What is a neural processing unit (NPU)?

A neural processing unit (NPU) is a specialized computer microprocessor designed to mimic the processing function of the human brain. They are optimized for artificial intelligence (AI) neural networks, deep learning and machine learning tasks and applications.

Differing from general-purpose central processing units (CPUs) or graphics processing units (GPUs), NPUs are tailored to accelerate AI tasks and workloads, such as calculating neural network layers composed of scalar, vector and tensor math.

Also known as an AI chip or AI accelerator, NPUs are typically used within heterogeneous computing architectures that combine multiple processors (for example, CPUs and GPUs). Large-scale data centers can use stand-alone NPUs attached directly to a system's motherboard; however, most consumer applications, such as smartphones, mobile devices and laptops, combine the NPU with other coprocessors on a single semiconductor microchip known as a system-on-chip (SoC).

By integrating a dedicated NPU, manufacturers are able to offer on-device generative AI apps capable of processing AI applications, AI workloads and machine learning algorithms in real-time with relatively low power consumption and high throughput.

Key features of NPUs

Neural processing units (NPUs) are well-suited to tasks that require low-latency parallel computing, such as processing deep learning algorithms, speech recognition, natural language processing, photo and video processing and object detection.

Key features of NPUs include the following:

- **Parallel processing:** NPUs can break down larger problems into components for multitasking problem solving. This allows the processor to run multiple neural network operations concurrently.
- **Low precision arithmetic:** NPUs often support 8-bit (or lower) operations to reduce computational complexity and increase energy efficiency.
- **High-bandwidth memory:** Many NPUs feature high-bandwidth memory on-chip to efficiently perform AI processing tasks requiring large datasets.
- **Hardware acceleration:** Advancements in NPU design have led to the incorporation of hardware acceleration techniques such as systolic array architectures or improved tensor processing.

How NPUs work

- Based on the neural networks of the brain, neural processing units (NPUs) work by simulating the behavior of human neurons and synapses at the circuit layer. This allows for the processing of deep learning instruction sets in which one instruction completes the processing of a set of virtual neurons.
- Unlike traditional processors, NPUs are not built for precise computations. Instead, NPUs are purpose-built for problem-solving functions and can improve over time, learning from different types of data and inputs. Taking advantage of machine learning, AI systems incorporating NPUs can provide customized solutions faster, without the need for more manual programming.
- As a standout feature, NPUs offer superior parallel processing, and are able to accelerate AI operations through simplified high-capacity cores that are freed from performing multiple types of tasks. An NPU includes specific modules for multiplication and addition, activation functions, 2D data operations and

decompression. The specialized multiplication and addition module is used to perform operations relevant to the processing of neural network applications, such as calculating matrix multiplication and addition, convolution, dot product and other functions.

- While traditional processors require thousands of instructions to complete this type of neuron processing, an NPU might be able to complete a similar operation with just one. An NPU will also integrate storage and computation through synaptic weights—a fluid computational variable assigned to network nodes that indicates the probability of a "correct" or "desired" result that can adjust or "learn" over time—leading to improved operational efficiency.
- While NPU development continues to evolve, testing has shown some NPU performance to be over 100 times better than a comparable GPU, with the same power consumption.

Key advantages of NPUs

Neural processing units (NPUs) are not designed, nor expected, to replace traditional CPUs and GPUs. However, the architecture of an NPU improves upon the design of both processors to provide unmatched and more efficient parallelism and machine learning. Capable of improving general operations (but best suited for certain types of general tasks), when combined with CPUs and GPUs, NPUs offer several valuable advantages over traditional systems.

Key advantages include the following:

- **Parallel processing:** As mentioned, NPUs can break down larger problems into components for multitasking problem solving. The key is that while GPUs also excel at parallel processing, the unique structure of an NPU can outperform an equivalent GPU with reduced energy consumption and a smaller physical footprint.
- **Improved efficiency:** While GPUs are often used for high-performance computing and AI tasks, NPUs can perform similar parallel processing with far better power efficiency. As AI and other high-performance computing become increasingly common and demand more energy, NPUs offer a valuable solution for reducing critical power consumption.
- **Real-time multimedia data processing:** NPUs are designed to better process and respond to a wider range of data inputs, including images, video and speech. Augmenting applications like robotics, Internet of Things (IoT) devices and wearables with NPUs can provide real-time feedback, reducing operational friction and providing critical feedback and solutions when response time matters most.