**Neural Processing Units: The Future of AI Acceleration**

Mark Timpson, Jaylin Mendoza, Chris Darnell, Kamdon Bird

Salt Lake Community College

CSIS 2810

Professor Hau Moy Kwan

December 4, 2024

### Neural Networks and NPUs - How Do They Work?

A neural processing unit (NPU), also known as an AI accelerator, is a specialized microprocessor designed to perform neural network and machine learning computations much more efficiently than general-purpose processors like CPUs. NPUs are often included in smaller electronics that have AI capabilities where computing power is limited, such as smartphones, wearable tech, and IoT (internet of things) devices. In these applications, maximizing computing power and efficiency are key. To perform AI tasks exceptionally well, NPUs use parallel processing, low-precision arithmetic, high-bandwidth memory, and hardware acceleration (Schneider & Smalley, n.d.).

At the core of current AI applications, neural networks are a model for machine learning that mimics the structure of the human brain. It consists of layers of nodes that simulate neurons. The nodes are interconnected by weighted connections, simulating synapses. Neural networks are trained by increasing or decreasing the strength of these connections based on expected output for given inputs. This allows for computer-generated predictions without explicit instructions (Malatino, 2023).

NPUs are specially designed to be fast and efficient at the common case of operations for the neural network. For example, NPUs handle inputs multiplied by weights - a matrix operation - through parallel processing. CPUs often do these tasks sequentially, which is much slower. NPUs are also designed to calculate “neuron” activation over many nodes, which is fundamental to the inference in neural networks. Standard CPUs are designed to process a variety of operations and arithmetic precisely, using relatively few, large processing cores. NPUs perform a select few, repetitive operations using low-precision arithmetic across thousands of processing elements that work in parallel, capturing the ideal form of multiprocessors (Patterson & Hennessy, 2017). This allows for special design to ensure maximum speed and efficiency for these operations.

### Benefits of Neural Processing Units

As we’re growing in this new age of AI, a critical question emerges: how will these NPUs revamp the future of computing, and what advantages will they offer to the average user? The key advancements of NPUs include their ability to deliver exceptionally high performance, low latency, and seamless compatibility with AI workloads, enabling more efficient data processing and real-time decision-making (Chambers, 2024). One breakthrough innovation that incorporates NPUs is the new generation of Windows computers, known as Copilot+ PCs, powered by the Snapdragon® X processors. This chip is currently the only one that’s compatible with an NPU that can execute 40 trillion operations per second (TOPS), which is a staple for AI processing speed. However, Intel has announced other systems on a chip (SoCs) that aim to exceed this performance threshold in the coming years, marking a significant leap forward in computing power (Nzube, 2024). Along with these revolutions, the Snapdragon® X allows for easy cloud management, improved energy efficiency, and better remote management (Miller, 2024).

The Copilot+ PCs introduce a range of new features designed to enhance the user experience. These include improved graphics, an improved OneNote tool that not only takes and summarizes notes but also converts them into to-do lists, and live captions that provide real-time audio transcription and translation (Miller, 2024). In return, this frees up the workload of GPUs and CPUs. These new AI-driven features can also greatly benefit businesses with efficient systems, rapid task execution, and dedicated processing power to crucial business operations. Essentially, using AI-powered computers like the Copilot+ isn't just a smart choice for individuals seeking better performance, it’s also great for businesses striving to cut costs, and stay ahead in an era where more businesses continue to adopt AI softwares (Chambers, 2024).

### Applications of Neural Processing Units

A benefit of incorporating NPUs into computing systems is a significant decrease in latency, which is hugely advantageous to applications that operate in time-sensitive situations. One such application is autonomous vehicles, or self-driving - where vast amounts of incoming data from multiple cameras, LiDAR, and other sensors for tracking multiple objects in real-time requires fast and accurate decision making and response times. In such scenarios, an NPU has been shown to provide a measurable improvement on latency, along with increased stability and lower error rates (Sheikh et al., 2024).

Much of the performance improvement gained from including an NPU coprocessor is due to the parallel processing opportunities that exist from such a hardware configuration, however an additional benefit of including NPUs in computing devices is the ability to offload processing of AI based tasks from cloud computing servers to on-device computation solutions.

Traditionally, large models require significant hardware resources to handle the processing load, but NPU coprocessors unlock the ability to run models on smaller portable devices. An example application of this is the on-device generative AI tools that are beginning to feature on smartphones, such as Apple Intelligence on the newest iPhones. Handling these generative tasks on the consumer device directly has the benefit of decreasing server load, network bandwidth usage, and results in a faster and improved user experience.

