Supplemental Discussion

MLPerf™ Tiny v0.7 Results Discussion

The following descriptions were provided by the submitting organizations as a supplement to help the public understand the submissions and results. The statements **do not reflect the opinions or views of MLCommons™**.

Alibaba

Alibaba Cloud Sinian Platform is a heterogeneous hardware acceleration and optimization platform, targeting high execution efficiency for machine learning and data-intensive applications. It is a unified platform to support both machine learning training and inferencing, but fully tailorable for cloud computing, edge computing, and IoT devices. Sinian makes it seamless to build, train, and deploy machine learning models without suffering the loss of performance portability.

In this round of submission to MLPerf™ Tiny v0.7, Alibaba team demonstrates the efficiency of our hardware and software co-optimization on accelerating MLPerf Tiny benchmarks on Alibaba T-HEAD's XuanTie C906 RISC-V CPU. At the model level, we leveraged the Sinian architecture-aware model optimizer (SinianML) to automatically compress the neural network while satisfying the requirements of model accuracy and/or AUC. At the processor architecture level, we performed vector-based optimizations available on the C906 CPU in the software. Overall, in the open division submission, Sinian achieved quite impressive results. For the four tiny benchmarks (visual wake words, image classification, anomaly detection, keyword spotting) on C906 RISC-V CPU, Sinian achieved an order of magnitude performance improvement (ranging from 9X to 33X) with respect to the same models without Sinian optimizations while still preserving the same accuracy requirements.

Andes

AndesCore™ is a series of high performance 32-bit/64-bit CPU core families from Andes Technology, specially designed for different emerging applications including AI on edge. AndesCore™ processors are based on the AndeStar™ V5 Instruction Set Architecture (ISA), which is compliant to the RISC-V technologies, and especially includes RISC-V DSP/SIMD Extension (AndesCore™ D25F and D45) and RISC-V Vector Extension (AndesCore™ NX27V) to boost the compute capabilities for different AI requirements.

RISC-V DSP/SIMD Extension (RVP) is to efficiently address requirements of low-volume data computation with low power consumption. By providing the compact SIMD (Single Instruction Multiple Data) and DSP (Digital Signal Processing) capabilities, it forms a very competitive basis for the TinyML, AIoT, and signal processing applications on edges and endpoints. RISC-V Vector Extension (RVV) targets high-volume data computation, no matter in the edge or cloud, it provides very scalable, efficient, and powerful compute capabilities for general AI, NN, and data processing applications.

The results of the MLPerf™ Tiny demonstrate the outstanding AndesCore™ performance, which is enabled by the “Andes NN Library” and the modified “TensorFlow Lite for Microcontroller”. “Andes NN Library” dramatically speeds up the development of Neural Network algorithms. It achieves a 17.3x speedup of MobileNet-v1 INT8 using AndesCore™ NX27V with 128-bit SIMD width and 128-bit vector length over the same core executing only RISC-V baseline (scalar) instructions. “TensorFlow Lite for Microcontroller” can execute all built-in NN models with Andes NN Library on Andes development boards.

Seventeen years in business and a Founding Premier member of RISC-V International, Andes Technology is a leading supplier of high-performance/low-power 32/64-bit embedded processor IP solutions, and the driving force in taking RISC-V mainstream. Andes’ fifth-generation AndeStar™ architecture adopts the RISC-V as the base. Its V5 RISC-V CPU families range from tiny 32-bit cores to advanced 64-bit cores with DSP, FPU, Vector, Linux, superscalar, and/or multicore capabilities.

hls4ml-FINN

The hls4ml-FINN team aims to democratize tiny but powerful AI by making codesign of optimized neural networks accessible on powerful and programmable FPGA hardware platforms. The hls4ml workflow originates from the Fast Machine Learning for Science community which focuses on developing tools for accelerating scientific discovery. FINN is an open-source project from AMD to enable the exploration of efficient deep learning acceleration on FPGAs. This joint submission is the product of an on-going collaboration to bring the FINN and hls4ml communities closer together with the goal of making FPGA-powered Tiny AI more realizable for all.

There are a number of unique features of the hls4ml-FINN submission. They support extreme flexibility in quantized neural networks precision to optimize performance. In fact, each solution from the team is at a different precision, from 1- to 12-bit operations. The resulting hardware implementations are configurable spatial dataflow architectures that are tailored for speed and efficiency. The code, from end-to-end, is openly available, and anyone interested can further explore the models that have been developed as a part of further design space exploration. The full workflow is provided from quantization-aware training in state-of-the-art frameworks like QKeras and Brevitas all the way to FPGA implementation including convenient Python APIs for introspection, validation, and deployment.

The team consists of researchers from AMD, CERN, Columbia University, Fermi National Accelerator Laboratory, Ruprecht-Karls-Universität Heidelberg, UC San Diego, and the University of Washington. The submissions are available on the TUL Pynq-Z2 platform with a Zynq-7020 SoC and the Digilent Arty A7-100T platform with an Artix-7 100T FPGA. The latter is the first submission on an FPGA-only platform. Open submissions are provided for the keyword spotting, image classification, and anomaly detection MLPerf™ Tiny benchmarks. The resulting submissions on FPGA hardware were optimized for performance and speed with latencies as low as 20 microseconds.

