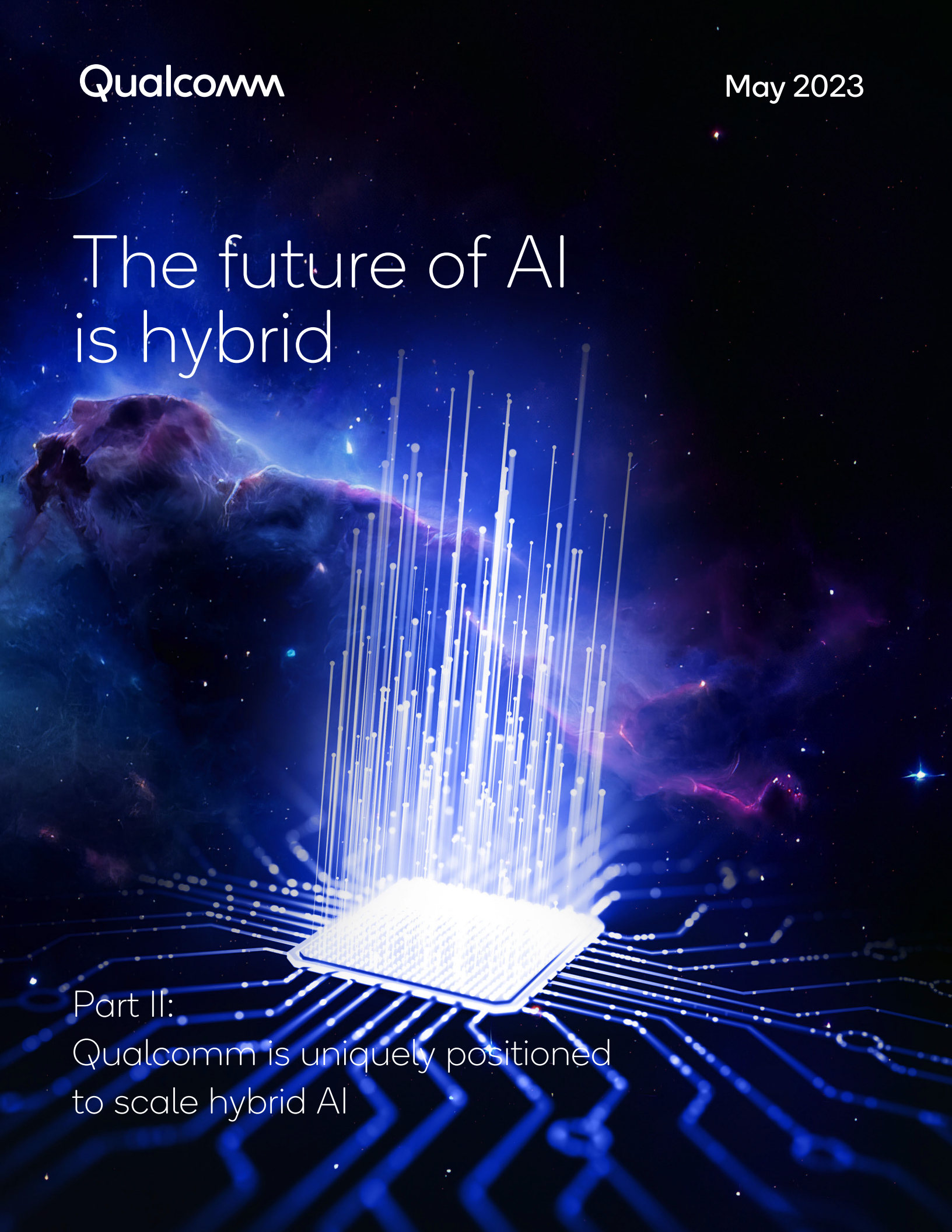
Qualcomm

May 2023

The future of AI is hybrid



Part II: Qualcomm is uniquely positioned to scale hybrid AI

Table of contents

1 Executive summary 3
2 Qualcomm Technologies is the leader in on-device AI 3
2.1 Sustained innovation 4
2.1.1 Our AI history 4
3 Our leadership in on-device generative AI 4
3.1 Pushing the boundaries of on-device and hybrid AI 5
3.2 Responsible AI 5
4 Superior on-device AI technology with full-stack optimization 6
4.1 Algorithms and model development 7
4.2 Software and model efficiency 7
4.2.1 Quantization 8
4.2.2 Compilation 9
4.3 Hardware acceleration 9
5 Unmatched global footprint and scale at the edge 11
5.1 Handsets 11
5.2 Automotive 12
5.3 PC and tablets 12
5.4 IoT 12
5.5 XR 12
6 Conclusion 12

