MARKET ANALYSIS

The AI PC Opportunity

Sponsored by intel

SUMMARY

The explosion of interest in AI, particularly generative AI, has created a level of excitement in the tech industry unlike almost anything that's come before it. Along with that excitement, however, has come a growing awareness that not all of the new capabilities that AI enables can be computed in the cloud. As a result, it's led to a new focus on running AI workloads on PCs and other client devices. PC semiconductor suppliers like Intel are responding to this interest with new chips and software that are optimized to run these types of applications. From the integration of new AI accelerator architectures, such as NPUs, in next-generation SOCs like Intel's Core Ultra, to improved efforts at leveraging the power of CPUs and enhanced GPUs, the company is making a concerted effort to realize the opportunities that are available in running both generative AI and other types of AI applications on PCs. What's becoming clear from these early efforts is that a simplistic focus on TOPS performance from an NPU doesn't offer an accurate assessment of what the experience of using AI on PCs is capable of achieving. Software support across multiple areas, including developer tools, APIs embedded into the operating system, Runtime AI frameworks, deployment tools and system-level drivers are all necessary to leverage the full potential of AI on client devices.

> "The most exciting impact of AI is that it's opened up people's minds with regards to what can be done with computing devices."—Bob O'Donnell, Chief Analyst

INTRODUCTION

One of the most exciting impacts of the renewed excitement around artificial intelligence (AI)-based applications is that it's opened up people's minds with regards to what can be done with computing devices. After what seemed like decades of potential promise and then ultimate disappointment when it comes to AI, the broad-based launch and nearly immediate impact of foundation models that can power generative AI applications like OpenAI's ChatGPT has launched us into an exciting new era of computing.

Generative AI (GenAI) is enabling new ways to think about computing, creativity, productivity, communications and more and inspiring people all over the world to leverage the technology in impactful ways. Plus, the effect is extending beyond GenAI to many existing "traditional" AI-focused applications. From image and video analysis and editing, to office productivity, meeting transcription and summarization, 3D modeling and texturing, image/video object removal, and more, AI-powered applications of all kinds are getting a whole new lease on life. On top of that, more "traditional" AI-based applications, such as background blurring and audio denoising are also being seen in a new light as they can start to leverage some of the AI-focused computing resources now becoming available.

Up until now, most of the focus for AI-powered computing has been on applications and services that run in the cloud. In reality, however, there are interesting new opportunities to run these kinds of applications directly on PCs and other client devices. In fact, not only is it quickly becoming possible to do so, in some situations the performance and the output can actually be better when running locally. In addition, there are tremendous advantages related to privacy and security when you can leverage data on your own device and not send any of it out to a public cloud environment.

Some of these new possibilities are happening because of the tremendous progress that on-device AI and GenAI solutions have made over the last few months. The rapid evolution and shrinking of open-source foundation models along with technology advancements such as model quantizing are suddenly making things that many industry observers didn't expect to happen on client devices for several years possible in the next few months. In fact, the pace of innovation happening on devices over the last few months is even faster than the overall speed at which GenAI advancements have been occurring—and that's saying something!

In addition to impressive technological advances, a big part of the reason for these dramatic speed-ups for on-device AI stem from very practical issues. Most notably, there's been a widely acknowledged recognition that—based on the incredible speed of adoption of GenAI tools as well as the huge array of new offerings coming online—the existing public cloud datacenter infrastructure simply can't support all the expected demand. In addition, there are significant concerns being raised about the power requirements that these cloud-based resources would demand. Finally, issues around cost, security and efficiency all point to the fact that running all—or even most—AI workloads in the cloud simply isn't a sustainable, long-term option. As a result, on device AI solutions are quickly becoming essential to ensure that the momentum around AI-powered applications can continue to grow. More of these AI workloads simply have to move onto PCs.

