Advances in Engineering Innovation (2024) Volume 10 EWA Publishing
Published online: 31 August 2024 DOI: 10.54254/2977-3903/10/2024109

LLMs at home? An evaluation on the feasibility of popularising On-device-ANI capable hardware in consumer grade devices

Yiding Wang

Wesley College Melbourne

Felix.Wang@wesleycollege.edu.au

Abstract. Artificial Narrow Intelligences (ANI) are rapidly becoming an integral part of everyday consumer technology. With products like ChatGPT, Midjourney, and Stable Diffusion gaining widespread popularity, the demand for local hosting of neural networks has significantly increased. However, the typical 'always-online' nature of these services presents several limitations, including dependence on reliable internet connections, privacy concerns, and ongoing operational costs. This essay will explore potential hardware solutions to popularize on-device inferencing of ANI on consumer hardware and speculate on the future of the industry.

Keywords: ANI, neural network, LLM, AI Accelerator, chip design, component efficiency, consumer applications

1. Introduction

Recently, Artificial Narrow Intelligences (ANI) have become increasingly and pervasively integrated into the lives of common consumers [1]. Transformer-based Large Language Models (LLM) and text-to-image models, such as ChatGPT, Midjourney, Stable Diffusion, and Character.ai, are witnessing widespread domestic applications. These products, which utilize Transformer and/or Diffusion-based models, have gained significant market popularity. For instance, OpenAI's ChatGPT boasts over 200 million monthly active users, according to data analysis from Backlinko [2]. However, a significant hurdle remains the typical 'always-online' nature of these services. Instead of running locally on users' devices, they require access to remote cloud servers to provide the computing power necessary for operating these neural network systems [3].

While hosting on cloud servers ensures broad accessibility, several limitations persist. These include dependence on a reliable internet connection, privacy concerns related to consumer data gathering, and ongoing operational costs due to bandwidth usage [4][5]. These aspects severely limit the utilization of ANI in broader fields, such as digital interactive mediums, especially works produced by smaller developers. The API fees per token and the always-online requirement create hurdles for consumers in these mediums that are otherwise absent in a one-time-purchase model or freeware-predominant environment. Furthermore, the need to connect to external servers adds another layer of privacy concern for users. These issues could be eliminated if consumers were able to run neural network models on their devices. However, current attempts at on-device artificial intelligence primarily involve simplifying or downscaling the models to demand less computing power, at the cost of limited capabilities. The software approach is favored as it is more economical and easier to popularize. However, software optimization alone is insufficient to enable the on-device operation of larger-scale neural networks, such as OpenAI's GPT-4, which is rumored to contain a total of 1.8 trillion parameters [6].

As such, this essay aims to explore the hardware aspect of on-device neural networks, analyzing various attempts and methods to popularize consumer-grade, on-device ANI-capable hardware capable of hosting more complex neural networks. This essay will analyze and evaluate the feasibility, advantages, and disadvantages of these approaches and proposed solutions from perspectives including but not limited to practicality, performance, scalability, and accessibility. While the potential for capital gain will be considered, the primary focus will be on the technical aspects of the products analyzed rather than the business aspect. The essay will concentrate on hardware optimization, acknowledging that software optimization is equally, if not more, important in popularizing on-device ANI compared to providing consumers with more powerful hardware.

The essay will focus on the consumer on-device hosting of pre-trained neural networks and occasional fine-tuning, excluding on-device training, which is much more computationally intensive, infeasible, and unnecessary for regular consumers. The analysis

will primarily draw on theoretical frameworks of hardware design and secondary research for data gathering, as practical experiments are not feasible at this stage.

2. GPU-based solution

One potential solution for designing ANI-capable consumer devices is by utilizing and improving an already-existing component on most personal computers – the Graphic Processing Unit (GPU). GPU and Central Processing Unit (CPU) are the two processing units most commonly found on modern consumer personal computers. The CPU is designed for more general-purpose computing, while the GPU's main functionality lies in graphic processing and rendering. Recently, however, the GPU has found many applications in the field of ANI training and hosting.

To efficiently process visual information, which typically consists of repetitive computation of simple data values in large quantities, most GPUs are made able to perform concurrent parallel processing and possess many more cores than an equivalent-level CPU. Each core is less powerful computationally than a CPU core but capable of executing operations concurrently [7][8]. In addition, to accommodate the large amounts of data GPUs have to process, they are usually designed with higher memory bandwidth, so that a larger quantity of data can be moved from VRAM (Video Random-Access Memory) to the cores. These qualities happen to make the GPU uniquely suited for accelerating neural networks, as much like video processing, running a neural network relies on a large number of matrix operations, which themselves are composed of small arithmetic operations. As such, the quality of a GPU finds much application in operating a neural net. For instance, compared to a CPU's sequential computation, a GPU's capability of concurrent parallel computation and large quantity of cores enables it to compute these matrix operations much more efficiently compared to most other types of processing units. Likewise, the higher memory bandwidth of the GPU ensures that more data can be transferred from system memory to the GPU's local VRAM with each fetch operation, resulting in fewer fetch operations on average, especially when handling large datasets such as matrices. This reduces the bottleneck period where the processing cores idle, waiting for data to be transferred. For these reasons, the GPU became a valuable resource for both neural network training and inferencing.

