Qualcomm

February 2024

Unlocking on-device generative AI with an NPU and heterogeneous computing

Table of contents

1 Executive summary

The generative AI revolution is here. With the growing demand for generative AI use cases across verticals with diverse requirements and computational demands, there is a clear need for a refreshed computing architecture custom-designed for AI. It starts with a Neural Processing Unit (NPU) designed from the ground-up for generative AI, while leveraging a heterogeneous mix of processors, such as the central processing unit (CPU) and graphics processing unit (GPU). By using an appropriate processor in conjunction with an NPU, heterogeneous computing maximizes application performance, thermal efficiency, and battery life to enable new and enhanced generative AI experiences.

The NPU is built from the ground-up for accelerating AI inference at low power, and it has evolved along with the development of new AI use cases, models, and requirements. A superior NPU design makes the right design choices and is tightly aligned with the direction of the AI industry.

Qualcomm is enabling intelligent computing everywhere. Our industry-leading Qualcomm® Hexagon™ NPU is designed for sustained, high-performance AI inference at low power. What differentiates our NPU is our system approach, custom design, and fast innovation. By custom-designing the NPU and controlling the instruction set architecture (ISA), we can quickly evolve and extend the design to address bottlenecks and optimize performance. The Hexagon NPU is a key processor in our best-in-class heterogeneous computing architecture, the Qualcomm® AI Engine, which also includes the Qualcomm® Adreno™ GPU, Qualcomm® Kryo™ or Qualcomm Oryon™ CPU, Qualcomm® Sensing Hub, and memory subsystem. These processors are engineered to work together and run AI applications quickly and efficiently on device. Our industry-leading performance in AI benchmarks and real generative AI applications exemplifies this.

We also enable developers by focusing on ease of development and deployment across the billions of devices worldwide powered by Qualcomm® and Snapdragon® platforms. Using the Qualcomm® AI Stack, developers can create, optimize, and deploy their AI applications on our hardware, writing once and deploying across different products and segments using our chipset solutions. Qualcomm Technologies is enabling on-device generative AI at scale.

2 Integrating processors into SoCs offers many benefits

Computing architectures have evolved over time, driven by growing user needs, new applications and device categories, and technology advancements. Initially, the central processing unit (CPU) could do most of the processing, but as computing demands increased, the need for new processors and accelerators emerged. For example, the early smartphone system consisted of discrete chips around the CPU for 2D graphics, audio, image signal processing, cellular modem, GPS, and more. Over time, the capabilities of these chips have been integrated into a single die known as a system-on-a-chip (SoC).

As an example, modern smartphone, PC, and automotive SoCs have integrated various processors — such as the CPU, graphics processing unit (GPU), and neural processing unit (NPU). **This integration in chip design provides many benefits, including improvements in peak performance, power efficiency, performance per area, chip size, and cost.**

For instance, having a discrete GPU or NPU in a smartphone or laptop would take up more board space and use more energy, affecting the industrial design and battery size. It would also result in increased data transfers across input/output pins, leading to lower performance, additional energy usage, and extra costs due to bigger boards and less shared memory efficiencies. Integration is further necessary for smartphones, laptops, and other portable devices requiring sleek industrial design within tight power and thermal constraints.

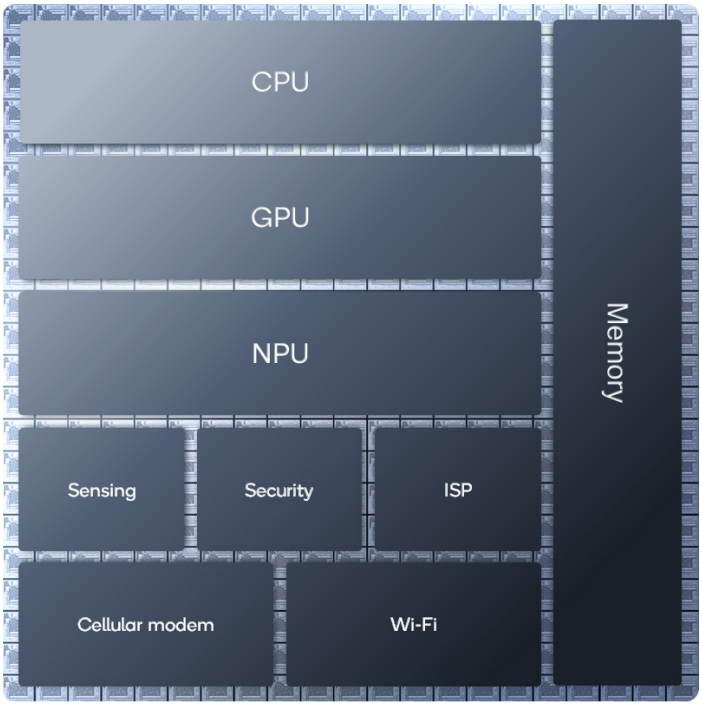

Figure 1: Modern SoCs consist of several processors integrated into the same die for improved peak performance, performance per watt, performance per area, industrial design, and cost.

3 Generative AI requires diverse processors

When it comes to AI, the integration of dedicated processors is not new. Smartphone SoCs have utilized NPUs for several generations to improve everyday experiences in the background — from superior photography and audio to enhanced connectivity and security. What is different is the growing demand for generative AI use cases across verticals with diverse requirements and computational demands. These use cases can be classified into three categories:

1. **On-demand** use cases are triggered by a user, require an immediate response, and include photo/video capture, image generation/editing, code generation, audio recording transcription/summarization, and text (email, document, etc.) creation/summarization. This includes creating a custom image while texting on your phone, generating a meeting summary on your PC, or using voice to locate the nearest gas station while driving your car.
2. **Sustained** use cases run for a longer period and include speech recognition, gaming and video super resolution, video call audio/video processing, and real-time translation. This includes using your phone as a real-time conversation interpreter while on a business travel overseas and running super resolution every frame while gaming on your PC.