THE IMPORTANCE OF A BALANCED SOC

Given this environment, a great deal of attention is being focused on the kinds of AI applications and workloads that can run on client devices such as PCs. This, of course, is directly related to the type of computing resources that modern PCs can offer. (As we'll discuss a bit later in this paper, there's also a lot more consideration to the software necessary to leverage that computing hardware.)

Thankfully, this year has seen the launch of several new PC SOC (System on Chip) architectures that can be leveraged to run AI workloads in more powerful and efficient ways than previous iterations. In particular, chips like Intel's new Core Ultra SOC (formerly codenamed "Meteor Lake") now include more flexible CPUs, more powerful GPUs, and a new type of component called NPUs (Neural Processing Units) that are specifically optimized for many types of AI workloads.

The addition of the NPU, in particular, is driving a great deal of interest in these new modern SOC architectures and the promise they have for enabling on-device AI. NPUs are designed to speed the performance of matrix multiplications and other mathematical equations that are often used by AI-powered applications or functions within larger applications. They're most useful for improving the performance of algorithms and other software components that consistently run in the background, as digital assistant applications and other "intelligence agents" often do.

As powerful as NPUs can be for some kinds of AI applications, however, they aren't a magic pill for all AI applications. In fact, many of the kinds of AI inferencing work done on PCs is still done with CPUs and other types of AI workload-related calculations can be done more efficiently by GPUs. The most important point to remember is that virtually all AI-powered processes that will run on a PC can actually be computed by any of the various architectural components of an SOC—CPU, GPU or NPU—there are just different levels of efficiency between them. Plus, on a chip like Intel's Core Ultra that has both more performant but power hungry P-Cores along with less performant but more power-efficient E-Cores, sometimes one CPU type is better for a given AI workload than the other, and vice versa. Generally speaking, CPUs are used for lightweight single inference low-latency AI, GPUs for AI intensive workloads, and NPUs for sustained AI and AI offload.

It’s a classic case of picking the right tool for a job. As many of us have undoubtedly learned in real-life, you can do a lot more with a hammer than it was originally designed for, but some tools make certain tasks a heck of a lot easier (and faster!).

Speaking of which, in addition to new capabilities it’s also important to discuss performance and efficiency benchmarks when it comes to running AI applications on PCs. The most common metric that many companies have started talking about is TOPS, or tera operations per second, a measuring mechanism originally designed for measuring mathematical calculations. In particular, there’s been a great deal of focus on the TOPS of a system’s NPU.

As it turns out, that isn’t the best measurement when it comes to comparing to real-world experience for several different reasons. First, many people don’t believe TOPS to be very reflective of true performance—it’s much more of a simplistic synthetic metric than something you can notice in real life. This is because TOPS measures the number of calculations performed, not the type of calculations, which can often have a much bigger impact on real-life performance. Another metric is TOPS/watt which looks at overall efficiency based on power consumption that occurs while performing certain calculations. While this is generally perceived to be better, it’s still not an ideal point of comparison.

The other big challenge is these ignore the basic fact that—as mentioned above—AI workloads can (and often do) occur on several different components of a complete PC system. As a result, there’s been more discussion about overall system TOPS, which combines the potential TOPS of the CPU, GPU and NPU in order to get a number that’s more reflective of systems that run several different types of AI applications.

Even system TOPS isn’t ideal, however, because it doesn’t always incorporate the experience of using different applications. Plus, another big problem with measuring AI performance on a PC is that traditional pure speed-based metrics don’t really make as much sense when it comes to AI. While some may care about how quickly a locally-running large language model (LLM)-powered chatbot can respond to your prompt, for example, most people won’t. Things like the quality of the response and the impact on battery life are likely a much more important influence on how people feel about the performance of one AI PC versus another.

And just to add yet one more challenge to the benchmarking of AI PCs, it turns out other factors can have significantly more influence on performance than the TOPS of any given component (or even the system as a whole). With many LLMs, because of the size of the data sets they use, the amount of system memory and the speed with which that memory can be accessed has a much larger impact on real-world performance than TOPS at any level ever will. To be more specific, faster access to more memory on systems with lower TOPS performance numbers can offer better real-world performance than other systems with higher TOPs specs.

