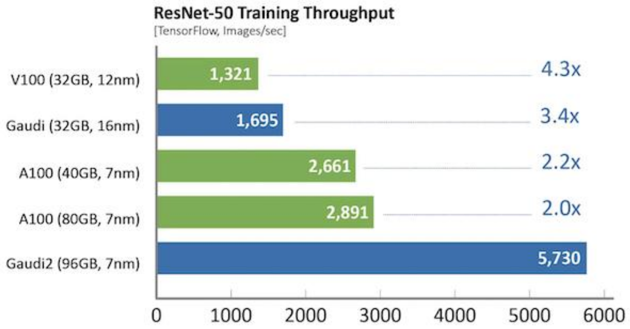

Figure 5: Training Throughput Comparison

Furthermore, Intel's upcoming Core Ultra 200V “Lunar Lake” SoC is set to offer more than 100 TOPS (tera-operations per second) for AI tasks, positioning it at the forefront of AI processing technology. This SoC will belong to the low-power segment and is best suited for applications where energy efficiency is prioritized. It will utilize the Lion Cove and Skymont architectures, with a low-power variant of the Xe2-LPG architecture.

8.2 AMD’s Promising Hardware

AMD has also focused on enhancing AI processing and high-performance computing with the launch of the Ryzen AI Engine. It is the first and only dedicated AI engine on an x86 Windows processor, specifically designed for processing AI inference models. The Ryzen AI Engine brings unprecedented efficiencies to personal computing, facilitating tasks such as work, collaboration, and innovation in a more seamless and connected manner. By offloading AI processing tasks from the main CPU, it enables more efficient and responsive computing experiences, particularly in applications that rely heavily on AI, such as real-time data analysis, generative AI, and advanced multimedia processing.

In the data center, AMD EPYC processors provide robust support for advanced AI engines, offering exceptional performance, scalability, and energy efficiency. Servers built with 4th Gen AMD EPYC processors deliver up to 2x the cores for heterogeneous and GPU processing compared to 5th Gen Intel Xeon processors. They also offer up to 29% more all-core turbo frequency per core, up to 50% more memory capacity, and up to 29% more memory bandwidth for heterogeneous GPU workloads.

9 Future Predictions

The market for ASIC chips dedicated to AI is projected to experience significant growth, with Morgan Stanley estimating an 85% annual increase between 2023 and 2027, reaching a market value of $30 billion. The increasing interest in custom silicon and AI hardware is reshaping the competitive landscape. Software-centric companies like OpenAI are exploring the development of their own AI processors, such as Neural Processing Units (NPUs), to reduce reliance on Nvidia's hardware.

While Custom Silicon is gaining traction, x86 architecture is expected to remain relevant in the market still. The primary reason is the compatibility of legacy software and the vast development ecosystem would ensure its sustained usage for some time. These still hold an advantage in certain security and high-performance compute applications due to their mature and established toolchains. The major application where x86 still remains unbeaten is the gaming market, in which ARM has not been able to tap into.

Despite the high costs associated with standalone chip development, industry experts believe that it is not prohibitively high. In 2022 alone, thousands of chip designs below 28 nm were completed, with nearly 100 designs at the most advanced nodes by a lot of independent non-conventional design houses. The continuous innovations in Electronic Design Automation (EDA) tools and the ecosystem business model have kept design costs manageable. AI-assisted workflows from companies like Synopsys and Cadence are crucial in this regard, lowering costs and speeding up time-to-market.

10 Conclusion

The ongoing evolution of computer architecture is driven by increasing demands of performance, efficiency, and customization. A fundamental shift in computing is in place, driven by the rise of custom silicon solutions tailored to meet the demands of modern workloads. The future promises further specialization and integration, fueled by the continued evolution of AI and ML workloads. As advanced hardware develops, custom silicon will increasingly deliver optimized performance and efficiency across a wide range of applications. This "Golden Age" is not just about faster chips and more powerful systems; it's about unlocking new possibilities and accelerating innovation across various domains. From scientific research and medical breakthroughs to personalized user experiences and sustainable computing, custom silicon is sure to reshape the technological landscape for generations to come.

REFERENCES

[1] N. Zhang, “Moore’s Law is dead, long live Moore’s Law!”. Accessed: May 25, 2024. [Online]. Available: https://arxiv.org/abs/2205.15011

[2] Y. Sun, N. B. Agostini, S. Dong, and D. Kaeli, “Summarizing CPU and GPU Design Trends with Product Data.” arXiv, Jul. 13, 2020. Accessed: May 25, 2024. [Online]. Available: http://arxiv.org/abs/1911.11313

[3] Q. Jiang, Y. C. Lee, and A. Y. Zomaya, “The Power of ARM64 in Public Clouds,” in *2020 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID)*, Melbourne, Australia: IEEE, May 2020, pp. 459–468. doi: 10.1109/CCGrid49817.2020.00-47.

[4] E. S. Chung, P. A. Milder, J. C. Hoe, and K. Mai, “Single-Chip Heterogeneous Computing: Does the Future Include Custom Logic, FPGAs, and GPGPUs?,” in *2010 43rd Annual IEEE/ACM International Symposium on Microarchitecture*, Atlanta, GA, USA: IEEE, Dec. 2010, pp. 225–236. doi: 10.1109/MICRO.2010.36.