3. **Pervasive** use cases constantly run in the background and include always-on predictive AI assistants, AI personalization based on contextual awareness, and advanced text auto-complete. This includes your phone suggesting a meeting with a colleague based on your conversation, or your tutor assistant on your PC adjusting study material based on your answers to questions.

These AI use cases have two key challenges in common. First, their demanding and diverse computational requirements are difficult to meet in power- and thermally-constrained devices using general-purpose CPUs or GPUs, which serve multiple needs on the platform. Second, they are constantly evolving, so implementing them in purely fixed-function hardware can be impractical. As a result, a heterogeneous computing architecture with processing diversity gives the opportunity to use each processor's strengths, namely an AI-centric custom-designed NPU, along with the CPU and GPU. For example, each excels at different tasks: the CPU for sequential control and immediacy, the GPU for streaming parallel data, and the NPU for core AI workloads with scalar, vector, and tensor math.

The CPU and GPU are general-purpose processors. Designed for flexibility, they are very programmable and have 'day jobs' running the operating system, games, and other applications, which limits their available capacity for AI workloads at any point in time. The NPU is built specifically for AI — AI is its day job. It trades off some ease of programmability for peak performance, power efficiency, and area efficiency to run the large number of multiplications, additions, and other operations required in machine learning.

By using the appropriate processor, heterogeneous computing maximizes application performance, thermal efficiency, and battery life to enable new and enhanced generative AI experiences.

4 NPU primer

The NPU is built from the ground-up specifically for accelerating AI inference at low power, and it has evolved along with the development of new AI use cases, models, and requirements. NPU design is heavily influenced by analyzing the overall SoC system design, memory access patterns, and bottlenecks in other processor architectures when running AI workloads. These AI workloads primarily consist of calculating neural network layers comprised of scalar, vector, and tensor math followed by a non-linear activation function.

Early NPUs in 2015 were designed for audio and speech AI use cases that were based on simple convolutional neural networks (CNNs) and required primarily scalar and vector math. Starting in 2016, photography and video AI use cases became popular with new and more complex models, such as transformers, recurrent neural networks (RNNs), long short-term memory (LSTM), and higher-dimensional CNNs. These workloads required significant tensor math, so NPUs added tensor accelerators and convolution acceleration for much more efficient processing. Having a large shared memory configuration and dedicated hardware for tensor multiplication not only significantly increases performance, but it also reduces memory bandwidth and energy consumption. For example, an NxN matrix multiplied by another NxN matrix would read in $2N^2$ values and would do $2N^3$ operations (individual multiplications and additions). This ratio of compute operations per memory access is N-to-one for a tensor accelerator and much less for scalar and vector accelerators.

In 2023, generative AI powered by large language models (LLMs), such as Meta's Llama 2-7B, and large vision models (LVMs), such as Stable Diffusion, significantly increased the typical model size by over an order of magnitude. Beyond compute requirements, this required significant memory and system design considerations to reduce memory data transfers for increased performance and power efficiency. Going forward, requirements for even larger models and multi-modal models are expected.

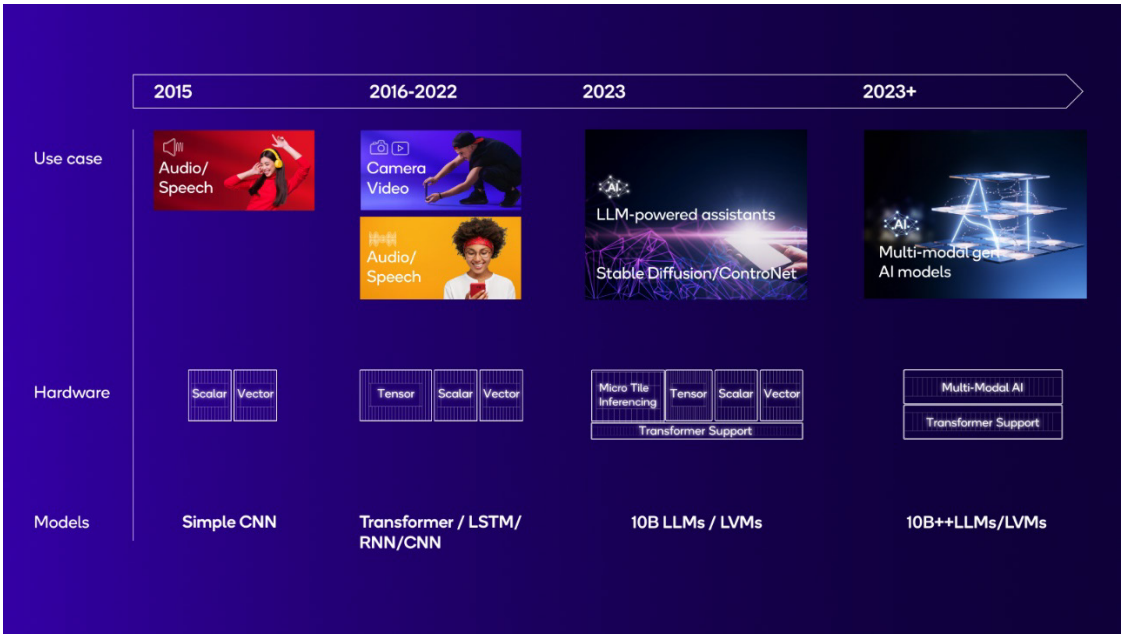

Figure 2: NPUs have evolved with the changing AI use cases and models for high performance at low power.

As AI continues to evolve quickly, many tradeoffs in performance, power, efficiency, programmability, and area must be balanced. A dedicated, custom-designed NPU makes the right choices and is tightly aligned with the direction of the AI industry.

