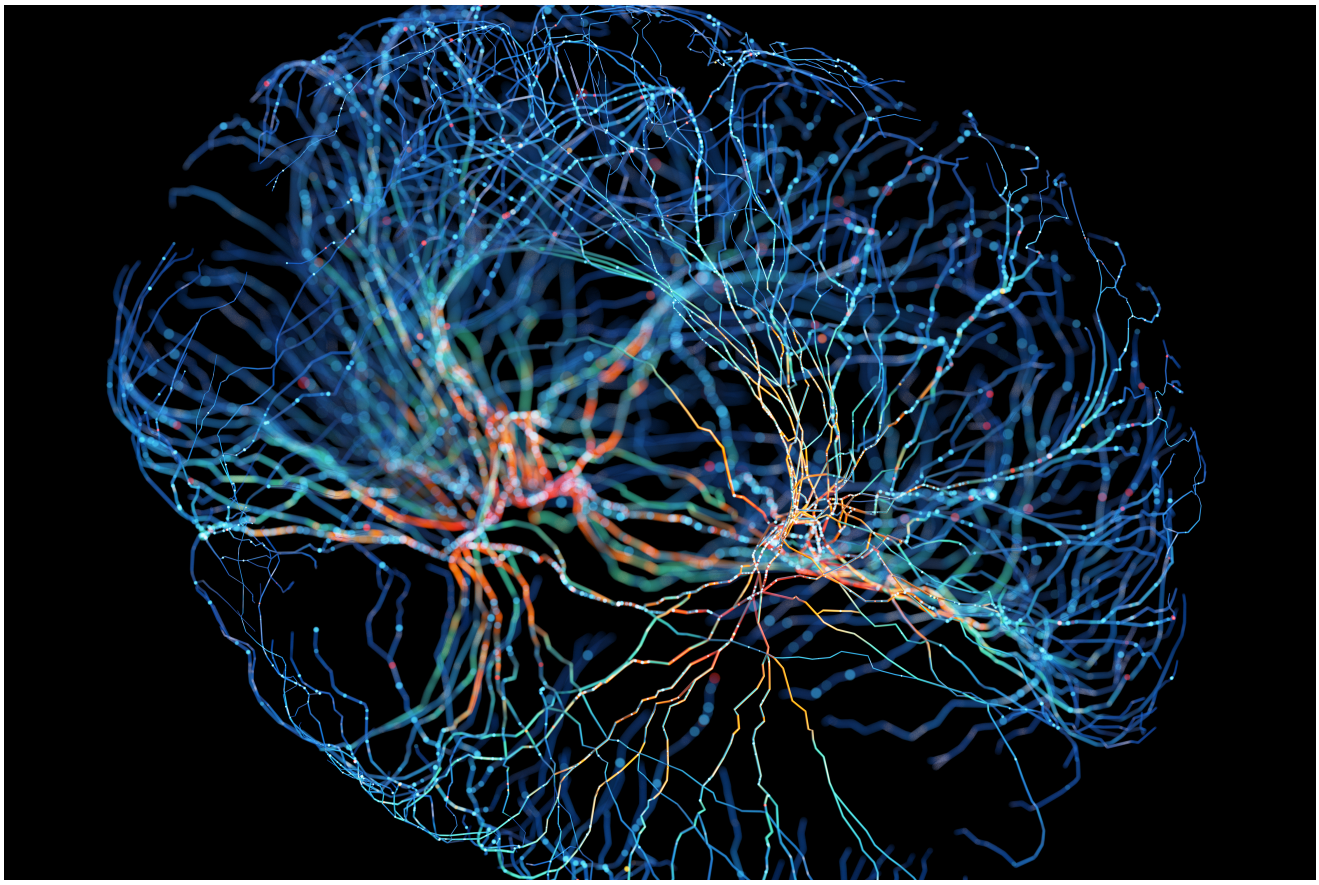
# What is a neural processing unit (NPU)?



Compute and servers

**27 September 2024**

**Authors**

**Josh Schneider**
Senior Writer
IBM Blog

**Ian Smalley**
Senior Editorial Strategist

# 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 (S

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-chi 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.

# NPUs vs. GPUs vs. CPUs

In the world of classical computer science, the central processing unit (CPU) is thought of as the “brain” of the computer. The CPU processes most traditional computing tasks and is responsible for a broad range of potential applications. While there are many different types, generally all CPUs perform operations in linear order, responding to requests in the order they come in.

From the 1950s to the 1990s, CPUs bore the brunt of practically all computer

processing, executing instructions to run programs, control systems, and manage input/output (I/O).