Another major factor contributing to the GPU's appeal is its already-established wide availability. GPUs had prolonged applications in media creation and video gaming before the full proliferation of fields such as machine learning [9], which in turn means that the production of GPUs is fully mature, and GPUs were widely accessible when the demand for computing power in the training and hosting of neural networks grew. This factor, combined with the GPU's superior efficiency at matrix calculations compared to other widely available options (CPUs), made the GPU an appealing choice for neural net operations. Compared to more specialized alternatives, such as AI accelerators or Neural Processing Units (NPUs), which are more efficient at parallel processing than GPUs, GPUs are still favored due to their established availability and matured production. This accessibility extends to consumers as well, providing them with a greater assortment of choices compared to more niche and specialized hardware such as NPUs. In the context of desktop applications, most modern desktop computers have GPUs as a core component, and users can easily slot the GPU into their motherboard without making special accommodations, providing ease of usage on top of ease of access. The widespread availability of GPUs means that consumers do not need to invest in highly specialized or expensive hardware to benefit from the capabilities of ANI. This broadens the potential user base and democratizes access to advanced computational resources, enabling everyday users to explore and leverage the power of neural networks without facing prohibitive costs or availability issues from more state-of-the-art components.

Additionally, many consumers view the generalist quality of GPUs as a significant advantage. Unlike specialist components, a GPU offers a broader range of utilities and applications. Industry professionals often seek out GPUs for their adequate parallel processing abilities in neural network development. However, consumers also value GPUs for their performance in video gaming, media consumption, and content creation. The capability to handle neural networking tasks is seen as an added bonus. As a result, consumers may generally be less enthusiastic about more specialized components like NPUs, which, despite their higher efficiency in parallel processing, lack the versatility of GPUs, being only able to fulfill one singular task. This preference for generalist hardware like GPUs ensures that they remain a popular choice for a wide range of users. The broader utility and established presence of GPUs in various applications make them a more attractive and practical option for most consumers.

However, while generalization may be appealing, it also limits the effectiveness of any GPU-based solution for popularizing local-hosting capable devices. To fulfill its intended purpose of graphic processing, a GPU cannot be solely optimized for neural network tasks, making it somewhat 'bloated' from the perspective of neural network hosting with features not needed for such purposes. This lack of specialization means that GPUs are generally less power-efficient compared to specialized components like NPUs. Furthermore, this inefficiency poses significant challenges for applications in the Internet of Things (IoT) and smartphones, where power consumption, battery life, and space are critical factors. The higher power demands of GPUs make them less suitable for these environments, and their performance must be significantly limited or capped for them to be applicable. Additionally, the multifunctional nature of GPUs can lead to resource competition within a system. When a GPU is tasked with both neural network processing and visual processing tasks, these activities can compete for the same resources, potentially causing bottlenecks and reducing overall system performance. This competition for resources can be particularly problematic in scenarios where high efficiency and quick processing times are essential, or when the two tasks need to work in conjunction, perhaps in multimedia work or video games that rely on real-time generative AI and have a high demand for graphic processing. While the generalist

nature of GPUs provides broad utility and appeal, it also introduces limitations in terms of power efficiency and resource allocation, making them less ideal for certain applications.

Currently, according to Steam's monthly user hardware survey, around 28.14% of participating Steam users possess a GPU of higher or equivalent quality to Nvidia RTX 3050, and roughly 23% of users have more than 12 Gigabytes of VRAM [10]. While consumers' computing power is still growing, with a higher and higher percentage of users using better GPUs as the chart would indicate, there is still a substantial gap to cross. It is also important to remember that Steam is a platform that is video game-centric, with an audience that naturally possesses a higher demand for better GPUs, and the platform focuses on the desktop environment. Across the broader market, many consumers use devices that do not match the high-performance standards seen among Steam users, nor do they necessarily have the incentive to upgrade their devices. In general computing, especially in work computers, mobile devices, and IoT applications, the demand for high-end GPUs is significantly lower, as the demand for high-fidelity graphics is lower. While there is clear progress in the adoption of more powerful GPUs among certain user groups, widespread application, particularly in power-sensitive and mobile environments, remains constrained. GPU solutions have a solid foundation when applied to desktop applications but may struggle to find solid footing in applying a universal solution to make consumer devices capable across different user groups and areas.