1 Executive summary

As we previously established, the future of AI needs to be hybrid, with AI processing distributed between the cloud and devices. A hybrid AI architecture, or running AI on device alone, offers benefits with regards to cost, energy, performance, privacy, security, and personalization – at a global scale.

Qualcomm is enabling intelligent computing everywhere. As the on-device AI leader, Qualcomm Technologies is uniquely positioned to scale hybrid AI with industry-leading hardware and software solutions for edge devices, spanning across billions of phones, vehicles, XR headsets and glasses, PCs, IoT, and more. Our hardware offers industry-leading performance per watt, exemplified by our mobile solutions being approximately 2X higher than the competition. Our perpetual flywheel of innovation — due to our fundamental research and our full-stack on-device AI optimization across AI applications, models, hardware, and software — keeps us at the forefront of on-device AI solutions.

Qualcomm Technologies also enables 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. With our technology leadership, global scale, and ecosystem enablement, Qualcomm Technologies is making hybrid AI a reality.

2 Qualcomm Technologies is the leader in on-device AI

Through our leadership in on-device AI which powers billions of edge devices, Qualcomm Technologies is enabling this new era of hybrid AI. Our scalable technology architecture allows us to utilize a single AI stack that is highly optimized and works across not only different end devices but also different models. Our AI solutions are designed to provide the highest performance per watt and make AI ubiquitous.

The Qualcomm® AI Engine, featured in our Snapdragon platforms and many of our other products, is at the core of this 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.[1]

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 fully leverage the Qualcomm AI Engine performance – with the aim to allow AI developers to create AI models once and deploy them everywhere across different products.

---

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

2.1 Sustained innovation

Our development of low-power, high-performance AI has led to a substantial device AI ecosystem spanning across established and emerging segments for smartphones, vehicles, XR, PCs and laptops, and enterprise AI. We have been utilizing AI for many years to support differentiation and command a premium for our chipset products, with key use cases around image and video capture, advanced connectivity, voice command, security, and privacy.

2.1.1 Our AI history

Qualcomm has been investing in AI research and development for more than 15 years. At Qualcomm AI Research[2], our mission is to create breakthroughs in fundamental AI research and scale them across industries and use cases. We are advancing AI to make its core capabilities – perception, reasoning, and action – ubiquitous across devices. Our notable AI research papers are influencing the industry and advancing power-efficient AI. By bringing together some of the best minds in the field, we are pushing the boundaries of what's possible and shaping the future of AI.

