PROWESS

Technical Research Study

How to Understand and Evaluate AI PCs for Your Apps and Workloads

If you're purchasing an AI PC, don't rely solely on TOPS or benchmark scores to gauge AI performance.

Commissioned by Intel

Executive Summary

AI workloads are becoming prevalent across every aspect of our work and our personal lives. AI is transforming business productivity, public and private research, analytics, content creation, and so much more. Given this explosion of AI workloads, it's no surprise that we're now seeing AI PCs on the market with hardware and software designed specifically to accelerate performance for AI-based applications.

If you're evaluating AI PCs for purchase, it can be challenging to know which ones are best suited to your particular use cases. This is particularly true if you're an IT decision-maker (ITDM) responsible for conducting a PC refresh for your organization. How can you determine which PCs will meet the needs of your productivity workers by supporting productivity assistants and enhancing collaboration? Which ones are best for handling AI-based analytics, video editing, or software development? And which ones can best support AI-based security features without dragging down performance?

Benchmarks can be useful, but traditional benchmarks aren't focused on AI workloads. New benchmarks are in development, but they might take time to reach the levels of maturity needed to offer a reliable indication of AI performance across real, application-based use cases. In the meantime, other approaches to evaluating AI performance might be more useful, depending on your specific workloads and needs.

In this paper, sponsored by Intel, Prowess Consulting helps unravel the jargon, explains what common benchmarks actually measure, and steps through other considerations needed to make an effective AI PC purchase for yourself or your organization.

Industry Landscape

AI has become the definitive industry disruptor of the 21st century. It's infiltrating most aspects of our home and work lives, spanning cars, appliances, smartphones, and PCs. It's revolutionizing medicine, research, finance, and manufacturing.

Businesses are responding to this dramatic shift by embracing AI in the workplace as a key driver for expanding service offerings and increasing productivity. According to a study by Forbes, 64 percent of businesses expect AI to increase productivity.[1]

This expectation is likely in response to the variety of PC apps available today with built-in AI features for collaboration, videoconferencing, content creation, and more. For example, Microsoft® Copilot brings generative AI features directly into Windows® and the Microsoft Edge® browser. Other apps offer real-time transcriptions for videos and videoconference calls, background blur for webcams, enhanced malware detection for security, text-to-image generation, editing tools for content creation, and even voice-operated personal assistants.

At first, most of these AI workloads were run in the cloud. But as they've expanded in functionality and computational complexity, many AI apps have shifted to run locally on users' PCs. The cloud offers simplicity and fewer storage constraints for large datasets, but it also adds challenges around responsiveness, privacy, security, and connectivity. By running AI workloads directly on a local PC, businesses can have greater control over user privacy and data security. Additionally, PC-based apps can significantly reduce latency and are less reliant on network connectivity and speed.

Highlights

Recommendations for choosing and evaluating an AI PC:

Purchase a device with a **CPU, GPU, and NPU**

Choose a device **built on x86 architecture**

Run benchmarks and **real-world workloads**

Test using **precision types that align with your applications**

This shift to PC-based AI workloads has kicked off a surge in the growth of new PCs featuring hardware and software optimized for AI. According to Canalys, by 2027, 60% of PCs shipped will have on-device AI capabilities.[2] AI PCs are distinguished from traditional devices by having some combination of more efficient central processing units (CPUs) with built-in accelerators, built-in graphics processing units (GPUs), and built-in neural processing units (NPUs) that can be better tuned to offload AI processing tasks. As the abilities of these new PCs expand, new apps are being developed to make use of those capabilities, leading to a "virtuous cycle" for AI workloads and the devices that run them.

These rapid developments have created an interesting challenge for both consumers and businesses looking to maximize their investments in new laptops: how do you evaluate a PC for AI capabilities and performance?