3. Specialised AI Accelerator

The alternative solution to using pre-existing hardware components is, of course, developing specialized new components. The development of AI Accelerators, also known as Neural Processing Units (NPUs), seeks to provide a specific and efficient solution for neural networks, both in industrial applications and consumer products. Compared to a GPU, NPUs are "tailored to execute specific operations required for neural networks" [11], allowing for specialization in processing machine learning algorithms. This is reflected in their architecture, which features smaller but more numerous cores, enabling faster parallel processing. Additionally, by offloading the task of operating a neural network from the CPU or GPU, an NPU can significantly improve the system's overall efficiency and performance. Numerous studies and ideas have focused on consumer applications of NPUs, many of which concentrate on on-device neural network systems [12][13].

The most straightforward advantage of deploying an NPU is its efficiency. With a higher number of cores and larger memory bandwidth, NPUs can outperform all other processing units [11]. Furthermore, NPUs achieve much higher energy efficiency [14], making them a more viable option for applications in mobile and smaller devices. However, because they are tailored for neural network inference, their performance in other areas is lackluster. NPUs must work alongside CPUs and GPUs to create a functional computing environment. From the perspective of many consumers, NPUs may be viewed negatively as an additional component to purchase, which could lead to fewer purchasing incentives. This downside can be avoided if NPUs are integrated into devices, such as mobile phones or IoT applications, or if NPUs are integrated into newer generation CPUs—an area Intel is interested in pursuing [7] and AMD has already explored. In this manner, consumers could benefit from the specialized processing power of NPUs without needing to purchase and install additional hardware, which would be more appealing to many consumers.

Another issue with NPUs is the cost of such components. As a proprietary piece of hardware, NPUs often come with higher associated research and development costs, which are reflected in their price tags. This increased cost can be a significant barrier to widespread consumer adoption, particularly in markets where price sensitivity is high. For individual consumers, the additional expense of purchasing an NPU might not be justifiable, especially if their primary computing needs do not heavily rely on neural network processing. In the context of mobile and IoT devices, integrating NPUs can also drive up the overall cost of the product. As such, manufacturers may struggle to balance the benefits of enhanced AI capabilities with the need to keep devices affordable to avoid dampening demand. However, the advantage of NPUs shines in their scalability, and NPUs are usually cheaper to manufacture than equivalent-level GPUs, due to being purpose-built and having less complex components [15]. As the initial development phase of NPUs passes and the product matures, they can become quite competitive and pose a potential solution to popularizing on-device ANI. Overall, while the cost of NPUs currently poses a challenge to their widespread adoption in consumer applications, ongoing advancements in technology and manufacturing processes hold the potential to make these powerful components more economically viable in the future.

There have been many discussions on the role NPUs will play in future consumer hardware, including how they may be integrated into the current hardware paradigm. Intel has predicted a "NPU/GPU combination" becoming popularized in the near future [7], and one study has explored a NPU-CPU combination system on a chip for on-device AI applications [14]. Systems on a chip (SoCs) present a unique approach to the issue. They typically describe a type of system design that integrates all integral parts of a computer (or other device) into a single chip. This type of design has several benefits, the most obvious being the reduction in size, allowing it to be employed in smaller mobile devices, such as mobile phones or handheld devices. Performance will also see improvement, as the close proximity of components on the chip can reduce latency and increase data transfer and communication speeds between them, without the need for relaying data through a motherboard. This is beneficial for on-device ANI, as memory bandwidth is often the bottleneck in ANI inferencing. By physically moving the components together, the increased memory transfer speed between system memory and the local memory of the NPU or GPU will largely alleviate this bottleneck. Apple has pioneered SoC design with their Apple M-series chips and is now developing their own on-device LLM [16].

However, despite these benefits, several challenges are associated with SoC integration. One major challenge is the limited flexibility of SoCs. Once integrated, the components cannot be easily upgraded or replaced, which can be a disadvantage in rapidly evolving fields like AI and machine learning. Consumers and developers are accustomed to the modularity of traditional desktop environments, where individual components such as NPUs can be upgraded independently. The fixed nature of SoCs means that users may need to replace the entire chip or device to benefit from newer technologies or improvements, potentially leading to higher long-term costs and electronic waste, which is already quite an issue in the electronics industry. Furthermore, thermal management is another critical issue. Integrating multiple high-performance components into a single chip generates significant heat, which can be challenging to dissipate effectively. Efficient cooling solutions are necessary to prevent overheating and ensure stable operation, especially in compact mobile devices where space for thermal management is limited. All in all, using SoC integrated with NPU as a solution to popularize on-device ANI is enticing and beneficial, but not entirely without issues.

