

WHAT IS NPU

KEY TAKEAWAYS

- **NPUs like Apple's ANE and Google's TPU are integrated into their smartphones and laptops for efficient AI processing.**
- **NPUs are specialized processors for AI tasks, different from CPUs and GPUs, enhancing performance with lower power consumption.**
- **Key NPU specs like TOPS, power efficiency, precision, memory bandwidth, and compatibility impact performance and user experience.**

With NPUs enabling on-device AI tech on the SoC (System on Chip) designs used in laptops, smartphones, and single-board computers, it's important to understand and compare their specifications to optimize performance and make informed purchasing decisions.

But how do you compare the specs of an NPU like other computer hardware?

What Is an NPU?

A **neural processing unit** (NPU is a special kind of processor that handles tasks related to artificial intelligence (AI) and machine learning. In devices like smartphones and laptops, an NPU makes processor-intensive tasks such as image processing, real-time video filters, voice recognition, augmented reality, and object recognition much more efficient.

You can find NPUs on some of the most popular handheld devices today. For example, Apple's latest iPhones and M-Series laptops all come integrated with their own NPU called ANE (Apple Neural Engine). Google uses a proprietary NPU called TPU (Tensor Processing Unit) on its Pixel devices and Coral prototyping kits, and Qualcomm is integrating its ARM NPUs into their latest SoCs for smartphones and laptops.

NPUs are becoming more common because we use AI in many devices, from smartphones to large data centers. As AI grows, we need hardware that can handle it better, and NPUs are the answer.

How NPUs Differ from CPUs and GPUs

For modern handheld, wearables, and other portable devices, the NPU, CPU, and GPU are integrated into a single SoC for faster data transfer, lower power consumption, and a smaller footprint.

- A CPU is a general-purpose processor for running software.
- A GPU is a specialized parallel processor that aids the CPU in running graphically intensive tasks.
- An NPU is an even more specialized processor that aids the GPU and CPU in machine learning and neural network processing tasks such as object detection, speech recognition, and AI processing.

Essentially, CPUs are great at handling many different applications but are not optimized for graphics and AI tasks. GPUs are better for graphics and AI tasks than CPUs because they can do many things simultaneously. NPUs excel at high-level parallel computations like GPUs but with lower power requirements, making them ideal for bringing the demanding calculations of AI processing to mobile devices.

How to Compare NPU Specs

When comparing NPUs, it's important to understand the key specifications that affect their performance. Here are some of the most important specs to look out for:

- **TOPS (Tera Operations Per Second):** This is a measure of how many trillion operations an NPU can perform per second. Higher TOPS values generally indicate better performance.
- **Power Efficiency:** Measured in TOPS per watt, this spec tells you how efficient the NPU is in terms of power consumption. Higher efficiency means longer battery life for mobile devices.
- **Precision:** NPUs can handle different levels of precision, such as 8-bit, 16-bit, and 32-bit operations. Lower precision can speed up computations and save power but might reduce accuracy. For many AI tasks, 8-bit precision is sufficient.
- **Memory Bandwidth:** This measures the amount of data the NPU can read or write per second. Higher memory bandwidth can help in faster processing of large datasets.
- **Supported Frameworks:** NPUs should support popular AI frameworks like TensorFlow, PyTorch, and Caffe. This ensures that developers can easily deploy their AI models on the NPU.

But as with most bits of computer hardware, higher is better. An NPU capable of 40 TOPS, like Qualcomm's Snapdragon X Elite chips, will have better AI capabilities than NPU capable of 10 TOPS, like Intel's Meteor Lake chips.

How an NPU Will Make Your Next Laptop or PC Better

Buying a laptop or PC with an NPU will have several key benefits:

- **Enhanced Performance for AI Tasks:** NPUs are specifically designed to handle AI tasks efficiently, which means faster performance for applications that rely on machine learning and neural networks.
- **Better Battery Life:** NPUs are optimized for low power consumption, which can help extend the battery life of laptops and mobile devices.
- **Improved User Experience:** Features like real-time language translation, voice assistants, augmented reality applications, and enhanced image processing can run more smoothly and efficiently with an NPU.
- **Offloading Work from CPU/GPU:** By handling AI-specific tasks, NPUs free up the CPU and GPU to perform other functions, leading to overall better system performance.
- **Future-Proofing:** As AI becomes more integrated into everyday applications, having an NPU ensures that your device is ready for future software and technology advancements.

NPUs are transforming the way AI tasks are handled in modern devices, and their role in computing will only grow as AI becomes an essential part of our daily lives. With NPUs gaining wider adoption, understanding and comparing NPU specifications can help you make informed decisions to optimize performance and future-proof your devices.