Considerations for Comparing and Evaluating AI PCs

There's no shortage of benchmark tests available today for a broad range of use cases, covering gaming, productivity, content creation, battery life, and more. But when it comes to AI, traditional benchmarks might not be relevant—or, in some cases, might be relevant only when combined with other benchmarks to more fully represent real-world tasks performed by users. As a result of this complexity, consumers and businesses need to have a clear understanding of the types of AI workloads they'll be running in order to most effectively evaluate which PCs would best support those workloads.

To help bring some clarity to this process, Prowess Consulting performed extensive research on specifications, terminology, benchmarks, and other considerations in this complex and rapidly evolving landscape. This paper steps briefly through these categories and then makes some recommendations on how to best measure performance for your needs.

The Limitations of Benchmark Tests

For years, PC and chip vendors, along with online tech reviewers, have relied on benchmark tests to compare performance between devices. Benchmarks can be useful when applied judiciously, but they can also be misleading if used improperly or in the wrong context.

Synthetic benchmarks, such as PassMark®, use predefined calculations and functions to test performance. This approach provides standardized tests that are easy to run and use for quick device comparisons. However, synthetic benchmarks might not provide an accurate depiction of performance for real-world scenarios. Other benchmarks, such as CrossMark® or UL Procyon® Office Productivity, rely on APIs provided by software or run actual software application code, which means they come closer to real-world performance. These benchmarks offer more useful results because they measure performance using the actual software that end users would be interacting with in their daily work or home lives. However, even these benchmarks have limited value for AI-based workloads, for several reasons.

First, measuring overall performance can be more complicated with AI because you need to consider both the real-world applications and the AI-based features within those applications. In addition, real-world benchmarks for AI are not as prevalent or mature as their non-AI equivalents. Benchmarks for traditional, non-AI workloads have developed organically over many years. AI, however, is still in the early days, with devices, processors, accelerators, software, and benchmark tests all in a state of flux.

Despite these challenges, some existing and emerging real-world benchmarks for AI can be employed for performance testing. We discuss these benchmarks further in the recommendations section of this paper.

The Industry Spin Around TOPS

Trillions of operations per second (TOPS) is a metric commonly used to simplify comparisons of processor performance in AI-capable devices. Specifically, TOPS represents the number of computing operations an AI chip (such as an NPU) can handle in one second. TOPS is useful in providing a single number intended to encapsulate an AI chip's computational capability, but its usefulness can be limited or even misleading. This is because the measurement doesn't differentiate between the types or quality of operations the chip can process.

Furthermore, the system on chip (SoC) in a modern AI PC typically includes multiple processors, such as a CPU, a GPU, and an NPU. A TOPS specification might relate to only one of those processors. As a result, if you primarily run a CPU-dependent workload, a higher TOPS number for an NPU is not necessarily helpful.

In addition, TOPS does not take factors such as software optimization, memory bandwidth, or specific use cases into account. This is analogous to core counts in PCs, where a higher core count doesn't always translate to higher performance if other considerations—memory, accelerators, and so on—are the limiting factors. In other words, TOPS is a convenient specification for quick comparisons, but it might not represent the performance you will see from real-world AI-based applications.

Numerical Precisions and Processor Optimizations

The way AI inferencing is performed within an application can have a significant impact on the application's performance and efficiency. This often comes down to the type of numerical precision used in AI models, how inference models are optimized for a specific processor (CPU, GPU, or NPU), and the built-in capabilities of those processors.

To unpack all of this, we need a brief explanation of mixed precision in AI inference models. Developers can use various data types and numerical precisions within a single AI model or algorithm in their code, based on how much weight they want to give accuracy of results versus computational efficiency and performance.

Some common numerical precisions are FP64, FP32, FP16, INT16, and INT8, where "FP" stands for floating point and "INT" stands for integer, as shown in Figure 1.