5 Our NPU: Sustained, high-performance AI at low power

With years of research and development dedicated to its advancement, the Hexagon NPU has evolved to address the rapidly changing requirements of AI. In 2007, the first Hexagon DSP was launched on the Snapdragon® Platform — the DSP control and scalar architecture was the basis for our future NPU generations.

In 2015, the Snapdragon 820 processor was announced and included our first Qualcomm AI Engine to support imaging, audio, and sensor operations. We added the Hexagon Tensor Accelerator to the Hexagon NPU in the Snapdragon 855 in 2018. The following year, we expanded the use cases for on-device AI on Snapdragon 865 to include AI imaging, AI video, AI speech, and always-on sensing.

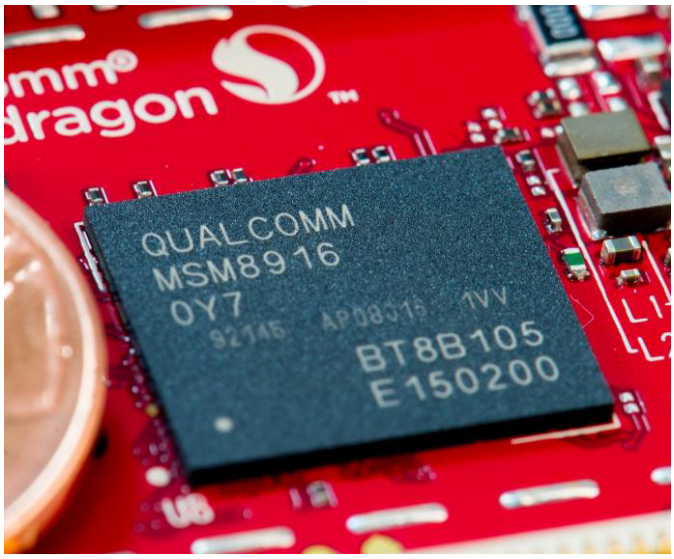

Figure 3: Snapdragon 820, launched in 2015, was the first to include the Qualcomm AI Engine.

In 2020, we achieved a major milestone with a revolutionary architecture update for the Hexagon NPU. We fused together the scalar, vector, and tensor accelerators for better performance and power efficiency. A large shared memory was dedicated for the accelerators to share and move data efficiently. **The fused AI accelerator architecture established a solid foundation for our NPU architecture moving forward.**

In 2022, the Hexagon NPU in Snapdragon 8 Gen 2 introduced a number of important enhancements. A dedicated power rail dynamically adapts and allocates power delivery according to the workload. Microtile inferencing leverages the Hexagon NPU scalar capacity to break networks into small microtiles that can be executed independently. This eliminates memory traffic between as many as 10 or more layers, maximizes utilization of the scalar, vector, and tensor accelerators within the Hexagon NPU, and minimizes power consumption. Native 4-bit integer (INT4) support enables improved power and memory bandwidth efficiency while doubling tensor throughput of INT4 layers and networks. The transformer network acceleration dramatically sped up inference for multi-head attention, which is used throughout generative AI, resulting in a staggering AI performance up to 4.35X faster on certain use cases with the MobileBERT model. Additional special hardware includes improved group convolution, activation function acceleration, and the performance of the tensor accelerator.

Our latest Hexagon NPU in Snapdragon 8 Gen 3 was designed for generative AI and is our best yet, delivering 98% faster performance and 40% improved performance-per-watt for sustained AI inferencing.[1] It includes micro-architecture upgrades across the NPU. The micro-tile inferencing was further upgraded for more efficient generative AI processing and reduced memory bandwidth. In addition, a dedicated power rail was added to the Hexagon Tensor Accelerator to enable maximum performance and efficiency for AI models that require different levels of scalar, vector, and tensor processing. The large shared memory also doubled its bandwidth. **These improvements, as well as our INT4 hardware acceleration, have resulted in the Hexagon NPU being the leading processor for inferencing large generative AI models on device.**

[1] Comparisons are made to the previous generation.

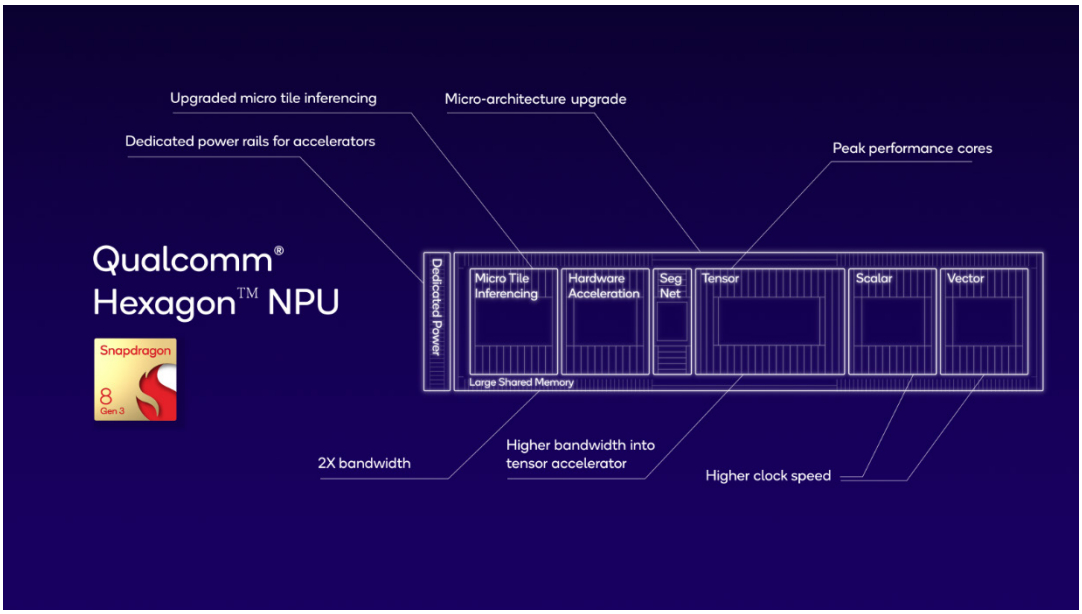

Figure 4: Hexagon NPU upgrades on Snapdragon 8 Gen 3 for leading generative AI performance at low power.