Demanding applications regularly pushed generation after generation of CPU designs to their hardware limits, often causing significant slowdown or even system failure. But with the advent of personal computer gaming and computer-aided design (CAD) in the 1980s, the industry required a faster, more efficient solution to rendering computer graphics.

The graphics processing unit (GPU) was initially created to offload demanding image-processing tasks from the main CPU. While GPUs tend to use fewer cores to perform linear operations, GPUs feature hundreds to thousands of cores with the ability to perform parallel processing—a process in which large tasks are broken down into smaller problems that can be solved simultaneously by multiple processors and/or cores.

Initially developed to handle video and image processing needs, the parallel processing capabilities of GPUs have made the hardware uniquely suited to other demanding computing applications, such as blockchain-related tasks and AI. While GPUs are not the only type of processor capable of performing parallel processing or parallel computing, they are well suited for parallelism. However, GPUs are not without their limitations and typically require extremely expensive power consumption to run more demanding operations. With GPUs, increased performance comes at an increased energy cost.

NPUs and other AI accelerators offer more efficient alternatives. Incorporating and improving on the advanced parallelism of GPUs, NPUs designed specifically for AI operations provide high performance with lower power consumption (and the added bonus of a smaller physical footprint).

## Comparing processors

- **Central processing units:** The “brain” of the computer. CPUs typically allocate about 70% of their internal transistors to build cache memory and are part of a computer’s control unit. They contain relatively few cores, use serial computing architectures for linear problem solving and are designed for precise logic control operations.

- **Graphic processing units:** First developed to handle image and video processing,

GPUs contain many more cores than CPUs and use most of their transistors to build multiple computational units, each with low computational complexity, enabling advanced parallel processing. Suitable for workloads requiring large-scale data processing, GPUs have found major extra utility in big data, backend server centers and blockchain applications.

- **Neural processing units:** Building on the parallelism of GPUs, NPUs use a computer architecture designed to simulate the neurons of the human brain to provide highly efficient high performance. NPUs use synaptic weights to integrate both memory storage and computation functions, providing occasionally less precise solutions at a very low latency. While CPUs are designed for precise, linear computing, NPUs are built for machine learning, resulting in improved multitasking, parallel processing and the ability to adjust and customize operations overtime without the need for other programming.

# NPU use cases

As an emerging technology, many leading computer and hardware manufacturers—including Microsoft, Intel, Nvidia, Qualcomm and Samsung—offer either stand-alone neural processing units (NPUs) or integrated variations, such as the Apple Neural Engine.

Incorporating NPUs into consumer-grade electronics offers a wide range of benefits, such as improved image recognition and optimization for AI-enabled cameras to better blur the background on video calls. Some additional applications for NPUs include the following.

## Artificial intelligence and large language models

As a type of AI accelerator, NPUs are purpose-built to improve the performance of AI and ML systems, such as neural networks. Complimenting GPUs, the improved parallelism of NPUs offer dramatic improvements for large language models that require low-latency adaptive processing to interpret multimedia signals, perform speech recognition and produce the natural language and art used in tools like AI chatbots and generative AI image and video applications.

## Internet of Things (IoT) devices

With exceptional parallel processing and self-learning capabilities, NPUs are well suited for networked IoT devices, such as wearables, voice assistants and smart appliances.

## Data centers

AI and machine learning has been a major boon for data centers seeking to optimize energy resources. High-performance and energy-efficient NPUs offer tremendous value for data centers offering better resource management for cloud computing.

## Autonomous vehicles

Autonomous vehicles like drones or self-driving cars and trucks benefit greatly from the real-time processing capabilities of NPUs, allowing for faster and better course correction based on multimedia sensor input. With unmatched parallel processing, NPUs can help autonomous vehicles interpret and process rapidly developing inputs, such as road signs, traffic patterns and even unexpected obstacles.

## Edge computing and edge AI

While cloud computing offers advanced off-site data and resource solutions for IoT, smart devices and other personal computing devices, edge computing and edge AI seek to bring critical data and compute resources physically closer to users. This reduces latency, mitigates energy consumption and bolsters privacy. Requiring less energy and offering a smaller physical footprint, NPUs are becoming a valuable component in edge computing and on-device AI.

## Robotics

Adept at handling tasks requiring machine learning and computer vision, NPUs offer critical support to the development of the robotics industry. AI-enabled robotics fr home assistants to automated surgical tools rely on NPUs to develop the ability to detect, learn from and react to their environments.

Ebook

### How to choose the right foundation model

Learn how to choose the right approach in preparing datasets and employing foundation models.

Read the ebook →

# Resources