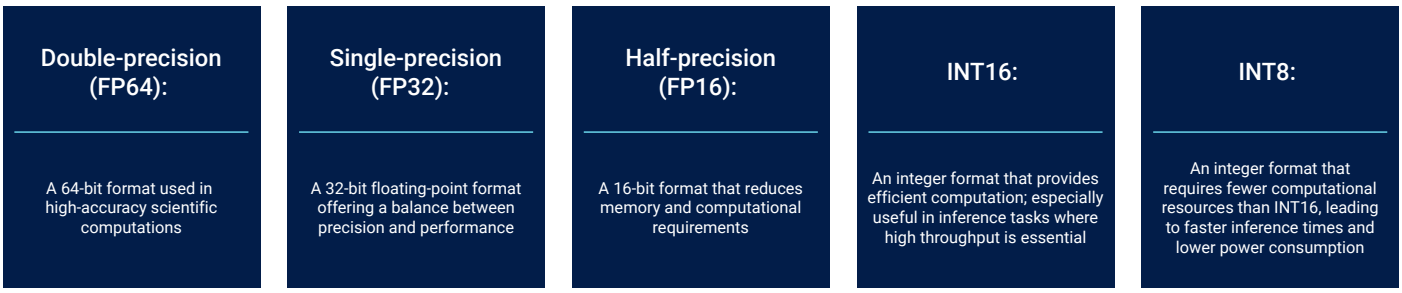

Figure 1 | Numerical precision models balance accuracy with performance and efficiency

Higher-precision models like FP64 or FP32 allow for more accurate representations of numerical values in complex AI models. Lower-precision floating-point or integer formats like FP16 or INT8 are used when speed and efficiency are of greater importance than achieving extreme levels of accuracy. Most existing AI models use mixed precision, combining INT8 with FP16, INT8 with FP32, or FP16 with FP32. Combining higher and lower precision models can accelerate computations and reduce memory footprint, which is crucial for large-scale AI models. (Incidentally, this highlights another reason why TOPS might not offer a good indicator of real-world performance: many vendors report TOPS based on INT8 precision, but as stated above, most AI models use mixed precision, not pure INT8.)

The choice between data types significantly impacts the performance of AI workloads on different hardware accelerators. CPUs, with their general-purpose architecture, can efficiently handle both INT8 and FP16 or FP32 computations, for example, but they might exhibit better performance with INT8 for some inference tasks due to its lower computational overhead. GPUs, which are renowned for their parallel processing capabilities, can make use of INT8, FP16, FP32, and even FP64 precision models for accelerating workloads. NPUs, which are specialized for AI computations, often support both INT8 and FP16 data types, allowing for flexibility in optimizing performance based on the specific requirements of AI applications.

AI performance also depends on whether a given precision model is supported by the hardware accelerator in the NPU. An NPU typically has both hardware acceleration and digital signal processor (DSP) capabilities. In NPUs, DSPs are often used for tasks such as image, audio, and video processing, and can be used for running neural networks and other AI models. If a given workload on one device relies on the DSP to run a precision model or other specific operation, the performance might not match the same workload running on another device that makes use of hardware acceleration provided by the NPU, even though both devices might have the same TOPS.

In general, however, AI PCs that are built with all three processors (CPU, GPU, and NPU) offer the most flexibility to provide optimized performance regardless of the data types employed by a particular application or workload.