What differentiates our NPU is our system approach, custom design, and fast innovation. In our system approach, we consider the architecture of each processor, the SoC system architecture, and the software infrastructure to deliver the best solution for AI. Making the right tradeoffs and decisions about hardware to add or modify requires identifying the bottlenecks of today and tomorrow. Through full-stack AI research and optimization across the application, neural network model, algorithms, software, and hardware, we find the bottlenecks. Since we custom-design the NPU and control the instruction set architecture (ISA), our architects can quickly evolve and extend the design to address bottlenecks.

This iterative improvement and feedback cycle enables continual and quick enhancements of not only our NPU but also the AI stack based on the latest neural network architectures. Through our own AI research as well as working with the broader AI community, we are also in sync with where AI models are going. Our unique ability to do fundamental AI research supporting full-stack on-device AI development enables fast time-to-market and optimized NPU implementations around key applications, such as on-device generative AI.

Accordingly, our NPU has been refined over generations and across many learnings to remove bottlenecks. For example, many of the architecture upgrades in the NPU for Snapdragon 8 Gen 3 help with accelerating large generative AI models. Memory bandwidth is the bottleneck for LLM token generation, which means that performance is limited by memory bandwidth rather than processing. We subsequently focused on memory bandwidth efficiency. Snapdragon 8 Gen 3 also supports one of the industry's fastest memory configurations, LPDDR5x at 4.8GHz and 77GB/s, to address rising memory demands for generative AI use cases.

Building our NPU from a DSP architecture was the right choice for improved programmability and the ability to tightly control scalar, vector, and tensor operations that are inherent to AI processing. Our design approach of optimized scalar, vector, and tensor acceleration combined with large local shared memory, dedicated power delivery systems, and other hardware acceleration differentiates our solution. **Our NPU mimics the neural network layers and**

operations of the most popular models, such as convolutions, fully-connected layers, transformers, and popular activation functions, to deliver sustained high performance at low power.

6 Heterogeneous computing: Leveraging all the processors for generative AI

Generative AI models suitable for on-device execution are becoming more complex and trending toward larger sizes, from one billion to 10 billion to 70 billion parameters. They are increasingly multi-modal, meaning that they can take in multiple inputs — such as text, speech, or images — and produce several outputs.

Further, many use cases concurrently run multiple models. For example, a personal assistant application uses voice for input and output. This requires running an automatic speech recognition (ASR) model for voice to text, an LLM for text to text, and a text-to-speech (TTS) model for a voice output. The complexity, concurrency, and diversity of generative AI workloads require harnessing the capabilities of all the processors in an SoC. An optimal solution entails:

1. Scaling generative AI processing across cores of a processor and across processors
2. Mapping generative AI models and use cases to one or more cores and processors

Choosing the right processor depends on many factors, including use case, device type, device tier, development time, key performance indicators (KPIs), and developer expertise. Many tradeoffs drive decisions, and the target KPI could be power, performance, latency, or accessibility for different use cases. For example, an original equipment manufacturer (OEM) making an app for multiple devices across categories and tiers will need to choose the best processor to run an AI model based on SoC specs, end-product capabilities, ease of development, cost, and graceful degradation of the app across device tiers.

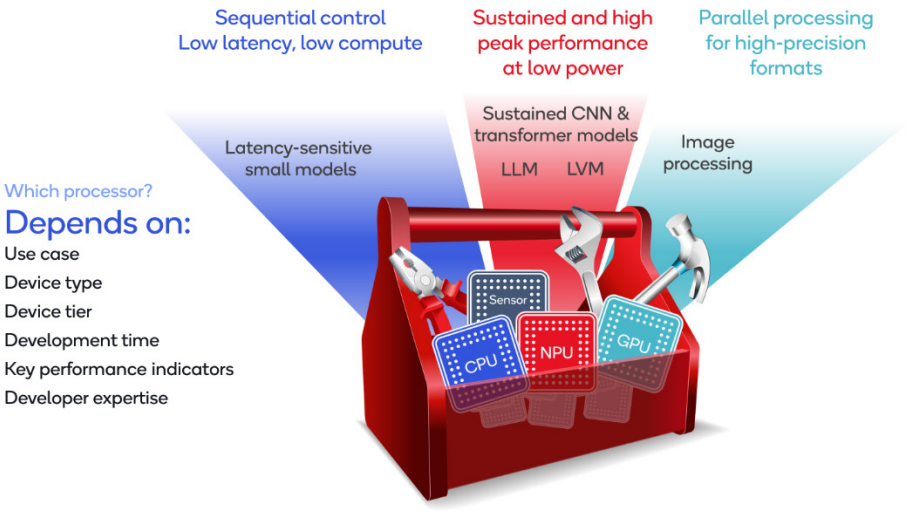

Figure 5: Choosing the right processor, like choosing the right tool in a toolbox, depends on many factors.

As previously mentioned, most generative AI use cases can be categorized into on-demand, sustained, or pervasive. For on-demand applications, latency is the KPI since users do not want to wait. When these applications use small models, the CPU is usually the right choice. When models get bigger (e.g., billions of parameters), the GPU and NPU tend to be more appropriate. **For sustained and pervasive use cases, in which battery life is vital and power efficiency is the critical factor, the NPU is the best option.**

Another key distinction is identifying whether the AI model is memory bound — performance is limited by memory bandwidth — or compute bound — performance is limited by the speed of the processor. Today's LLMs are memory bound for the text generation, so focusing on memory efficiency on the CPU, GPU, or NPU is appropriate. For LVMs, which could be compute or memory bound, the GPU or NPU could be used, but the NPU provides the best performance per watt.

