Supplemental Results Discussion for MLPerf Tiny v1.3

This information is under embargo until 9/17/25 8:00AM PT

Qualcomm

At the heart of Snapdragon mobile platforms lies our Qualcomm AI Engine equipped with the powerful Qualcomm Hexagon NPU, which is dedicated to AI processing. Nevertheless, the need for low power AI capabilities is increasing exponentially. Having that in mind, in this round of submission we chose to focus on another significant core of the Qualcomm AI Engine that is responsible for ultra-low power AI processing: the Qualcomm Sensing Hub.

The Qualcomm Sensing Hub is a multi-core architecture that consists of dedicated micro NPUs, a DSP, an always-sensing ISP plus its own memory to process contextual data streams including voice, audio, sensors, and connectivity allowing for ultra-lower power AI processing.

With the latest Snapdragon mobile and compute platforms, we are pushing the boundaries of the Sensing Hub's horsepower. In this submission, we showcase the impressive inference performance of the next generation Snapdragon mobile platform Sensing Hub across common vision and audio use cases, such as Keyword Spotting, Visual Wake Words, and Image Classification. Notably, all results achieve latencies well under 0.15ms, well-surpassing our previous MLperf Tiny v1.2 submission with Snapdragon 8 Gen 3.

To learn more about the Qualcomm Sensing Hub latest upgrades on the next-gen Snapdragon mobile platform, please stay-tuned for the upcoming Snapdragon Summit from September 23rd to September 25th.

Supplemental Results Discussion for MLPerf Tiny v1.3

This information is under embargo until 9/17/25 8:00AM PT

ST

STM32Cube.AI is STMicroelectronics tool to optimize AI models to run efficiently on resource-constrained MCU architectures. When using the tool one can easily evaluate, optimize and deploy Machine Learning and Deep Neural Networks on STM32 microcontrollers.

Because AI/ML and data science practices are constantly evolving, ST now also proposes this tool as a Command Line Interface (CLI) with the ST Edge AI Core. This comes on top of the already available expansion pack (X-CUBE-AI) for STM32CubeMX desktop tool and online platform (STM32Cube.AI Developer Cloud) for people willing to build a robust AI/ML pipeline. ST Edge AI Core supports all STM32 microcontrollers and enables the use of hardware accelerators for AI when available.

The latest release of STMCube.AI (10.2) brings continuous improvement in NN layers and topology support while pushing AI optimizations and broadening the support for new STM32 devices. Indeed, results are now published for the STM32U3, the latest member of ST ultra-low power (uLP) MCU family. STM32U3 is the first STM32 based on near-threshold design which drastically reduces the dynamic consumption of the final application. The STM32U3 is cutting energy cost by almost 6x versus the STM32L4 and by 2.5x versus the STM32U5. This new device offers the perfect trade-off between performance and energy efficiency which is critical for tiny ML applications.

Since memory footprint is of prime importance in tiny edge AI applications on MCUs, these performance improvements have been achieved with no compromise on code size, and RAM and flash footprints remain highly optimized.

Thanks to the new performances achieved, STM32Cube.AI strengthens its position as one of the leaders in AI model optimization for STM32 MCUs across the board.

Supplemental Results Discussion for MLPerf Tiny v1.3

This information is under embargo until 9/17/25 8:00AM PT

Syntiant

Syntiant Corp. develops advanced deep learning processors, ML models and sensors for always-on edge AI applications in voice, audio, vision, and general sensing, enabling total solutions across diverse industries ranging from consumer to automotive. Its Syntiant Neural Decision Processor™ (NDP) platform combines purpose-built neural silicon with an edge-optimized pipeline and training stack to deliver private, responsive, ultra-low-power on-device machine learning.

In the MLPerf® Tiny v1.3 benchmark, Syntiant's NDP120 led the new streaming keyword-spotting test, achieving both the lowest energy per inference and highest throughput. On-device profiling shows the streaming workload uses a meagre ~2% duty cycle at 50 MHz, leaving substantial capacity to run additional neural networks concurrently, such as noise cancellation, beamforming or other sensor-fusion tasks. This achievement builds on the company's strong performances in prior MLPerf Tiny rounds (v0.7, v1.0, v1.1, v1.2) consistently delivering low latency performance at the lowest energy across multiple classification tasks.

Built using the Syntiant Core 2™ programmable deep learning architecture, the NDP120 is designed to natively run multiple Deep Neural Networks on a variety of architectures, such as CNNs, RNNs and fully connected networks. It proved its versatility by taking on the new streaming benchmark. The next-generation Syntiant Core 3™, now sampling to customers, offers even higher capacity and is optimized for computer-vision applications.

The company's modeling toolchain provides a straightforward path from reference or customer models to deployment on NDP-class devices. It offers compatibility with common layers and flexible pre/post-processing on the integrated DSP with typical networks using only a fraction of on-chip resources.

Learn more at **www.syntiant.com**

Supplemental Results Discussion for MLPerf Tiny v1.3

This information is under embargo until 9/17/25 8:00AM PT

Kai Jiang

Not Submitting Supplemental.