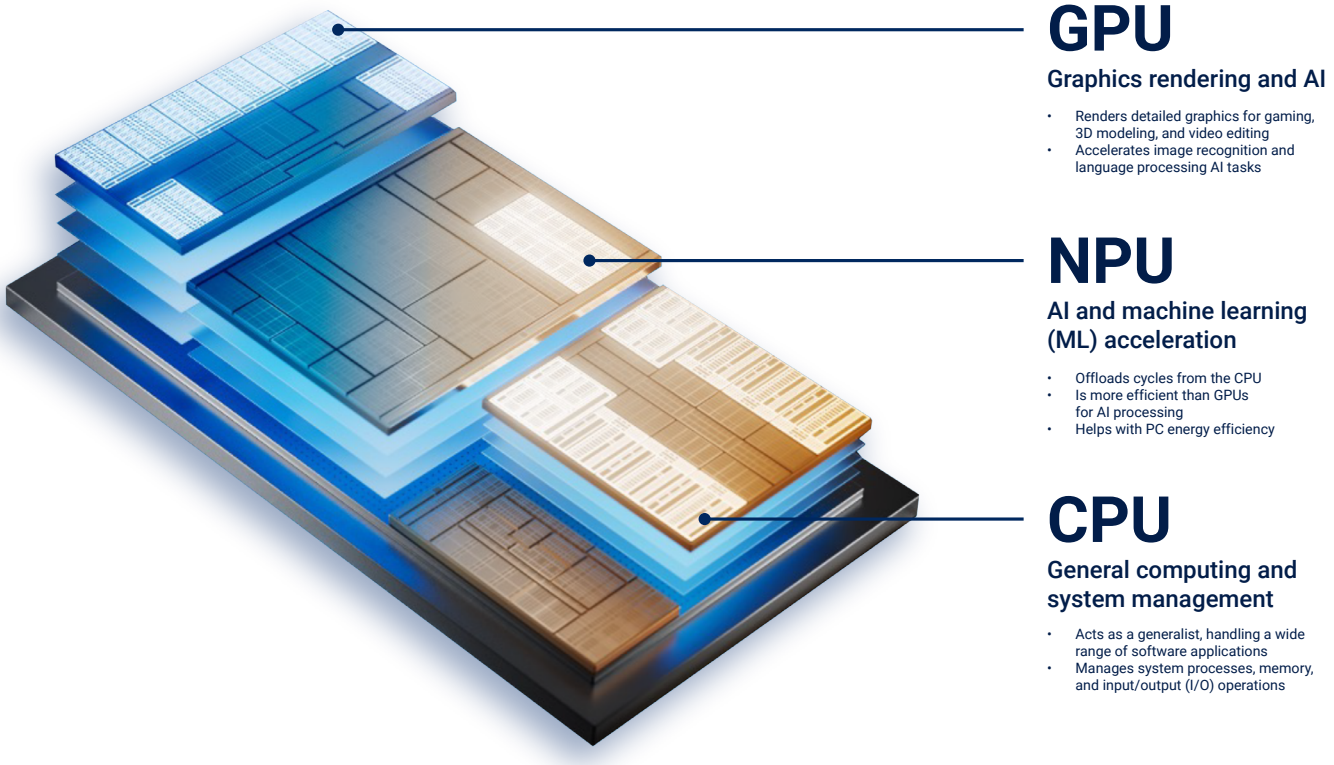

Figure 2 | AI PCs can include up to three processors that can help with AI processing

Frameworks and Tools

Developers can use specific frameworks and toolkits to create and optimize AI inferencing in applications. For example, Open Neural Network Exchange (ONNX) helps facilitate model interoperability, allowing developers to switch between frameworks and precision formats.

Another option, the OpenVINO™ toolkit, optimizes deep learning models specifically for Intel® hardware. The OpenVINO toolkit makes use of Intel CPUs, GPUs, and NPUs to benefit from hardware-acceleration technologies such as Intel® Deep Learning Boost (Intel® DL Boost), which enhances INT8 and mixed-precision performance for AI on the CPU and GPU.

Microsoft® DirectML is another framework for AI workloads. As part of the Microsoft DirectX® family, DirectML enables high-performance machine learning (ML) inference on Windows devices and supports various precision types.

For PCs built on AMD hardware, the AMD ROCm™ framework enables developers to harness GPUs for AI workloads, and AMD Ryzen™ AI software includes tools and libraries for NPUs on AMD Ryzen AI PCs. However, our research found that the OpenVINO toolkit offers greater flexibility and a single framework across CPU, GPU, and NPU (with support for CPU, GPU, and NPU optimizations) and has a broader ecosystem of support than the AMD offerings.