AI Software Tools

Having advanced hardware capabilities are important, but as with most everything in the computing world, they're meaningless without the right software tools. AI/ML/DL models, algorithms and development frameworks are particularly critical because AI-based software is a rapidly evolving field.

As mentioned earlier, there have been tremendous advancements in bringing the extremely large cloud-based foundation models that have helped power the world of GenAI applications and services down to a size that fits and runs natively on PCs.

Smaller versions of large-scale models such as Meta's Llama 2, Google's new Gemini, and many others are creating options to run foundation models with less than 10 billion parameters directly on PCs without the need for a cloud connection. The wealth of open-source models and growth of marketplaces such as Hugging Face is also creating opportunities for developers to build models that are specifically designed to run on client devices. On top of that, there's been a great deal of exciting work done recently to quantize larger models down to fit within the confines of PC's resources. Taken together, these developments and the many more that are bound to follow have quickly turned on-device AI from short-term science fiction to real-time science reality.

In the case of AI workloads running on PCs, various system and application-level software components also play very important roles in the quality of the on-device AI experience on one PC versus another. On Windows-based PCs, for example, certain elements of the operating system play a critical role in doling out AI-based workloads and functions to the various components of a system's hardware. DirectML, in particular, serves as the "traffic cop" for AI-based applications in Windows, directing various software elements or sub-routines within a given app to the best hardware element on a PC's SOC. The latest version of DirectML incorporates a number of software refinements that Intel specifically created and gave to Microsoft to enhance the overall software ecosystem for AI-powered applications on PCs (including systems running other vendors' SOCs). Things like DirectML are very important for improving the performance in real world environments where there may be multiple AI-powered applications or agents running simultaneously. When that happens there's a critical need for balancing the mix of different software components running on different SOC hardware elements to enable different power and performance outcomes.

In addition to these system-level enhancements, getting the best possible performance from a given application typically entails working directly with software developers to ensure that their code is optimized for a given architecture. This is where's Intel's size and its huge range of on-staff software developers can often give them an advantage because of their ability to reach out to and work with a large number of ISVs (independent software vendors) working on AI-powered PC applications. A related effort is Intel's new AI Software Initiative, where they are working with the top 100 AI-focused application developers to ensure that their applications work as efficiently on Intel silicon as they possibly can.

Finally, the last, but often overlooked, part of the software story is development tools. Software developers often leverage tools from CPU vendors when creating their applications. A development environment like Intel's OpenVINO can go a long way towards speeding up and improving their efforts. OpenVINO comes with a "model zoo" of over 200 pretrained AI models that have been built and tested to run on PCs (and most efficiently on Intel SOCs). In addition, OpenVINO includes a model conversion API that lets developers bring new public or open-source models into OpenVINO, giving them more flexibility and options when it comes to building their AI-powered applications. OpenVINO also supports models trained in Pytorch and TensorFlow and serves as an integrated backend for Hugging Face Optimum and Pytorch's torch.compile, giving application developers a wide range of potential options.

PC AI Applications

Speaking of which, we're already starting to see a number of PC applications and system-level functions that leverage AI functionality. Microsoft'sWindows Studio Effects features—which have been specifically optimized to run on the NPUs on PCs that have them—offer enhanced video background blurring and improved audio noise reduction in real-time messaging functions. Best of all, they do so in a significantly more efficient way than if the same functions were to run on a PC's CPU or GPU.

Even more exciting are the possibilities that get enabled by something like Rewind.ai, which Intel demonstrated at their recent Innovation event. As its name implies, Rewind.ai records everything you do and say with your PC, from emails to documents to chats to online meetings and more, and gives you GenAI-powered summarizations and access to all that information. It hints at the kind of truly digital assistant that many of us have dreamed of since the early, significantly less powerful (and useful) iterations of things like Cortana, Siri, Alexa and more.