4. Conclusion

In conclusion, the integration of Artificial Narrow Intelligence (ANI) into consumer devices offers numerous benefits, including enhanced privacy, reduced bandwidth, and lower server costs, while also presenting substantial challenges. GPUs, with their established presence and versatile capabilities, provide a practical solution for many users, particularly in desktop environments. However, their lack of specialization and higher power consumption render them less suitable for mobile and IoT applications.

Conversely, NPUs offer highly efficient and specialized processing power for neural network tasks, making them an attractive option for enhancing AI capabilities in various devices. Yet, their higher costs and limited versatility pose barriers to widespread adoption. Integrating NPUs into System on a Chip (SoC) designs presents a promising avenue, offering improved performance and energy efficiency. However, this approach also faces challenges related to flexibility, thermal management, and the complexities of advanced manufacturing.

Ultimately, as numerous studies have indicated, the future of on-device ANI will likely involve a combination of these approaches, leveraging the strengths of both GPUs and NPUs while addressing their respective limitations. As technology continues to evolve and manufacturing processes advance, it is anticipated that more efficient, cost-effective, and versatile solutions will emerge, making advanced AI capabilities more accessible to a broader range of consumers.

References

[1] Mökander, J., Sheth, M., Watson, D. S., & Floridi, L. (2023). The switch, the ladder, and the matrix: Models for classifying AI systems. *Minds and Machines, 33*(1), 221–248. https://doi.org/10.1007/s1102302209620y

[2] Backlinko Team. (2024, June 4). ChatGPT statistics 2024: How many people use ChatGPT? *Backlinko*. https://backlinko.com/chatgpt-stats

[3] Zhang, M. (2023, January 26). ChatGPT and OpenAI's use of Azure's cloud infrastructure. *Dgtl Infra*. https://dgtlinfra.com/chatgpt-openai-azure-cloud/

[4] Ham, M., Woo, S., Jung, J., Song, W., Jang, G., Ahn, Y., & Ahn, H. J. (2022, January 16). Toward among-device AI from on-device AI with stream pipelines. *ArXiv.org.* https://doi.org/10.48550/arXiv.2201.06026

[5] Moon, J. J., Lee, H. S., Chu, J., Park, D., Hong, S., Seo, H., Jeong, D., Kong, S., & Ham, M. (2024, January 4). A new frontier of AI: On-device AI training and personalization. *ArXiv.org.* https://doi.org/10.48550/arXiv.2206.04688

[6] Schreiner, M. (2023, July 11). GPT-4 architecture, datasets, costs and more leaked. *THE DECODER.* https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/

[7] Intel. (2024). CPU vs. GPU: What's the difference? *Intel.* https://www.intel.com/content/www/us/en/products/docs/processors/cpu-vs-gpu.html

[8] Amazon Web Services, Inc. (2024). GPU vs CPU - Difference between processing units. *Amazon Web Services, Inc.* https://aws.amazon.com/compare/the-difference-between-gpus-cpus/

[9] Frankenfield, J. (2021, September 7). Graphics processing unit (GPU). *Investopedia.* https://www.investopedia.com/terms/g/graphics-processing-unit-gpu.asp

[10] Steam. (2024, June). Steam hardware & software survey. *Steampowered.com.* https://store.steampowered.com/hwsurvey/Steam-Hardware-Software-Survey

[11] Peru, G. (2023, December 27). What is an NPU? Here's why everyone's talking about them. *Digital Trends.* https://www.digitaltrends.com/computing/what-is-npu/

[12] Xu, D., Zhang, H., Yang, L., Liu, R., Huang, G., Xu, M., & Liu, X. (2024, July 8). Empowering 1000 tokens/second on-device LLM prefilling with mllm-NPU. *ArXiv.org.* https://doi.org/10.48550/arXiv.2407.05858

[13] Microsoft. (2024). All about neural processing units (NPUs) - Microsoft Support. *Support.microsoft.com.* https://support.microsoft.com/en-au/windows/all-about-neural-processing-units-npus-e77a5637-7705-4915-96c8-0c6a975f9db4

[14] Li, Q., Zuo, D., Feng, Y., & Wen, D. (2024). Research on high-performance Fourier transform algorithms based on the NPU. *Applied Sciences, 14*(1), 405. https://doi.org/10.3390/app14010405

[15] Daimagister. (2024). Are NPUs the key to disrupting Nvidia's AI dominance? | Resources. *DAI Magister.* https://www.daimagister.com/resources/npu/

[16] Al Bawaba. (2024). Apple goes against the AI tide, developing on-device large language model. https://www.proquest.com/newspapers/apple-goes-against-ai-tide-developing-on-device/docview/3043117121/se-2?accountid=6387