Qualcomm Technologies, Inc.

5775 Morehouse Drive, San Diego, CA 92121

www.qualcomm.com

Exploring the AI capabilities of the Snapdragon 888 Mobile Platform

By Jeff Gehlhaar, VP of Technology, Qualcomm Technologies & Hsin-I Hsu, Sr. Product Manager, Qualcomm Technologies

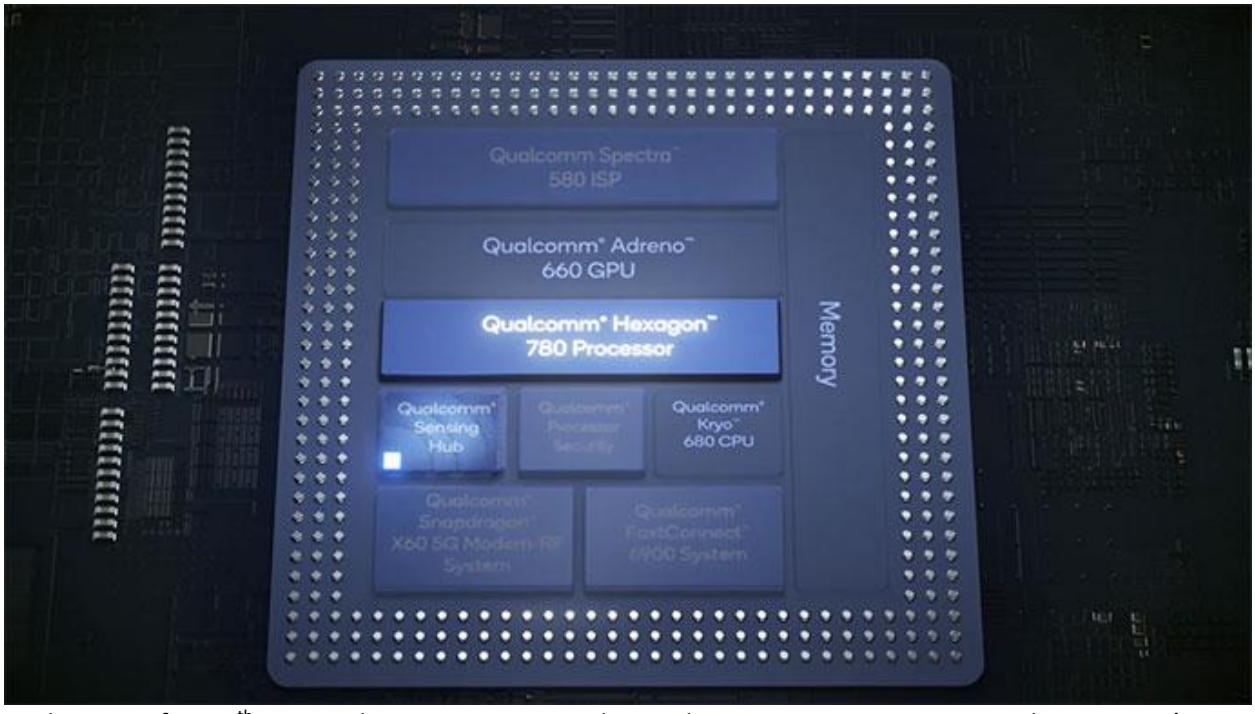
Day two of our annual Snapdragon Tech Summit Digital 2020 focused on a deep dive of the powerful Qualcomm Snapdragon 888 Mobile Platform — specifically, its groundbreaking AI capabilities.

At last year's Summit, we introduced the 5th gen Qualcomm AI Engine and real-time translation, a world's first, on the Snapdragon 865 5G Mobile Platform, with all AI processed on the device. We also demonstrated how our AI is inside your favorite social media apps, powering the cool and fun lens filters many of us enjoy using.

And we're helping transform industries beyond mobile. For example, our AI-enabled Qualcomm Robotics RB5 Platform is powering a robot ping-pong system that is helping train China's top players, and we continue to innovate with AI in autonomous driving.

AI is getting more complex, and it requires higher performance AI processing. That's why we're bringing you a more powerful, sophisticated AI on the Snapdragon 888 Mobile Platform, with the introduction of the 6th gen Qualcomm AI Engine.

6th gen Qualcomm AI Engine and the redesigned Qualcomm Hexagon processor



At the core of our 6th gen Qualcomm AI Engine is the Qualcomm Hexagon processor. This year, we're introducing the Hexagon 780 processor. It's completely redesigned and features our biggest leap in architecture and performance in years. We call it the *fused AI accelerator architecture*. In past generations, we used scalar, vector, and tensor accelerators. For this new generation, we're removing the physical distances between the accelerators and fusing them together, so it's now on one big AI accelerator. We've also added a dedicated large shared memory across the three different accelerators, so they can share and move data efficiently. The shared memory is 16x larger than its predecessor, and the result is an impressive hand-off time between the accelerators in the nanosecond range — up to 1000x faster in certain use cases.