In a completely different vein, the latest version of Adobe's Lightroom and Vegas' Magix incorporate GenAI image and video enhancement technology and can use a local PC NPU to accelerate their efforts. We've also started to see the introduction of other GenAI-powered image generation tools that don't require a cloud-based connection to function. Like any locally running application, this greatly improves privacy and security when using those applications because none of your information can be captured in the cloud.

It’s also important to think about the potential implications that the recent reductions in LLM model size might have on even more common PC applications. While both Microsoft’s latest M365 and Google’s latest Workspace productivity suites both currently leverage the cloud for most of their GenAI functionality, the opportunity to perform some of those functions directly on the PC with these smaller LLMs is tantalizing close. Plus, even more exciting is the opportunity to do customizations of these LLMs with just your own (or just your company’s) data. This ability to do further customizations based on locally stored or company intranet stored data can create tools that are even more capable and better optimized than anything that can be accessed in the cloud. In addition, these operations can be performed more quickly if all the data they’re using is stored on the local device. In fact, this is probably one of the most important reasons for and powerful reasons why on-device AI applications make so much sense.

Another interesting option that companies are starting to explore is leveraging the concept of hybrid AI, where certain aspects of the work are done in the cloud and other portions are done on the PC. For example, imagine a scenario where an image editing program creates a screen-friendly lower resolution version of an image on the PC, but then separately creates a higher resolution version via a cloud-based model. The lower resolution version can be edited quickly on the PC, but the cloud-based version is the one that ultimately gets saved. In a business environment such as the highly regulated healthcare industry, there have also been some early examples of companies doing things such as generating customized emails about medical procedures via multiple models. In these cases, the private personally identifiable information is manipulated on a local model on the PC, while the more generic form letter parts of the email are generated with a large, cloud-based LLM. The final message is then put together by merging these two elements into the generated email. These and many other examples that we’ll likely see in 2024 highlight the more seamless nature of how generative AI work is going to be happening in the future. They also demonstrate how the PC is going to end up playing a much bigger role in generative AI than at first may be apparent.

To be clear, there are some PC-based applications that will require powerful NPUs to run efficiently (or potentially even at all), but the vast majority of applications built to run on PCs will leverage the NPU as an accelerator if it’s available. It’s conceptually similar to the role that GPUs play. On systems with more powerful discrete GPUs, certain functions may run faster or certain games can run in higher resolution modes or at faster frame rates than those with integrated graphics solutions, but in most all cases, they all still run. Over time, as more powerful NPUs become available across a broader range of the PC installed base, we’ll undoubtedly see software developers take more advantage of these capabilities. But as with most technological advancements, these evolutions take time to come to fruition.

CONCLUSIONS

There’s little doubt that GenAI and AI in general has opened an entire new vista of opportunities that can be achieved from our computing devices. While most people were resigned to having to leverage a cloud-based connection to get access to the power of these tools, it’s quickly becoming clear that on-device AI isn’t just possible, it’s necessary. And, in the not-to-distant future, it’s going to be even better than what the cloud-based experiences currently offer.

For these reasons and more it’s easy to see why so many people are so excited about what’s happening in the world of computing. As many have said, it feels like a once-in-a-generation type of opportunity to start doing things in an exciting and new way.

And despite early doubts, it’s now very clear that PCs are going to play an extremely important role in these efforts going forward. From exciting advancements in silicon architectures to important developments in PC-based software applications and tools, the PC is arguably on the cusp of being reborn in a new and exciting way. Admittedly, there are still a number of questions regarding how to best measure the performance enhancements we're about to see and it’s fair to say that it might be time for an entirely different way to think about benchmarks and other measurements.

Regardless of how these issues get resolved, it’s definitely an exciting time in the PC industry and that’s something worth appreciating.