[5] A. Firoozshahian *et al.*, “MTIA: First Generation Silicon Targeting Meta’s Recommendation Systems,” in *Proceedings of the 50th Annual International Symposium on Computer Architecture*, Orlando FL USA: ACM, Jun. 2023, pp. 1–13. doi: 10.1145/3579371.3589348.

[6] B. Nikolic, E. Alon, and K. Asanovic, “Generating the Next Wave of Custom Silicon,” in *ESSCIRC 2018 - IEEE 44th European Solid State Circuits Conference (ESSCIRC)*, Dresden: IEEE, Sep. 2018, pp. 6–11. doi: 10.1109/ESSCIRC.2018.8494310.

[7] D. Reed, D. Gannon, and J. Dongarra, “Reinventing High Performance Computing: Challenges and Opportunities.” arXiv, Mar. 04, 2022. Accessed: May 12, 2024. [Online]. Available: http://arxiv.org/abs/2203.02544

[8] C. Singasani, “From Concept to Completion: The Power of Custom SoCs and End-to-End Synopsys Solutions,” Synopsys.

[9] M. Riga, “An Era of Affordability for the Custom System on-Chip (SoC)”, ARM.

[10] K. Bulusu, “Custom Silicon: The Key to Innovation and Competition in the Future of Computing.” 2023. https://medium.com/@kbulusu/custom-silicon-the-key-to-innovation-and-competition-in-the-future-of-computing-49c212c0fb7

[11] G. Ciccomascolo, “Future of AI Is Also A Chip Size Matter: What’s ASIC And Why It’s Relevant Now.” 2024. https://www.ccn.com/analysis/future-ai-chip-size-matter-asic/

[12] K. Morales, “Classic Moore’s Law Scaling Challenges Demand New Ways to Wire and Integrate Chips.” 2022. https://www.appliedmaterials.com/us/en/blog/blog-posts/classic-moores-law-scaling-challenges-demand-new-ways-wire-and-integrate-chips.html

[13] M. Awad, “Harnessing the Power of the Ecosystem in the Era of Custom Silicon on Arm.” 2023. https://newsroom.arm.com/news/arm-total-design-ecosystem

[14] M. Awad, “How Arm Total Design is built around 5 key building blocks.” 2023. https://www.edn.com/how-arm-total-design-is-built-around-5-key-building-blocks/

[15] Apple, “Apple announces Mac transition to Apple silicon.” 2020. https://www.apple.com/newsroom/2020/06/apple-announces-mac-transition-to-apple-silicon/

[16] J. Salter, “Hands-on with the Apple M1—a seriously fast x86 competitor.” 2020. https://arstechnica.com/gadgets/2020/11/hands-on-with-the-apple-m1-a-seriously-fast-x86-competitor/

[17] AWS, “AWS Inferentia.” https://aws.amazon.com/machine-learning/inferentia/

[18] AWS, “AWS Trainium.” https://aws.amazon.com/machine-learning/trainium/

[19] K. Sato and C. Young, “An in-depth look at Google’s first Tensor Processing Unit (TPU).” 2017. https://cloud.google.com/blog/products/ai-machine-learning/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu

[20] A. Vahdat, “The past, present and future of custom compute at Google.” 2021. https://cloud.google.com/blog/topics/systems/the-past-present-and-future-of-custom-compute-at-google

[21] A. Vahdat, “Introducing Google Axion Processors, our new Arm-based CPUs.” 2024. https://cloud.google.com/blog/products/compute/introducing-googles-new-arm-based-cpu

[22] P. McLellan, “Domain-Specific Computing 2: The End of the Dark Ages.” 2019. https://community.cadence.com/cadence_blogs_8/b/breakfast-bytes/posts/dsc2

[23] Qualcomm, “Snapdragon X Elite Product Brief.” 2024.

[24] Meta, “Our next-generation Meta Training and Inference Accelerator.” 2024. https://ai.meta.com/blog/next-generation-meta-training-inference-accelerator-AI-MTIA/

[25] C. Trueman, “Meta’s upgraded MTIA AI chips offer 3.5x performance boost.” 2024. https://www.datacenterdynamics.com/en/news/metas-upgraded-mtia-chips-offer-35x-performance-boost/

[26] Habana AI, “Intel Gaudi Accelerator.” https://habana.ai/products/gaudi/

[27] P. Alcorn, “Intel says Lunar Lake will have 100+ TOPS of AI performance — 45 TOPS from the NPU alone meets requirement for next-gen AI PCs.” https://www.tomshardware.com/pc-components/cpus/intel-says-lunar-lake-will-have-100-tops-of-ai-performance-45-tops-from-the-npu-alone-meeting-requirement-for-next-gen-ai-pcs

[28] AMD, “The Future of AI PCs Gets Even Better with AMD”. https://www.amd.com/en/products/processors/consumer/ryzen-ai.html

[29] Y. Boon, “One Size Doesn’t Fit AI”. https://www.nb.com/en/gb/insights/insights-one-size-doesnt-fit-ai