A personal assistant that offers a natural voice user interface (UI) to improve productivity and enhance user experiences is expected to be a popular generative AI application. The speech recognition, LLM, and speech models must all run with some concurrency, so it is desirable to split the models between the NPU, GPU, CPU, and the sensor processor. For PCs, agents are expected to run pervasively (always-on), so as much of it as possible should run on the NPU for performance and power efficiency.

7 Qualcomm AI Engine: Best-in-class heterogeneous computing for generative AI

The Qualcomm AI Engine, which is comprised of several hardware and software components, accelerates on-device AI on Snapdragon and Qualcomm platforms. In terms of integrated hardware, the Qualcomm AI Engine has a best-in-class heterogeneous computing architecture consisting of the Hexagon NPU, Adreno GPU, Kryo or Qualcomm Oryon CPU, Qualcomm Sensing Hub, and memory subsystem — all engineered to work together and run AI applications quickly and efficiently on device.

Figure 6: The Qualcomm AI Engine consists of the Hexagon NPU, Adreno GPU, Qualcomm Kryo or Qualcomm Oryon CPU, Qualcomm Sensing Hub, and memory subsystem.

7.1 The processors of the Qualcomm AI Engine

Our latest Hexagon NPU offers significant improvements for generative AI, delivering 98% faster performance and 40% improved performance per watt. It includes micro-architecture upgrades, enhanced micro-tile inferencing, reduced memory bandwidth, and a dedicated power rail for optimal performance and efficiency. These enhancements, along with INT4 hardware acceleration, make the Hexagon NPU the leading processor for on-device AI inferencing.

The Adreno GPU, besides being the powerhouse engine behind high-performance graphics and rich user experiences with low power consumption, is designed for parallel processing AI in high-precision formats, supporting 32-bit floating point (FP32), 16-bit floating point (FP16), and 8-bit integer (INT8). The upgraded Adreno GPU in Snapdragon 8 Gen 3 yields 25% improved GPU power efficiency, enhanced AI, gaming, and streaming. Llama 2-7B can generate more than 13 tokens per second on the Adreno GPU.

As mentioned in the prior section, CPUs can perform well for low-compute AI workloads that require low latency. On Snapdragon® X Elite Compute Platform, the Qualcomm Oryon CPU — the new CPU leader in PCs — delivers up to 2X faster CPU performance versus the competition, matching competitor peak performance with one-third of the power.

It is essential to also have an always-on processor to handle contextualization for pervasive generative AI applications. The integrated Qualcomm Sensing Hub is an extremely efficient always-on AI processor for smaller neural networks and pervasive applications, such as contextual awareness and sensor processing, that need to run 24/7 — often at less than 1 milliamp (mA) of current. The newly updated Qualcomm Sensing Hub in Snapdragon 8 Gen 3 is 3.5X better than its predecessor, has a 30% increased memory, and features two next-generation micro NPUs for enhanced AI performance. The Qualcomm Sensing Hub has a dedicated power rail, which allows it to run while the rest of the SoC is off, saving significant power.

All the processors complement each other, and together they enable much more efficient AI processing.

7.2 Our system approach to AI heterogenous computing

Applying a system approach to our heterogeneous computing solution is essential since heterogeneous computing encompasses the entire SoC, which has three layers — the diverse processors, the system architecture, and the software. The holistic view allows our architects to evaluate key constraints, requirements, and dependencies between each of these layers and then make the most appropriate choices for the SoC and end-product usage, such as designing the shared memory subsystem or deciding what datatypes each processor should support. Since we custom design the entire system, we can make the appropriate design tradeoffs and use that insight to deliver a more synergistic solution.

This custom-design approach differentiates our solution and allows us to insert new AI instructions or hardware accelerators into each processor. We strive to maintain processor diversity while evolving the architecture for heterogeneous computing features since it is the diversity that provides the benefits. If all the processors have similar architectures, then the SoC becomes a homogeneous system. In contrast, most other chipset vendors typically license several third-party processors and then piece them together. These processors may not necessarily fit together well and were not necessarily designed for the same constraints or market segment.

The Qualcomm AI Engine, featured in our Snapdragon platforms and many of our other products, is at the core of our on-device AI advantage. **The result of many years of full-stack AI optimization, the Qualcomm AI Engine provides best-in-class on-device AI performance at extremely low power to support use cases today and in the future.** We have shipped more than 2 billion products featuring the Qualcomm AI Engine — powering an unmatched range of device categories including smartphones, XR, tablets, PCs, security cameras, robots, vehicles, and more.[2]

7.3 Case study: Avatar AI personal assistant using heterogeneous computing

At Snapdragon Summit 2023, we demonstrated a voice-powered AI personal assistant with a real-time animated on-screen avatar on a smartphone powered by Snapdragon 8 Gen 3. This application had many complex workloads with varied computing requirements running at the same time. **Utilizing the diversity of processors in the SoC and running the appropriate workload on the most suitable processor was key to achieving a good user experience.**