We've also made improvements on the accelerators themselves. The scalar accelerator is 50 percent more powerful; the tensor accelerator is now two times faster versus that in the Snapdragon 865, and the Qualcomm Hexagon Vector eXtensions (HVX) now supports additional data types.

Other parts of the Qualcomm AI Engine are upgraded as well. Our Qualcomm Adreno 680 GPU now offers a 43 percent AI performance boost and includes new instruction sets like 4-input mixed precision dot product and wave matrix multiply for faster floating-point calculation.

Last year, we announced Snapdragon 865 at 15TOPS, one of the highest TOPS in the industry. This year, with Snapdragon 888, we're introducing the highest TOPS performance on mobile —a whopping 26 TOPS.

At Qualcomm Technologies, our expertise lays in delivering powerful performance at ultra-low power consumption, so while our processor is extremely powerful, we managed to do it with extreme efficiency. Our performance per watt on the Hexagon 780 processor is an impressive three-fold improvement versus the previous generation.

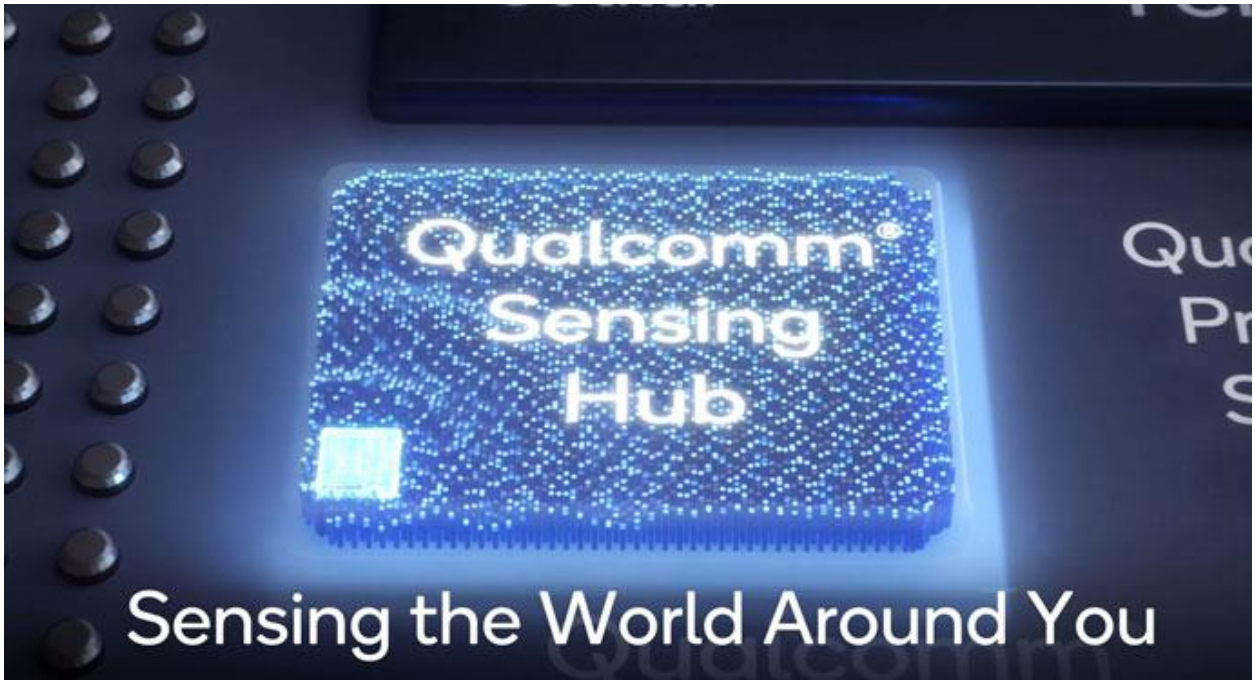
What can all this AI horsepower do? This year, we'll be demonstrating a brand new AI use case that fully utilizes the 6th gen Qualcomm AI Engine: Tetris.AI's super movie app.

For example, you'll be able to erase a character and put yourself inside a movie scene or a video that you recorded and interact with the other people/characters inside. You can even see this in real time in preview

mode even before you start acting and recording. The Qualcomm AI Engine is running and accelerating Tetras.AI's video instance segmentation and fusion algorithms, at 30 fps, up to 4K resolution.

2nd gen Qualcomm Sensing Hub

There are neural networks running continuously on your device 24/7, ready for action. These tiny, always-on AI algorithms require a completely different hardware and that's why we created the Qualcomm Sensing Hub.



With Snapdragon 888, we're introducing the 2nd generation Qualcomm Sensing Hub, and it's smarter than ever. First of all, we added a dedicated always-on, low-power AI processor and we're seeing a mind-blowing 5x AI performance improvement.

The extra AI processing power on the Qualcomm Sensing Hub allows us to offload up to 80 percent of the workload that usually goes to the Hexagon processor, so that we can save even more power. All the processing on the Qualcomm Sensing Hub is at less than 1 milliamps (mA) of power consumption.