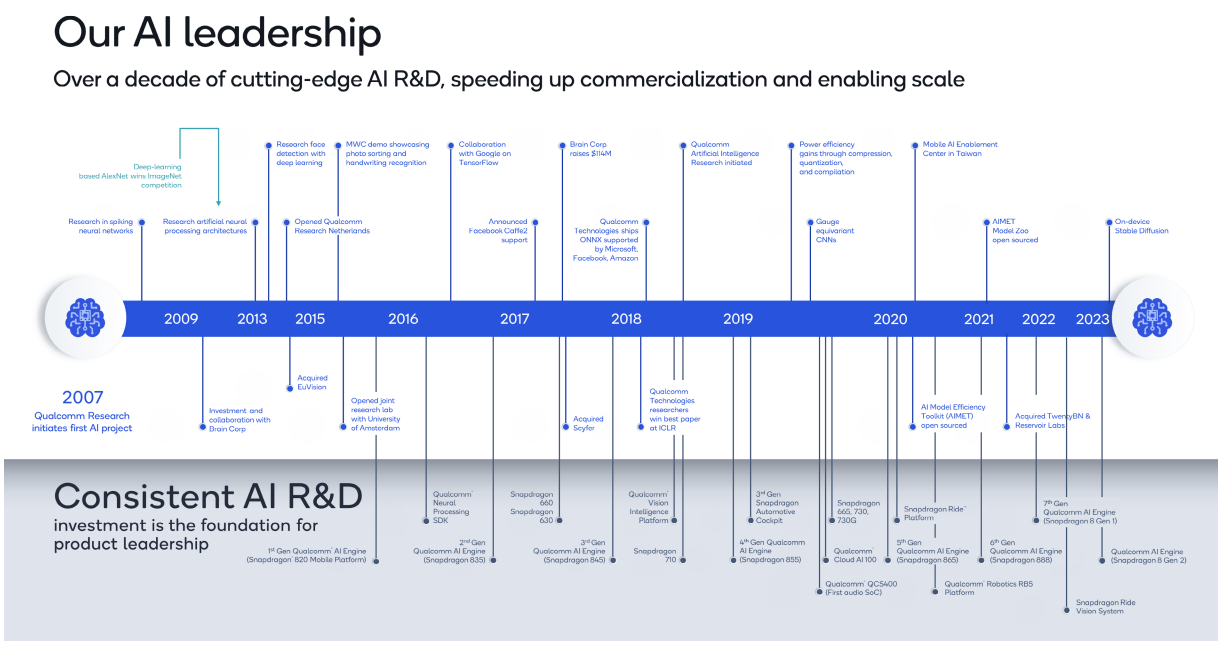


Figure 1: Our consistent AI R&D investment is the foundation for product leadership.

3 Our leadership in on-device generative AI

Our AI research team has explored generative AI for several years. Generative AI dates back to generative adversarial networks (GANs) and variational auto encoders (VAEs). We initially explored whether generative models could compress well and generate improved perceptual-pleasing artifacts. We used VAE technology to create better video and speech codecs, keeping the model size small at less than 100 million parameters. We also extended the generative AI ideas to wireless to replace the channel model to be more effective in communication systems.

[2] Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

More recently, we increased the scale of on-device generative AI models to more than 1 billion parameters, such as Stable Diffusion, and have plans for models with tens of billions of parameters in the future. We are also researching how to use generative AI models as universal agents to structure computations and use language to represent tasks and actions. We are studying how we can further exploit this capability by adding perceptual input, such as visual and audio, and the ability to interact with an environment, for example to generate a command for a robot or run software.

3.1 Pushing the boundaries of on-device and hybrid AI

Qualcomm Technologies is uniquely capable to provide the processing performance needed in edge devices to run generative AI, such as large language models (LLMs), at low power. For generative AI to be adopted broadly, inferencing cannot only be run in the cloud as it is now – there must be significant AI processing on devices. Ultimately, AI processing needs to use both the cloud and the device for generative AI to become a part of everyday life. Down the line, we see users choosing their next phone, PC, or vehicle upgrade largely based on AI capabilities.

We already lead in on-device AI inference with regards to AI hardware acceleration and software solutions to ease development, like the Qualcomm AI Stack. We can run models with over 1 billion parameters on device today and anticipate this growing to over 10 billion parameters in the coming months.

Our AI acceleration architecture is flexible and robust for potential changes in generative AI model architectures. As LLMs and other generative AI models continue to evolve, our AI stack and technology can advance accordingly. Being able to easily develop hybrid AI apps is key, and our common AI architecture across our portfolio and AI tools are designed for this future.

3.2 Responsible AI

We strive to create AI technologies that bring positive change to society. Our vision for on-device AI is based on transparency, accountability, fairness, managing environmental impact, and being human-centric. We aim to act as a responsible steward of AI, considering the broader implications of our work and taking steps to mitigate any potential harm. Our on-device AI solutions are designed to enable enhanced privacy and security, essential to a robust and trustworthy AI ecosystem.

Qualcomm monitors and engages governments globally on regulatory frameworks, guidelines, and best practices, including intergovernmental guidance, like the Organisation for Economic Co-operation and Development's Recommendation on AI, and regional and national frameworks like the European Union's AI Act and the U.S. National Institute of Standards and Technology AI Risk Management Framework. These regulations and guidelines provide important legal and ethical considerations for the responsible development and deployment of AI technologies. Compliance with AI regulations and best practices is a fundamental aspect of our commitment to ethical and responsible AI innovation, and we will continuously align our practices with the evolving landscape of AI governance.