Integrated NPUs can have a direct impact on personal health as well, as these dedicated processing units find their way into medical devices such as defibrillators, biometric monitoring devices, and other advanced diagnostic equipment. The increase in processing capability due to NPU usage in such systems can lead to faster and more accurate diagnosis, leading to better patient outcomes and overall health.

### The Current Challenges of Neural Processing Units

Neural Processing Units are designed for specific tasks like neural network computation and machine learning, making them essential for AI-driven applications, particularly in edge computing and IoT devices. However, they face limitations when applied to broader technology uses. Their main drawbacks include limited use cases, integration complexity, high costs, and compatibility issues with existing systems.

Unlike CPUs and GPUs, NPUs excel in specific computations but struggle with general computing tasks like running office applications or managing user interfaces. For general computing, GPUs are still the go-to, especially with applications like blockchain, where parallel processing is key.

Integrating NPUs into existing systems is challenging. Companies often need to redesign both hardware and software, which requires expertise in AI and custom-built solutions that not every organization can provide. Even newer systems may not support retrofitting NPUs easily.

NPUs are also costly. They’re custom-built for specialized tasks, which means significant investment in research and development. This makes them less accessible for smaller companies, who may struggle with the high cost of integration, maintenance, and support.

Adopting NPUs also requires applications to be reworked, demanding both time and resources. Since NPUs are a new technology, their software ecosystem is still limited, further complicating integration.

While NPUs shine in AI and machine learning, their specialized nature, high cost, and integration challenges make them less suited for general purpose computing. However, in AI environments, they’re still the most efficient solution for the tasks they were designed to perform.

**References**

Patterson, D. A., & Hennessy, J. L. (2017). Computer organization and design: The hardware/software interface (5th ed., p. 502). Morgan Kaufmann.

Schneider, J., & Smalley, I. (n.d.). What is a neural processing unit (NPU)? IBM. Retrieved December 3, 2024, from <https://www.ibm.com/think/topics/neural-processing-unit>

Malatino, K. (2023, September 27). The future of AI is wide, deep, and large. Rensselaer Polytechnic Institute. Retrieved December 3, 2024, from <https://news.rpi.edu/content/2023/09/27/future-ai-wide-deep-and-large>

Chambers, P. (2024, July 2). *Ai Golden Child: How AI PCs Will Change the Game for Enterprises*. The Fast Mode. <https://www.thefastmode.com/expert-opinion/36153-ai-golden-child-how-ai-pcs-will-change-the-game-for-enterprises>.

Miller, J. (2024, October 16). *Neural Processing Units (NPUs): The dawn of AI pcs*. CIO. <https://www.cio.com/article/3566907/neural-processing-units-npus-the-dawn-of-ai-pcs.html>.

Nzube, S. (2024, July 2). *NPUs: How They Are Changing the Computing Game with the New Copilot+ PCs*. Auslogics Blog | Tech Wisdom Blog. <https://www.auslogics.com/en/articles/npus-how-they-are-changing-the-computing-game-with-the-new-copilot-pcs/>.

Sheikh, R. A., Borkar, P. G., Sable, D. G., & Sugandhi, M. Y. (2024, 11 11). The Role of Neural Processing Units in Enhancing Real-Time Processing and Decision-Making in Autonomous Systems. *International Journal of Humanities Social Science and Management (IJHSSM)*, *6*(6), 213-217. <https://ijhssm.org/issue_dcp/The%20Role%20of%20Neural%20Processing%20Units%20in%20Enhancing%20Real%20Time%20Processing%20and%20Decision%20Making%20in%20Autonomous%20Systems..pdf>

Morrison, R. (2023, February 8). IBM launches Vela, its new cloud-native AI supercomputer. Tech Monitor <https://www.techmonitor.ai/digital-economy/ai-and-automation/ibm-vela-cloud-native-ai-supercomputer?cf-view>

Schneider, J., & Smalley, I. (2024, October 10). *NPU vs GPU: Understanding the differences*. IBM.<https://www.ibm.com/think/topics/npu-vs-gpu>

Clarke, H. (2024, May 22). *AI hardware showdown: CPU vs GPU vs NPU*. Harrison Clarke.<https://www.harrisonclarke.com/blog/ai-hardware-showdown-cpu-vs-gpu-vs-npu>