In order to give developers easier access the Qualcomm Sensing Hub, we're working with Google and its TensorFlow Micro framework, so it can be optimized and accelerated on both the Hexagon processor and the AI Processor in the Qualcomm Sensing Hub.

Another new feature on the Qualcomm Sensing Hub is the ability to collect and decipher data from all different cores and create contextual awareness use cases. For the very first time, we're able to collect

connectivity data like 5G, Wi-Fi, Bluetooth, and location streams to create even more appealing use cases. The 2nd gen Qualcomm Sensing Hub will bring a suite of new always-on and contextually aware use cases to next year's smartphones. Our work with Audio Analytic, for example, can allow your phone to recognize the acoustic scene around you, enabling a range of intelligent new capabilities such as matching ring volume to your environment. And with trinamiX, your phone can analyze your skin conditions and recommend the right moisturizer, using a combination of on-device AI and cloud AI via our amazing 5G capabilities.

AI software

By now, you're aware of the incredible AI hardware capabilities at the heart of our new processor, but our hardware must be enabled by equally strong software. This is why we also completely ramped up our AI software.

Qualcomm Technologies was the first to commercialize an on-device AI SDK, the Qualcomm Neural Processing SDK. It's now powering incredible AI experiences in over 500 million Android devices worldwide. This year's improvements in the Qualcomm Neural Processing SDK include support for additional models and even expanded support for Windows 10 AI use cases on laptops powered by Snapdragon 888.

We created our ultra-fast Qualcomm AI Engine so that AI applications can take full advantage of the hardware acceleration. We introduced Hexagon NN Direct on Snapdragon 865 to give developers direct access to Hexagon from their applications. With the 6th gen Qualcomm AI Engine, we made a significant upgrade to this approach for Snapdragon 888 – bringing the power of direct APIs across the whole mobile platform.

We're introducing the Qualcomm AI Engine direct with our 6th gen Qualcomm AI Engine, moving to the next chapter of our AI software. With this solution, we extend and enhance the capabilities of our AI software solutions to provide developers with access directly to the hardware, and not only for the Hexagon 780 processor, but also for the Adreno GPU and Qualcomm Kryo CPU.

The Qualcomm AI Engine direct has been built from the ground up to, for the first time, bring a unified AI API across the whole Snapdragon platform. In addition, we're making this API backward compatible, which will support the previous 5th gen Qualcomm AI Engine, so developers and OEMs can take advantage of this solution across Snapdragon platforms and leverage both the 5th and 6th gen Qualcomm AI Engines. Not only that, but we're also focused on modularity and extensibility, expanding on our user-defined operator concept to bring new capabilities for developers to create their own AI solutions, accelerated on Snapdragon.

With the introduction of Snapdragon 888, we're collaborating with Hugging Face, a leader in innovative natural language processing NLP solutions. We're utilizing the power of the 6th gen Qualcomm AI Engine to enable and accelerate the robust NLP library, Hugging Face transformers, on Snapdragon, for precision and responsiveness. This enables your email client to give you autocomplete suggestions as you type, your AI

Qualcomm

voice assistant to better understand your questions, and language translation apps to work so much faster and be more accurate.

In 2019, as a part of our 5th gen Qualcomm AI Engine, we introduced the concept of user defined operators. This enabled developers to write custom operators in OpenCL or use the Qualcomm Hexagon SDK and then plug them into the Qualcomm Neural Processing SDK. However, even for developers that are already experienced with Hexagon, to create operators, you often needed to write complex and long routines in low-level languages.

As part of our commitment to enable access to the 6th gen Qualcomm AI Engine, this year, we announced that we've extended TVM, an open-source compiler for AI accelerators, with support for Hexagon. Now, custom operators can be written in a few short lines of Python, compiled for Hexagon, and plugged directly into the Qualcomm AI Engine direct framework.

Lastly, we added additional support to the AI Model Efficiency Toolkit (AIMET) for even better quantization of networks, with little or no loss in accuracy, using post-training techniques such as Adaround, and quantization aware training with range learning. We've also included support for RNN and LSTM networks. With the addition of support for mixed precision networks, you'll be able to maximize power/performance tradeoffs while maintaining accuracy. As with TVM, we open sourced the AIMET on Github, and we invite open collaboration with our researchers.

We have a great example to showcase what AIMET can do. We're continuing to work with one of the most popular social media apps on the planet, Snapchat, to enable AIMET on its popular lenses.

Snapchat is using our AIMET, quantizing an array of its AI lenses models to improve accuracy and performance for face detection. That way, you can apply the latest lens filters instantly, and they'll be super responsive and smooth.

At Qualcomm Technologies we're committed to transforming cutting-edge hardware and software solutions. The Snapdragon 888 Mobile Platform is one of those solutions, and with it, we're taking on-device AI to another level. We're extremely proud of the advancements made by our engineers to bring powerful AI experiences to users everywhere.

Learn more about our work with AI: https://www.qualcomm.com/products/artificial-intelligence