Finally, as part of our participation and leadership in industry collaborations, standard body organizations and consortia, we contribute and advocate for AI standards, data and privacy protections, robust cybersecurity. Qualcomm has long recognized the importance of having

robust and comprehensive standards to guide the responsible development and deployment of new technologies.

Working collaboratively to develop robust and effective AI standards is a critical step toward building a sustainable and trustworthy AI ecosystem.

4 Superior on-device AI technology with full-stack optimization

We do full-stack AI research and optimization across the application, neural network model, algorithms, software, and hardware. A heterogeneous computing approach takes advantage of hardware – such as CPU, GPU, and AI accelerators – and software – such as Qualcomm AI Stack – to accelerate on-device AI. We have teams working jointly and across all these disciplines to develop the most optimized solutions.

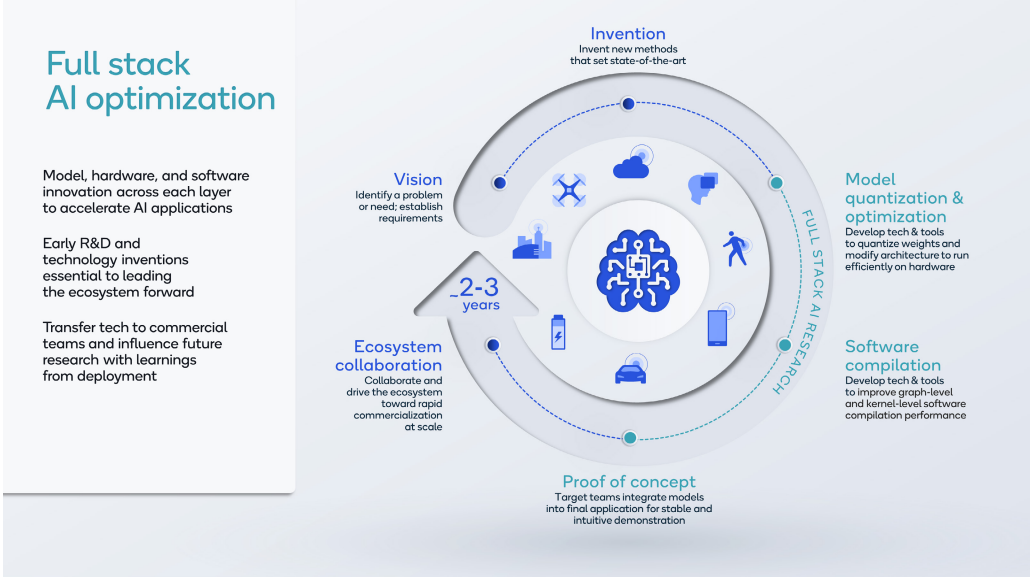


Figure 2: Our full-stack AI research and optimization enables continuous improvement and leading power-efficient solutions.

This flywheel of innovation allows us to continually improve our AI stack across hardware, software, and algorithms based on the latest neural network architectures. Our unique ability to do fundamental AI research supporting full stack on-device AI development enables fast time-to-market and optimized implementations around key applications, such as on-device generative AI.

Our world's first demonstration of Stable Diffusion on an Android smartphone highlights the benefits of our approach. All the full-stack research and optimization that went into making it possible to run Stable Diffusion completely on device in under 15 seconds is already flowing into the Qualcomm AI Stack and will improve future hardware designs. Furthermore, optimizations allowing Stable Diffusion to run efficiently on phones can also be used for other platforms such as laptops, XR devices, and virtually any other device powered by out technologies.

4.1 Algorithms and model development

Our research team develops and modifies neural network architectures for improved efficiency without sacrificing accuracy. Action recognition and super resolution are examples.

Traditional deep-learning models designed for action recognition process video sequences frame by frame, layer by layer. While this leads to accurate results, it is compute intensive, high latency, and power inefficient. Our FrameExit model, which is publicly available, automatically learns to process fewer frames for simpler videos and more frames for complex ones, saving power and improving performance. Beyond our model architecture innovation, our full-stack AI optimizations included state-of-the-art quantization techniques and a novel compiler stack. We demonstrated this on a mobile device and achieved an up to 5-time reduction in compute and latency (on average) when compared to other methods on commonly used action recognition benchmarks.