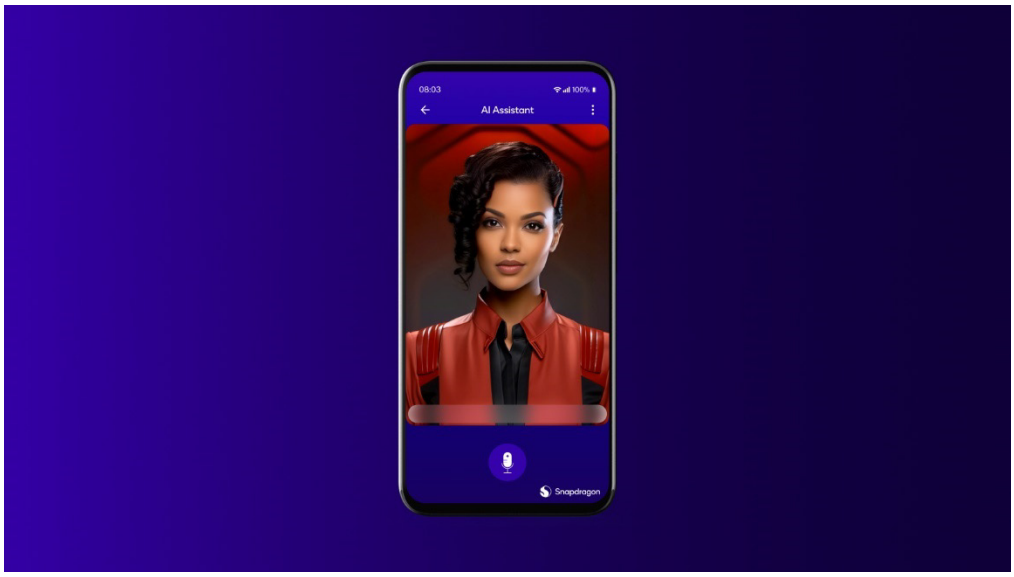

Figure 7: The avatar AI assistant includes many complex workloads.

Let's examine how the workloads were distributed:

1. As the user speaks to the assistant, voice is converted to text through Whisper, OpenAI's generative AI model for ASR. This ran on the Qualcomm Sensing Hub.
2. The assistant then used Llama 2-7B as its LLM to generate a text response. This ran on the NPU.
3. The text is then converted to speech using an open-source TTS model running on the CPU.

[2] https://www.qualcomm.com/products/mobile/snapdragon/smartphones/mobile-ai

4. In parallel, the avatar must be rendered in sync with the speech for a compelling user interface. The audio is used to create blendshapes so that the mouth and facial expressions properly animate. This traditional AI workload runs on the NPU.
5. The final avatar rendering is run on the GPU. For all these steps, data is being efficiently transferred throughout the memory subsystem and kept on chip as much as possible.

This personal assistant demo utilized all the diverse processors of the Qualcomm AI Engine to efficiently process the generative and traditional AI workloads.

Figure 8: The personal assistant with an avatar utilizes all the diverse processors of the Qualcomm AI Engine.

8 Leading AI performance on Snapdragon

Great hardware and software are required for leading performance. While tera operations per second (TOPS) is indicative of hardware potential, it is software that enables accessibility to and overall utilization of the hardware. AI benchmarks offer a better indication of performance, but the ultimate measure is in real applications with measurements of peak performance, sustained performance, and performance per watt. While generative AI benchmarks and applications are still in their infancy, below is an analysis of competitive AI metrics available today that show leading performance on Snapdragon Platforms.

8.1 Leading AI performance on smartphones with Snapdragon 8 Gen 3

On MLCommon's MLPerf Inference Mobile V3.1 benchmark, Snapdragon 8 Gen 3 has leading performance compared to other smartphone competitors. For example, on MobileBert, which is a generative AI language understanding model, Snapdragon 8 Gen 3 is 17% better than competitor A and 321% better than competitor B.[3] On Ludashi's AIMark V4.3 benchmark, Snapdragon 8 Gen 3 is 5.7X and 7.9X the overall score of competitor B and

[3] Qualcomm Technologies ran and collected data on phones powered by Snapdragon and competitor B. Competitor A scores are self-reported.

competitor C, respectively. On AnTuTu's AITuTu benchmark, Snapdragon 8 Gen 3 is 6.3X the overall score of competitor B.

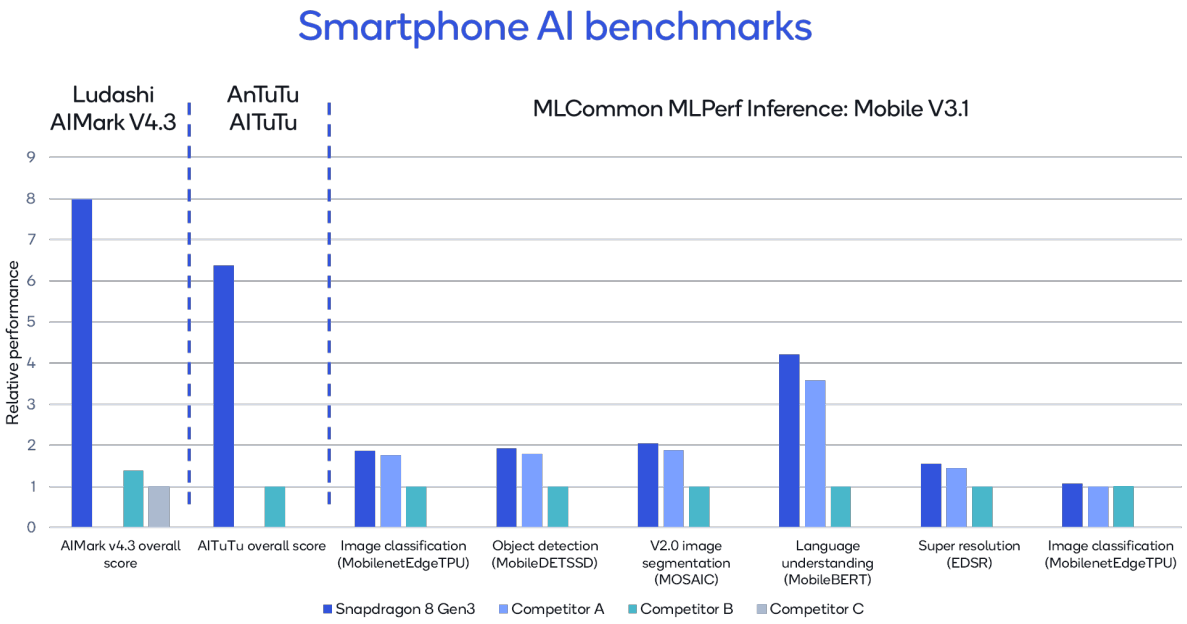

Figure 9: Snapdragon 8 Gen 3 has leading smartphone AI performance on AIMark, AITuTu, and MLPerf.

