One-for-All Chip: Same, Similar, Different—What You Want, When You Want It, All in One Chip

Nitesh Narayana GS
UPC Barcelona
nitesh@ac.upc.edu

Xavier Martorell
UPC Barcelona
xavi.martorell@upc.edu

1 Introduction

Computer chips have come a long way, from early designs like the Manchester "Baby" [23] to innovations like the Intel 4004 [16] and modern processors such as Intel Ultra [26]. These advancements have enabled humanity's journey from the Moon landing [1] to the exploration of Mars! At the core of these breakthroughs lies continuous progress in integrated circuit design.

However, with Moore's Law[27] nearing its limits, traditional scaling is slowing, and architectural advancements have become the primary differentiator across chip generations. Despite incremental improvements, billions of 'new gen' chips are still produced each year. This aggressive manufacturing, without fully accounting for material and energy impacts, risks leading us toward a WALL-E[25]-like future rather than opening new Interstellar[17][1] frontiers.

To ensure a sustainable and efficient computing future, chip design must become more intelligent and modular. Innovations, whether incremental or revolutionary, should be adaptable to chips as their physical area reaches limits. To this end, we propose the **One-for-All Chip**[2], a highly modular, reconfigurable architecture focused on sustainability and energy efficiency—one that we believe will define the future of chip design.

2 Computer Chips in All Different Angles

Computer chips come in a variety of types, use cases, shapes, sizes, and more. In this section, we explore how they can be similar, the same, and yet different in many ways.

Different Use cases (Shapes): Computer chips come in various shapes tailored to different use cases. For general-purpose computing, we have CPUs. For graphics rendering, high-performance computing applications, and AI tasks, Graphics Processing Units (GPUs) are utilized. For data processing, Data Processing Units (DPUs) are commonly used, and for specialized AI processing, we have Neural Processing Units (NPUs). Each of these chip types serves distinct roles based on their specific requirements.

Chips - Similar *but* Different: Within each category of chips, there are various models that, while similar, differ significantly in key aspects. For example, x86 CPU chips from Intel[7], ARM[6] CPU chips from Apple[4], Ampere[3], Nvidia[8], and Qualcomm[11], and RISC-V[12] CPU chips from Ventana[14] and Shakti[13] all share the common goal of general-purpose computing, yet differ fundamentally in their Instruction Set Architectures (ISAs) and the strategic objectives of their respective vendors. Similarly, GPUs from Nvidia (with CUDA[9]), Apple (with MSL[5]), and Intel (with oneAPI[10]) all aim to serve the same purpose of graphics, AI, and high-performance computing, but each employs unique approaches and technologies. These examples highlight the redundancy of chip types within each category, yet the fundamental differences that drive innovation within them.

Table 1. Comparison of CPU Microarchitecture generations

Component	Gen-X	Gen-Y	Gen-Z
ROB Size	224	224	352
BTB Size	1.5K	1.5K	2.25K
Uop Cache	1.5K entries	2.25K entries	2.25K entries
L1 Cache	32KB/core	32KB/core	32KB/core
L2 Cache	256KB/core	1MB/core	1.25MB/core
Node	14nm	14nm	10nm

Table 2. Comparison of Core Elements Across Chips

Core Element	CPU	GPU	NPU	DPU
ALU	✓	✓	✓	✓
Control Unit	✓	✓	✓	✓
Registers	✓	✓	✓	✓
Pipeline Stages	✓	✓	✓	✓
Interconnect Units	✓	✓	✓	✓
Caches (L1, L2)	✓	✓	✓	✓
Memory Controllers	✓	✓	✓	✓
Fetch & Decode Units	✓	✓	✓	✓
Vector Processors	✓	✓	✓	✗
FPU	✓	✓	✓	✗
Branch Predictor	✓	✗	✗	✗

Chips - Same *and* Different: Even within the same category, chips can share similarities yet exhibit significant differences. For example, consider different generations of processors from a top CPU vendor, as shown in Table 1. Upon closer inspection, it becomes evident that while the 'Gens's are based on similar core designs,'Gen-Y' integrates additional resources from 'Gen-X'. Similarly, 'Gen-Y', with even more resources, manifests as 'Gen-Z'. This indicates that at the core, different generations are still the same.

Chips - Different *but* Similar: At a high level, CPUs, GPUs, NPUs, and DPUs may appear to be vastly different. However, a deeper dive into the underlying architecture, as illustrated by Table 2, reveals that these chips share many fundamental components.