Super resolution clarifies, sharpens, and upscales an image to higher resolution for applications like gaming and video playback on high-resolution screens. Although AI-based super resolution achieves impressive visual quality compared to traditional approaches, enabling it in real time on mobile devices is challenging. We optimized across the full AI stack, including the algorithm with our Q-SRNet model, the software with 4-bit integer (INT4) quantization, and the Snapdragon 8 Gen 2 hardware with INT4 acceleration. We achieved the world's first on-device demo of real-time super resolution using an INT4 model, which dramatically improved latency and power consumption. In fact, compared to INT8, INT4 performance and power efficiency improve by 1.5 to 2 times.

4.2 Software and model efficiency

The Qualcomm AI Stack is designed to help developers write once and run AI loads everywhere across our hardware. The Qualcomm AI Stack, from top to bottom, supports popular AI frameworks such as TensorFlow, PyTorch, ONNX, and Keras, and runtimes including TensorFlow Lite, TensorFlow Lite Micro, ONNX runtime, and more. Additionally, it includes inferencing software development kits (SDKs) like our popular Qualcomm® Neural Processing SDK 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, 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.

The Qualcomm AI Stack also includes Qualcomm® AI Studio, supporting a complete model workflow from design to optimization, deployment, and profiling. It brings together all the tools that we offer into a graphical user interface along with visualization tools to simplify the developer experience, enabling them to see their model development in action including AI Model Efficiency Toolkit (AIMET), AIMET Model Zoo, model analyzers, and neural architecture search (NAS).[3]

[3] AIMET and AIMET Model Zoo are products of Qualcomm Innovation Center, Inc.

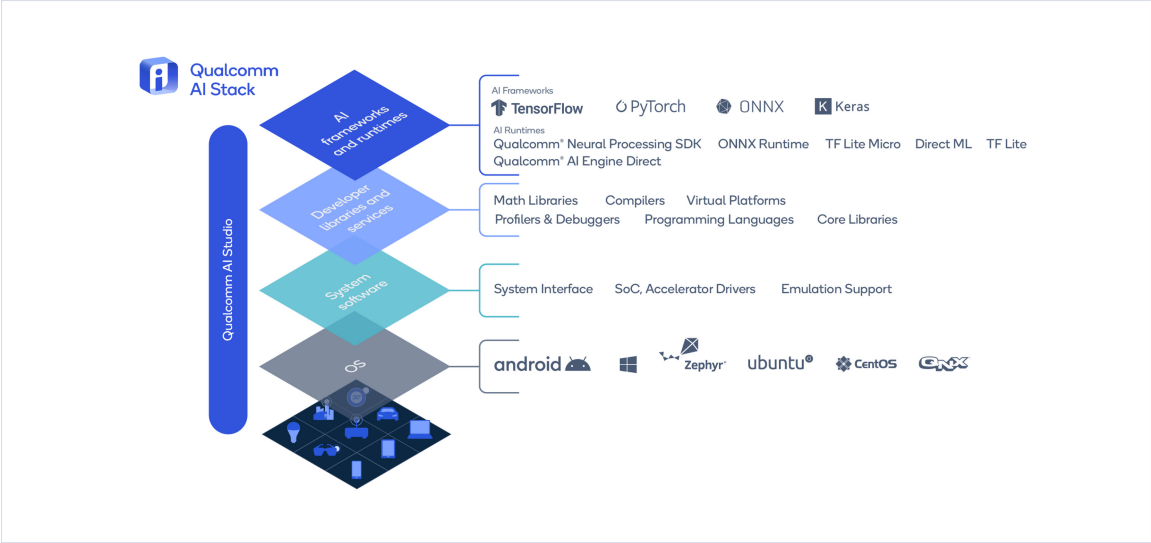


Figure 3: The Qualcomm AI Stack aims to help developers write once and run everywhere, achieving scale.

We are focused on AI model efficiency research for improved power efficiency and performance. A small and fast AI model is not useful if it provides low-quality or inaccurate results. So, 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 that have already been optimized for mobile devices by the industry.