At Snapdragon Summit, we demonstrated two generative AI applications that are indicative of real-world performance for the common architectures of LLMs and LVMs. **On Snapdragon 8 Gen 3, our personal assistant demo ran Llama 2-7B at up to 20 tokens per second. Fast Stable Diffusion generated a 512x512 image in less than 0.6 seconds without losing much accuracy.[4] Our Llama and Stable Diffusion metrics are leading in smartphones.**

8.2 Leading AI performance on PCs with Snapdragon X Elite

On Snapdragon X Elite, the 45 TOPS of the Hexagon NPU is significantly higher than the number of TOPS on the NPUs of competitors' latest X86 chips. **On UL's Procyon AI Benchmark for Windows, Snapdragon X Elite has leading performance compared to other PC competitors.** For example, the overall score for the benchmark on Snapdragon X Elite is 3.4X and 8.6X the overall score of X86 competitor A and B, respectively.

[4] CLIP (Contrastive Language-Image Pre-training) score, which is a measurement of accuracy, is close to the baseline model.

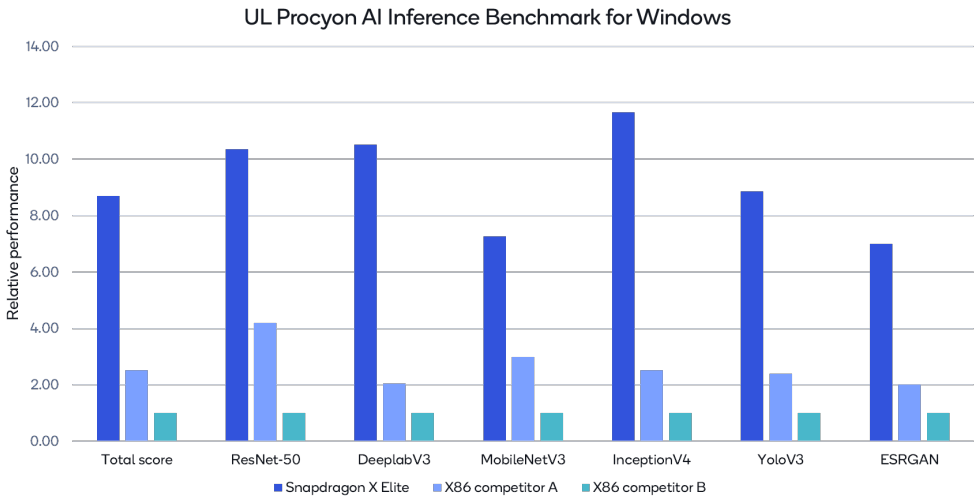

Figure 10: Snapdragon X Elite has leading laptop AI performance on the Procyon benchmark.

On Snapdragon X Elite, Llama 2-7B ran at up to 30 tokens per second on the Qualcomm Oryon CPU. Fast Stable Diffusion generated a 512x512 image in less than 0.9 seconds without losing much accuracy. Our Llama and Stable Diffusion metrics are leading in laptops.

9 Accessing our AI processors through our software stack

It is not enough just to have great AI hardware. Making AI acceleration via heterogeneous computing accessible to developers is essential for on-device AI to scale. The Qualcomm AI Stack unifies our complementary AI software offerings into a single package. OEMs and developers can create, optimize, and deploy their AI applications on our products and utilize the Qualcomm AI Engine performance — with the aim to allow developers to create AI models once and deploy them everywhere across different products.

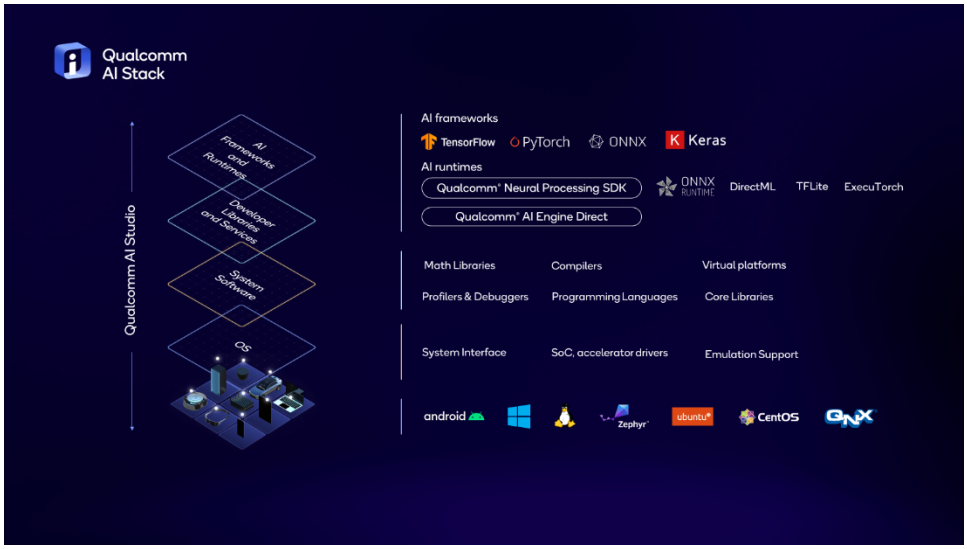

Figure 11: The Qualcomm AI Stack aims to help developers write once and run everywhere, achieving scale.

The Qualcomm AI Stack, from top to bottom, supports popular AI frameworks — such as TensorFlow, PyTorch, ONNX, and Keras — and runtimes — such as TensorFlow Lite,

TensorFlow Lite Micro, ExecuTorch, and ONNX runtime — as well as delegators for these runtimes that can be directly coupled with the Qualcomm AI Engine direct SDK (software development kit) for faster development.