[1] at least the world towards the end of the movie!
[2] Inspired by One For All [19], a modular superpower from My Hero Academia [18], capable of accumulating and utilizing multiple abilities at will.

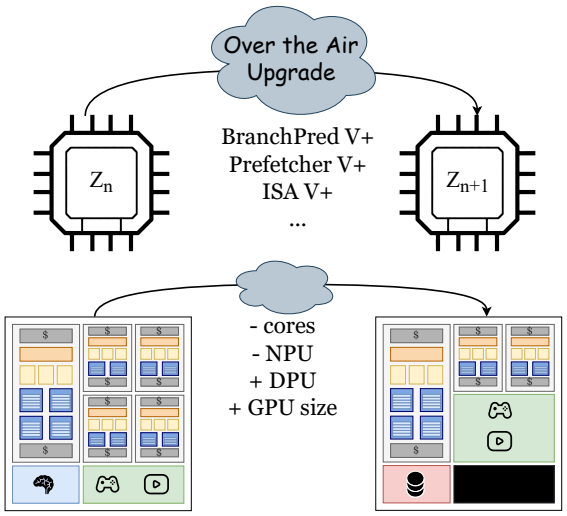

Figure 1. One-for-All Chip Concept.

Consequence of Differences: As we've seen, there are cases where fundamental components are similar across different chips, and other instances where chips of the same category differ significantly in their architecture. Furthermore, varying use cases necessitate the creation of entirely new chips, which adds significant costs for even small differences. For example, the price difference between Intel's i5, i7, and i9 processors within the same generation—$296, $436, and $562 respectively [2]—illustrates this disparity. Additionally, add-on features, such as integrated graphics, can further increase the cost for the same processor, with prices rising from $532 (without integrated graphics) to $562 (with integrated graphics). Such cost disparities are also evident over generations in GPUs and other chips. While the end-product cost difference might seem small, the overall cost to design, manufacture, and deliver new chips—from concept to final product—can be massive, even for minor architectural improvements. This makes the current process of chip development increasingly unsustainable.

When similar chips stop being different: The primary factor that differentiates 'similar' chips across generations has traditionally been the process node. However, as noted by Moore's Law extrapolation in NN GS et al. [28], the scaling of processor nodes is expected to slow down and hit a limit due to the physical limits of materials (e.g., atomic size [21, 22] and quantum tunneling[20]). As we approach these physical limits, future chip generations will likely differ primarily in terms of architectural innovations rather than process node advancements. With chip area potentially reaching a plateau beyond Moore's Law, one must question whether it still makes sense to have a new chip for each generation. Moreover, could this shift prompt us to reconsider the need for entirely different 'chips' for different usecases?

3 One-for-All Chip

The Great Power: Imagine a single, highly reconfigurable chip that acts as a "black box," reconfigurable to your exact needs. With the One-for-All Chip, you can select an architecture—Intel or Apple—just like choosing an operating system, and configure it to your preferences. You can add or remove features at will, such as enhancing the branch predictor or adjusting the ISA features based on the latest architecture upgrades delivered "over-the-air", allowing seamless evolution without hardware replacements (Fig. 1). The chip also enables dynamic core adjustments: need more CPU, GPU, or NPU cores? Simply add or remove them. Switch between architectures to optimize for power efficiency or performance, adapting to your changing demands. You can tailor the chip's behavior for specific needs by selecting the right configuration, as illustrated in Fig. 1. For parts of the chip you don't need, you can completely switch them off at the element level (darkened part in the Fig.1), similar to the AMOLED's pixel-off concept[15], ensuring maximum energy efficiency.

The Rewards: The design process becomes more sustainable with lower costs and faster upgrades to advanced architectures, allowing both users and companies to access rapid, low-cost improvements without the need for full hardware replacements. Energy efficiency is easily achieved by selectively switching off chip elements. The industry moves towards a unified approach to architecture and software, enhancing compatibility and development efficiency. Reconfigurability and modularity take the center stage, allowing chips to adapt more effectively to diverse needs.

The Greater Responsibility: For unification and modularization to be successful, companies across both hardware and software domains must collaborate. This includes standardizing elements like the ISA, compilers, interconnects, chip design, and memory systems, among others. Reconfigurable chip design research should advance to the point where it can match the performance and capabilities of current traditional chips, even at the device level. Additionally, the software and workflows, similar to tools like Vivado[24], that support reconfigurability must be fast, lightweight, and efficient. Furthermore, "over-the-air" upgrades for chips would require a new approach to testing and validation, akin to the rigorous processes used for software releases.