For users of devices built on Qualcomm® processors, the Snapdragon Neural Processing Engine (SNPE) is a Qualcomm® Snapdragon® software-accelerated runtime designed for the execution of deep neural networks. SNPE helps optimize software to make use of the Snapdragon CPU, Qualcomm® Adreno™ GPU, or Qualcomm® Hexagon™ DSP.

If you or your organization are not developing AI-based applications, the frameworks might not seem relevant to a discussion of AI PCs. However, the AI PC you choose could directly impact how well your AI workloads have been optimized, based on the available frameworks and underlying PC silicon. For example, a PC with a CPU, GPU, and NPU that also exposes hardware acceleration for AI can take full advantage of the optimizations provided in frameworks such as the OpenVINO toolkit.

Recommendations for Evaluating an AI PC

Prowess Consulting has established several recommendations for evaluating AI PCs based on the factors discussed above. First, for ideal performance and efficiency, select a PC with a built-in CPU, GPU, and NPU. We also recommend selecting an AI PC that supports frameworks that are commonly used by the software ecosystem for optimizing AI performance with mixed-precision AI inferencing workloads.

If you're interested in measuring performance for your AI applications and workloads, we recommend that you don't rely solely on pure specification numbers (like TOPS) or synthetic benchmarks like PassMark as reliable indicators of real-world performance. Wherever possible, run real-world benchmarks or test your specific applications with a focus on the expected AI use cases for your workers. Keep in mind that a combination of benchmarks and workloads will provide a more holistic view of performance than just a single benchmark. Also, if you're running benchmarks that let you configure the precision model, select the one that best aligns with the applications you'll primarily be running in your environment.

Unfortunately, the ecosystem of AI-specific real-world benchmarks is still early in development. Based on our research, the most relevant benchmarks available at the time of publication of this study are from UL Procyon. For example, the UL Procyon AI Image Generation Benchmark can be useful, but note that it targets only GPUs. The UL Procyon Image Generation Benchmark measures inference performance using Stable Diffusion® workloads with FP16 precision. UL Procyon also offers an AI Computer Vision Benchmark, which runs machine-vision tasks using common neural networks running on CPUs, GPUs, and supported NPUs.

Geekbench® ML is another option for evaluating AI workload performance on the CPU, GPU, and NPU. This benchmark uses computer vision and natural language processing (NLP) ML tests to measure and compare cross-platform performance.

Over time, we can expect to see more AI benchmarks that are focused on end-user experiences using real-world applications. In the meantime, consider conducting your own platform-level comparisons running apps and workloads that represent typical use cases for your workers. It's also important to evaluate performance and power needs for your workers and organization because this will influence where you run and test your workloads. Is peak performance the primary goal? Performance/power? These considerations will dictate the best AI engine to specify for a given workload or benchmark.

Keep in mind that AI performance is not the only factor in evaluating a PC—even an AI PC—for purchase. Traditional apps and workloads still account for a significant percentage of processor utilization for your workers. Therefore, a broader approach to testing both overall performance and AI-specific tasks will give you a more holistic view of an AI PC's capabilities. For example, you could run the UL Procyon Office Productivity Benchmark to compare the performance of Microsoft® Word, Excel®, PowerPoint®, and Outlook® software. Then run Windows Studio Effects to test support and performance of AI-based features like background blur in collaboration apps such as Microsoft Teams®. You could also run the UL Procyon AI Computer Vision Benchmark or Geekbench ML for additional AI-focused data. Additionally, you could run UL Procyon benchmarks concurrently with Windows Studio Effects to gain a more complete picture of overall system performance.