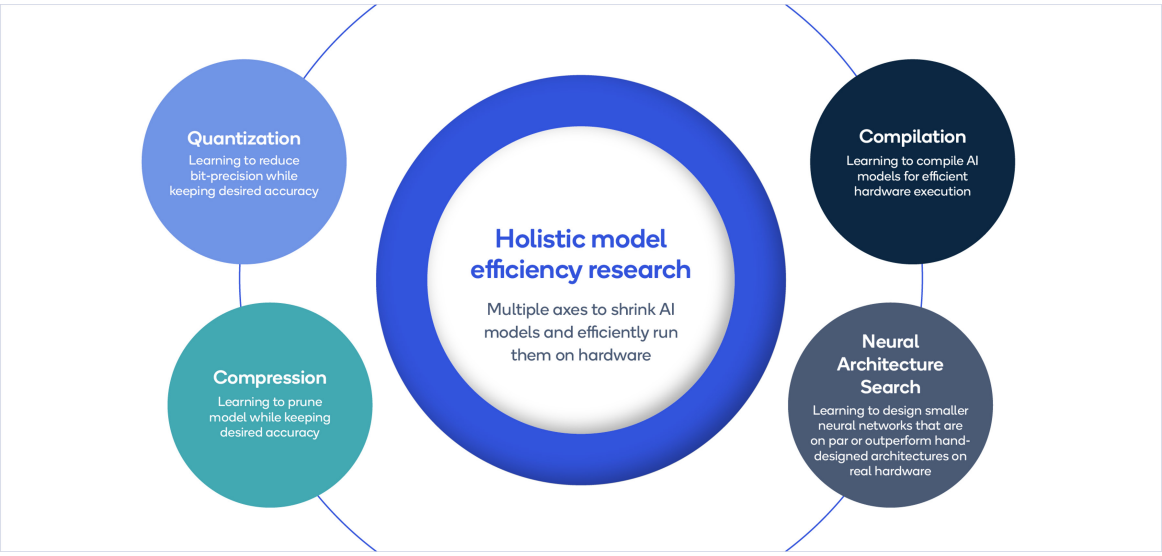


Figure 4: Qualcomm AI Research takes a holistic approach to AI model efficiency research.

4.2.1 Quantization

Quantization for efficient integer inference is a key focus area. Over the past several years, we have shared our leading AI research on quantization, including post-training quantization (PTQ) techniques like Data Free Quantization and AdaRound, and joint quantization and pruning

techniques, like Bayesian Bits, through papers and demonstrations. Quantization not only increases performance and decreases memory storage requirements, it also saves power by allowing the model to efficiently run on our dedicated AI hardware and consume less memory bandwidth. For example, quantizing from FP32 to INT4 can increase performance-per-watt savings by up to 64X in memory and compute.

With respect to generative AI, transformer-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-bounded. Several research works, ours included, have shown that 4-bit weight quantization is not only possible for LLMs, but is also optimal and possible to do in the PTQ setting. This efficiency boost surpasses what is possible with floating-point.

Our AIMET provides quantization tools developed from techniques created by Qualcomm AI Research and now incorporated into the Qualcomm AI Studio. With quantization aware training and/or further quantization research, many generative AI models can be quantized to INT4. 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.

4.2.2 Compilation

Compilers are a key component of the AI stack for efficiently running AI models at the highest performance and lowest power. The AI compiler converts an input neural network into code that runs on target hardware, while optimizing for latency, performance, and power. Compilation consists of tiling, placement, sequencing, and scheduling steps of a computation graph. Our expertise in traditional compiler techniques, polyhedral AI compilers, and AI research in combinatorial optimization for compilers has led to state-of-the-art results.

For example, the Qualcomm AI Engine direct framework sequences the operations to improve performance and minimize memory spillage based on the hardware architecture and memory hierarchy of the Qualcomm® Hexagon™ Processor. Our optimizations help reduce DRAM traffic, significantly reducing runtime latency and power consumption.

4.3 Hardware acceleration

Our hardware offers industry-leading performance per watt – approximately 2X higher than mobile competition.