4 Conclusion

We envision that the future of chip design requires a fundamental shift in methodology, with a strong emphasis on sustainability and energy efficiency. In this context, we believe that focusing on the development of the **One-for-All** chip will not only revolutionize chip design but also drive meaningful innovations in both computer architecture and software. As demonstrated by the **One-for-All** chip, we are approaching a future where modular and reconfigurable chips and devices are within reach.

Acknowledgments

We also extend our gratitude to the ASPLOS Student Travel Grant Committee, as well as Anupama G and Sathyanarayana GK, for making it possible for us to present our work in person. Most importantly, we are deeply grateful to S. Rajendra for his critical advice, without which this idea would not have come to fruition.

References

[1] 1966 (accessed April 18, 2024). Apollo Guidance Computer. https://en.wikipedia.org/wiki/Apollo_Guidance_Computer.
[2] 1968 (accessed April 18, 2024). Intel. https://www.tomshardware.com/news/full-10th-gen-comet-lake-cpu-tray-pricing-listed.
[3] (accessed April 18, 2024). Ampere Computing. https://amperecomputing.com/.
[4] (accessed April 18, 2024). Apple. https://www.apple.com/.
[5] (accessed April 18, 2024). Apples Metal Shading Language. https://developer.apple.com/metal/Metal-Shading-Language-Specification.pdf.
[6] (accessed April 18, 2024). ARM. https://www.arm.com/.
[7] (accessed April 18, 2024). Intel. https://en.wikipedia.org/wiki/Intel.
[8] (accessed April 18, 2024). Nvidia. https://www.nvidia.com/.
[9] (accessed April 18, 2024). Nvidia CUDA. https://en.wikipedia.org/wiki/CUDA.
[10] (accessed April 18, 2024). Optimize Your GPU Application with the Intel® oneAPI Base Toolkit. https://www.intel.com/content/www/us/en/developer/tools/oneapi/training/gpu-optimization-workflow.html#gs.is2yam.
[11] (accessed April 18, 2024). Qualcomm. https://www.qualcomm.com/.
[12] (accessed April 18, 2024). RISC-V. https://riscv.org/.
[13] (accessed April 18, 2024). Shakti, Open Source Processor Development Ecosystem. https://shakti.org.in/.
[14] (accessed April 18, 2024). Ventana Micro. https://www.ventanamicro.com/.
[15] (accessed Feb 18, 2025). AMOLED. https://en.wikipedia.org/wiki/AMOLED.
[16] (accessed Feb 18, 2025). Intel 4004. https://myheroacademia.fandom.com/wiki/One_For_Allhttps://myheroacademia.fandom.com/wiki/One_For_All.
[17] (accessed Feb 18, 2025). Interstellar. https://www.warnerbros.co.uk/movies/interstellar.
[18] (accessed Feb 18, 2025). My Hero Academia. https://heroaca.com/.
[19] (accessed Feb 18, 2025). One For All- My hero academia. https://myheroacademia.fandom.com/wiki/One_For_All.
[20] (accessed Feb 18, 2025). Quantum Tunnelling. https://en.wikipedia.org/wiki/Quantum_tunnelling.
[21] (accessed Feb 18, 2025). Single Atom Transistor. https://en.wikipedia.org/wiki/Single-atom_transistor.
[22] (accessed Feb 18, 2025). These Transistor Gates Are Just One Carbon Atom Thick Researchers may have hit a hard limit. https://spectrum.ieee.org/smallest-transistor-one-carbon-atom.
[23] (accessed Feb 18, 2025). Timeline of Computer History. https://www.computerhistory.org/timeline/computers/.
[24] (accessed Feb 18, 2025). Vivado Design Suite. https://www.amd.com/en/products/software/adaptive-socs-and-fpgas/vivado.html.
[25] (accessed Feb 18, 2025). WALL-E. https://www.pixar.com/wall-e.
[26] (accessed March 23, 2025). Intel Ultra Architecture. https://download.intel.com/newsroom/2024/client-computing/Lunar-Lake-Architecture-Fact-Sheet.pdf.
[27] Gordon E Moore. 1965. Cramming More Components onto Integrated Circuits. *Electronics* 38, 8 (1965), 1–4.
[28] **Nitesh Narayana GS** and Abhijit Das. April 27- May 1, 2024. Apparate: Evading Memory Hierarchy with GodSpeed Wireless-on-Chip. In *Wild and Crazy Ideas (WACI) Session in ASPLOS 2024.*