Additionally, the AI Stack includes our Qualcomm® Neural Processing SDK for inferencing with versions for Android, Linux, and Windows. Our developer libraries and services support the latest programming languages, virtual platforms, and compilers.

At a lower level of the software stack, our system software includes the basic real-time operating system (RTOS), system interfaces, and drivers. Spanning across different product lines, we also have a rich variety of OS support, including Android, Windows, Linux, and QNX, and deployment and monitoring infrastructure like Prometheus, Kubernetes, and Docker.

For direct cross-platform access to the GPU, OpenCL and DirectML are supported. For the CPU, which is often the first place to start for AI programming due to its ease of programmability and presence in all platforms, our LLVM compiler infrastructure optimizations enable accelerated and efficient AI inference.

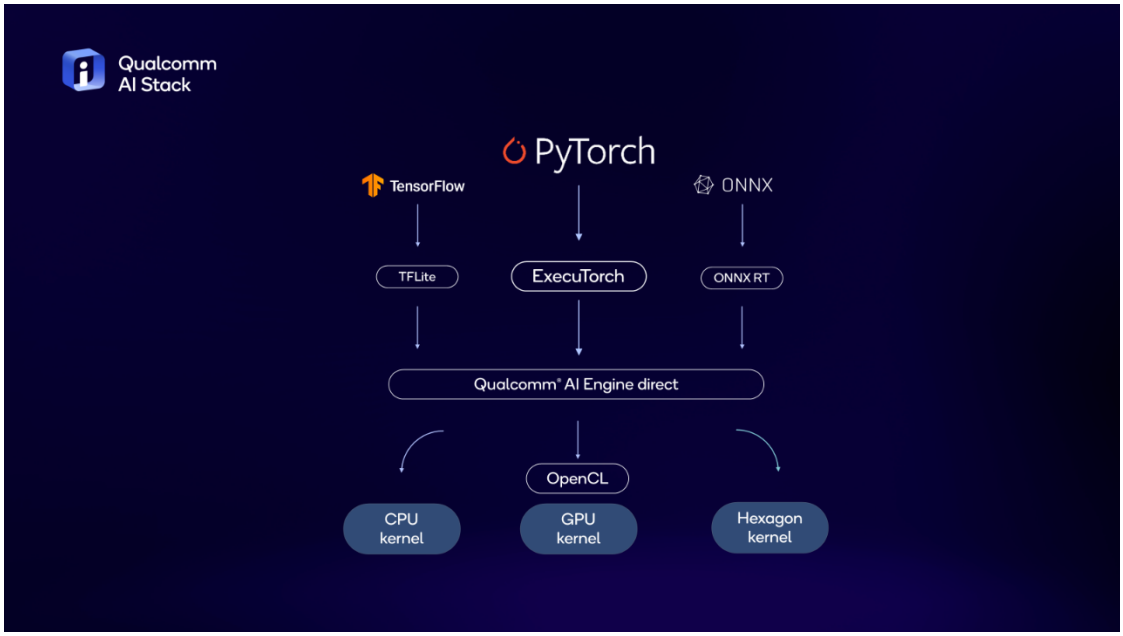

Figure 12: The Qualcomm AI stack supports key frameworks and runtimes.

We are focused on AI model optimization for improved power efficiency and performance. A small and fast AI model is not useful if it provides low-quality or inaccurate results. We take a holistic and principled approach — across quantization, compression, conditional compute, neural architecture search (NAS), and compilation — to shrink AI models and run them efficiently without sacrificing much accuracy, even those models that have already been optimized for mobile devices by the industry.

For example, quantization is beneficial for performance, power efficiency, memory bandwidth, and storage. The Hexagon NPU natively supports INT4, and our AI Model Efficiency Toolkit (AIMET)[5] provides quantization tools developed from techniques created by Qualcomm AI Research to limit accuracy loss while reducing bit precision. For generative AI, transformer-

[5] AIMET is a product of Qualcomm Innovation Center, Inc.

based LLMs — such as GPT, Bloom, and Llama — tend to benefit greatly from the jump in efficiency when quantized to 8-bit or 4-bit weights, as they are memory bound.

With quantization-aware training and/or further quantization research, many generative AI models can be quantized to INT4. In fact, INT4 has been the trend and is becoming the norm for LLMs, especially across the open-source community and with the desire to run large-parameter models on edge devices. **Support for INT4 allows for even higher power savings without compromising accuracy or performance — delivering up to 90% better performance and 60% better performance per watt compared to INT8 for running more efficient neural networks. Low-bit integer precision is essential for power-efficient inference.**

10 Conclusion

Heterogeneous computing with diverse processors is essential for maximizing performance and power efficiency in generative AI applications. **The Hexagon NPU, specifically designed for sustained, high-performance AI inference, offers superior performance, power efficiency, and area efficiency compared to the competition. The Qualcomm AI Engine, comprised of the Hexagon NPU, Adreno GPU, Qualcomm Kryo or Qualcomm Oryon CPU, Qualcomm Sensing Hub, and memory subsystem, provides a best-in-class heterogeneous computing solution for generative AI across on-demand, sustained, and pervasive use cases.**

By custom designing the entire system, we make the appropriate design tradeoffs and use that insight to deliver a more synergistic solution. Our iterative improvement and feedback cycle enables continual and quick enhancements of not only our NPU but also the AI stack based on the latest neural network architectures. Our leading performance in AI benchmarks and generative AI applications for smartphones and PCs is a result of our differentiated solution and full-stack AI optimizations.

The Qualcomm AI Stack enables developers to create, optimize, and deploy AI applications across different products, making AI acceleration on the Qualcomm AI Engine accessible and scalable. **The combination of technology leadership, custom silicon designs, full-stack AI optimization, and ecosystem enablement sets Qualcomm Technologies apart to drive the development and adoption of on-device generative AI.**

For more information, visit us at:
qualcomm.com