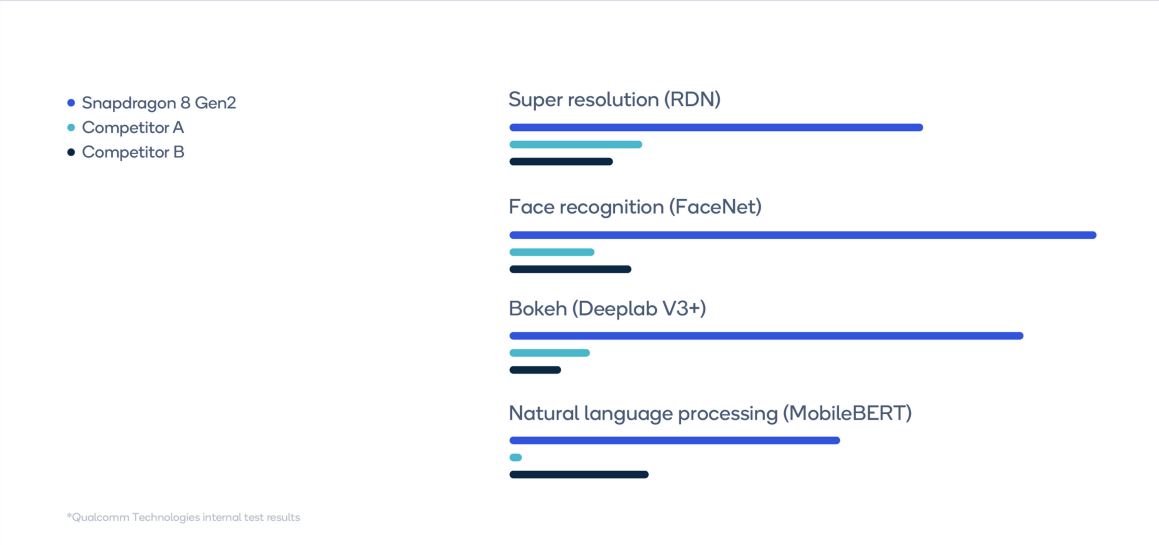


Figure 5: Snapdragon 8 Gen 2 provides leading AI performance per watt compared to mobile competitors.

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 hardware, the Qualcomm AI Engine has a heterogeneous computing architecture consisting of the Hexagon Processor, Qualcomm® Adreno™ GPU, and Qualcomm® Kryo™ CPU – all engineered to run AI applications quickly and efficiently on device. Through this heterogeneous computing approach, developers and OEMs can optimize AI user experiences on smartphones and other edge devices.

With years of research dedicated to its advancement, the Hexagon Processor is the most critical piece of the Qualcomm AI Engine. It has evolved to address the changing requirements of AI. In 2007, the first Hexagon Processor was launched on Snapdragon. In 2015, the Snapdragon 820 processor was announced and included our first dedicated mobile Qualcomm AI Engine to support imaging, audio, and sensor operations. We added a tensor accelerator to the Hexagon Processor 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 hub.

In 2022, Snapdragon 8 Gen 2 provided groundbreaking AI, integrated across the entire system, powered by our fastest, most advanced Qualcomm AI Engine to date. Users can experience faster natural language processing with multi-language translation or have fun with AI cinematic video capture. The latest Hexagon Processor introduced a dedicated power delivery system, which adapts power according to the workload. Special hardware improves group convolution, activation function acceleration, and the performance of the Hexagon Tensor Accelerator. Support for micro-tile inferencing and INT4 hardware acceleration provide even higher performance while reducing power and memory traffic. The transformer acceleration dramatically speeds up inference for multi-head attention that is used throughout generative AI, resulting in a staggering AI performance up to 4.35X on certain use cases with MobileBERT.

5 Unmatched global footprint and scale at the edge

Qualcomm Technologies has an immense footprint at the edge with an installed base of billions of user devices powered by Snapdragon and Qualcomm platforms – and many hundreds of millions of devices powered by our platforms enter the market each year.[4]

Our AI capabilities span a wide range of products – including mobile, vehicles, XR, PC, and IoT. We develop AI acceleration solutions, like the Qualcomm AI Engine, along with all other key IP innovations and technologies for the premium tier, usually on a yearly cadence as part of our scalable technology architecture, and quickly cascade capabilities across segments and down to mainstream and entry-level tiers.