1. **Choose an AI PC with a CPU, a GPU, and an NPU** for maximum flexibility, performance, and efficiency.
2. **Choose a device built on x86 architecture** for the widest application compatibility.
3. **Identify key apps and workloads to evaluate** for performance and power/performance, based on your needs.
4. **Run a combination of benchmarks and real-world workloads** to test performance, efficiency, and compatibility for your specific work environment.
5. **Test using the precision type and processor that best align with your applications**. Focus on mixed precision with FP16 and FP32, since these are the precision types used by most applications.

The Ecosystem Is Evolving Rapidly

AI PCs are disrupting the personal computing ecosystem with new processors, software, and accelerators. These devices are built to run increasingly complex AI workloads directly on the PC instead of in the cloud. This approach has several benefits, from reduced latency to increased data security and privacy.

With several AI PCs to choose from, buyers might feel overwhelmed when it comes to understanding which device will perform best for their workloads. Although several benchmark vendors are trying to establish their software as ideal for measuring AI performance, it's still too early to consider comparing platforms based solely on existing benchmarks or TOPS. Today's options are incomplete because the technology is new and the ecosystem is evolving at a rapid pace. For example, whether an application runs best on the CPU, GPU, or NPU varies and can change over time.

To complicate things further, some language models might be more memory-bound than compute-bound, in which case a higher TOPS specification wouldn't necessarily equate to better performance, compared to memory speed. For example, a device with 100 TOPS wouldn't necessarily run a large language model (LLM) any faster than a device with 20 TOPS if both devices are configured with 120 GB/s memory.

Even though the current landscape is evolving and complex, there are several ways AI PC buyers can best meet their needs for both traditional and AI-dependent workloads.

As stated earlier, the primary consideration for purchasing an AI PC is to ensure that it has three processor components—a CPU, a GPU, and an NPU—thus providing a comprehensive platform for on-device processing of AI workloads.

We also recommend selecting a system built on x86 architecture because this platform currently provides the broadest ecosystem for application and feature compatibility, compared to other options. For example, Apple includes an SoC with a CPU, a GPU, and an NPU for broad AI support, but users are limited to apps and tools within the Apple ecosystem. For some users, those limitations might be acceptable. But for others—especially workers at enterprise and small-to-medium-size businesses (SMBs)—the lack of Windows-based applications and open-source tools would likely be a deal-breaker.

Similarly, there are new ARM®-based systems built on Qualcomm Snapdragon processors that have broader compatibility with Windows for ARM and can run some apps natively. However, many apps were compiled for x86 architecture and run only through a translation layer. Others have only limited support, are missing native drivers, or don't run at all. Things are changing rapidly in this space, but the safest recommendation as of the publication of this paper is to rely on the broader, proven architecture and ecosystem offered by x86 PC vendors such as Intel and AMD. As major software vendors such as Adobe and Microsoft provide upgrades and performance improvements to their applications, they're likely to prioritize x86 platforms first. They're also more likely to test compatibility on AI PCs from top vendors, such as Dell Technologies, HPE, Lenovo, Samsung, and ASUS, because that's where the mass adoption is.

The Bottom Line

It would be convenient to rely on a single specification like TOPS or on simple benchmarks to determine performance, but only a combination of benchmarks with relevant, real-world workloads can give a true indication of how a specific AI PC will perform for given use cases.

Regardless of how you test performance, we recommend that you select a PC with the broadest support for AI—both in terms of hardware-acceleration capabilities and in terms of software and ecosystem support. Based on our research, AI PCs built on Intel® Core™ Ultra processors are a good choice for most users because these devices offer the combined power of a CPU, a GPU, and an NPU, and because they support proven AI frameworks such as the OpenVINO toolkit and a broad ecosystem of tools and software across the industry.

[1] Forbes Advisor. "How Businesses Are Using Artificial Intelligence In 2024." April 2023.

[2] Canalys. "Now and next for AI-capable PCs: Revolutionizing computing: AI PCs and the market outlook." January 2024.

0724/230293