As such, Qualcomm Technologies is uniquely positioned to enable hybrid AI to scale globally.

Figure 6: Devices powered by Snapdragon platforms can scale hybrid AI to billions of units across segments and tiers.

5.1 Handsets

Snapdragon is the leading mobile platform driving premium Android experiences – including more than two billion processors shipped with AI capabilities. Snapdragon platforms are also leading in AI benchmarks for mobile platforms – e.g., taking the top-20 positions in industry leading AI Benchmark.[5]

In the second quarter of 2023, TechInsights, a leading market research firm, forecasts Qualcomm Technologies to remain the leader in AI-capable smartphone processors shipments with more than 40% of unit share – far beyond others such as Apple (25%) and MediaTek (24%).[6]

[4] Counterpoint Research, May '23
[5] Based on ai-benchmark.com scores as of May '23
[6] TechInsights, Apr. '23

5.2 Automotive

Qualcomm Technologies is a leader in cockpit and in-vehicle infotainment solutions with all major global automakers selecting the Snapdragon® Cockpit Platforms to power their digital cockpit systems. Many of these automakers have already launched programs into production or are currently designing platforms with our solutions. These automakers include Honda, Mercedes, Renault, Volvo, JLR, Stellantis, BMW, GM/Cadillac, Great Wall Motor, Mahindra, Togg, Toyota, XPeng, GAC, Jetour, NIO, and WM Motor.

With the addition of the latest generation of the Snapdragon Cockpit Platform, our automotive solutions aim to provide best-in-class in-vehicle user experiences, as well as safety, comfort, and reliability, raising the bar for digital cockpit solutions in the connected car era.

Our Snapdragon Ride™ Platform offers an extended product roadmap, featuring the first-announced scalable and automated driving SoC platforms built on 5nm process technology, with an expanded software ecosystem with industry-proven stacks for vision perception, parking, and driver monitoring.

5.3 PC and tablets

Snapdragon Compute Platforms integrate the Qualcomm AI Engine for powerful on-device acceleration delivering better quality, performance, and efficiency of the latest applications. Beyond generative AI uses like text, image, and video creation, our AI Engine has traditional AI uses ranging from faster threat detection for improved security to eye contact and noise suppression for enhanced video conferencing. Leveraging the Hexagon Processor offers improved performance and efficiency for long battery life, while keeping other system resources like the CPU and GPU free to help users be more productive.

5.4 IoT

Qualcomm Technologies is a major technology provider for IoT, with more than 16,000 customers across verticals. The AI processing capabilities embedded in our IoT chipsets and platforms allow for on-device analysis of data, such as video, in efficient and actionable ways – driving innovation and transformation across multiple segments, including robotics, intelligent cameras, retail, and city infrastructure.

5.5 XR

XR devices, such as VR headsets and AR glasses, are infused with our on-device AI and Snapdragon Spaces™ Technology to provide more immersive experiences and better adapt to the surrounding world.

To date, more than 65 XR devices have launched using Snapdragon platforms – including many of the most popular devices from brands including Meta, Pico, and Lenovo.

6 Conclusion

Hybrid AI is inevitable. The cloud and devices will work together to deliver next-generation user experiences through powerful, efficient, and highly optimized AI capabilities. Our leadership in on-device AI uniquely positions us for the move toward a hybrid architecture – with many workloads moving from the cloud toward edge devices, thus requiring high performance along with superb power efficiency. Early investments in research and product development allow

Snapdragon platforms today to support generative AI models featuring more than 1 billion parameters – and support for 10 billion parameters or more is already in sight.

We have an unmatched footprint at the edge with an installed base of billions of devices worldwide powered by Snapdragon and Qualcomm platforms, enabling the opportunity to scale for generative AI and positively impact countless lives. Qualcomm Technologies is set to support developers, OEMs, and other ecosystem innovators in building new generative AI applications and solutions quickly and cost-effectively. The combination of technology leadership, global scale, and ecosystem enablement sets Qualcomm Technologies apart to drive the development and adoption of hybrid AI.

Qualcomm

Follow us on:
For more information, visit us at:
qualcomm.com