Plumerai

Plumerai makes deep learning tiny and radically more efficient to enable inference on small, cheap and low-power hardware. This makes it possible to embed intelligent, battery-powered sensors everywhere and create a future where we know more things faster. Our research has been published at top conferences such as NeurIPS and MLSys and we are backed by world-class investors. Plumerai is a full stack AI developer: we collect our own data, develop our own intelligent data pipeline, design and train our own models, deploy using our own ultra-efficient inference engine, and run on standard CPUs or our own hardware accelerators. Plumerai offers turnkey solutions for camera-based people detection for Arm Cortex-M microcontrollers, Arm Cortex-A-based chips, and FPGAs. Plumerai's people detection AI gets integrated into smart home, doorbell, and security cameras. In addition, our people detection technology enables smart lighting, occupancy detection, air conditioning systems and more.

An efficient inference engine is a key component to bring AI tasks to ultra-small and low-power devices at the edge. Plumerai's inference engine has been optimized to run any 8-bit neural network model in the smallest memory and power footprint. Plumerai's MLPerf™ Tiny Inference benchmark results show our inference engine running four different neural networks on tiny off-the-shelf Arm Cortex-M microcontrollers from STMicroelectronics and Infineon. Plumerai's inference engine enables companies to run their AI software on any Cortex-M microcontroller in a minimal memory footprint and with ultra low latency, without affecting the accuracy of the AI.

Plumerai's inference engine unlocks new applications such as computer vision to run on tiny and battery-powered devices. In addition, Plumerai's inference engine enables AI tasks that typically run on much larger systems to now run on small edge devices. We are very proud of our MLPerf Tiny benchmark results that function as an independent endorsement of the highest efficiency that our inference engine reaches.

Renesas

Renesas Electronics Corporation a premier supplier of advanced semiconductor solutions, is committed to building a comprehensive AI ecosystem by delivering a wide range of software and hardware building blocks that will work out of the box with the RA Family of 32-bit Arm® Cortex®-M microcontrollers and the RX Family of 32-bit MCUs, with Renesas' proprietary RX CPU core, Both families are supported by the comprehensive e2studio IDE tool.

RA MCUs deliver the ultimate combination of optimized performance, security, connectivity, peripheral IP, and easy-to-use Flexible Software Package (FSP) to address the next generation of embedded solutions. RX MCUs deliver high CPU performance due to their sophisticated architecture for embedded systems.

In the MLPerf Tiny v0.7 benchmarks, Renesas demonstrates that four AI models can be executed on RA6M4 MCU Group (Arm Cortex-M33 core based, 200 MHz) and RX65N MCU Group (Renesas’s RXv2 core based, 120 MHz) at a very cost efficient single voltage rail supply configuration at 3V, proving that general purpose MCUs can fit into various AI use cases such as signal, voice and imaging with appropriate inference time and energy cost.

Open-source software environment such as TensorFlow for Micro and Arm CMSIS NN kernels, provide benchmark scores and project files that allow embedded AI engineers to select appropriate MCUs from the wide variety of Renesas MCU portfolio and porting their application code easily using the e2 studio IDE.

Note that for optimization of Neural Network performance Renesas modified the ARM CMSIS NN to fit into Renesas’s own RX core. Embedded AI engineers can get appropriate support by using a Tiny v0.7 project file.

Renesas will continue to improve AI performance on MCUs by expanding its product portfolio and easy to use software development tools to integrate AI efficiently and easily.

Silicon Labs

Silicon Labs (NASDAQ: SLAB), a leader in secure, intelligent wireless technology for a more connected world, announced an end-to-end machine learning (ML) development platform for its existing Series 1 and Series 2 wireless SoCs, including its newly introduced EFR32BG24 and EFR32MG24 products. The EFR32BG24 and EFR32MG24 feature integrated AI/ML hardware acceleration providing up to 4x faster processing and up to 6x lower power consumption for ML workloads as compared to running those same workloads on the Cortex core. The Silicon Labs GSDK environment now supports TensorFlow Lite Micro with standard CMSIS-NN as well as HW accelerated kernels. Using Silicon Labs' AI/ML software toolkit, designers can enhance embedded applications with AI/ML capabilities even in ultra-low-power wireless IoT devices.

Silicon Labs is publishing MLPerf™ Tiny v0.7 benchmark results on its new EFR32MG24. The results highlight the benefits of the on-chip acceleration that efficiently serves the growing needs of AI/ML-enhanced low-power wireless IoT solutions with low energy consumption for inferencing. In addition, battery life is conserved by letting the device go back into sleep mode earlier because the main CPU is freed to execute other tasks, or if it isn't needed, put into sleep mode outright. The easiest way to reproduce results is using the MG24-DK2601B developer kit, available in April.

STMicroelectronics

Introduced at CES 2019, STM32Cube.AI has become an industry reference for building, validating, and deploying tiny Machine Learning and neural network models on STM32 microcontrollers. Today, more than 60,000 customers using STM32 microcontrollers can benefit from this advanced technology for free. Our close collaboration with several customers has allowed us to focus on the right technical considerations for their product, understand their challenges, and provide solutions that improve the way they effectively develop and deploy neural networks. With a steady improvement in the number of layers and topologies supported, we are now measuring the results of this fruitful journey.

Our goal is to provide the most appropriate combination of hardware and software to meet the expectations of embedded AI engineers in terms of ease of use, low power consumption, performance, and cost. The combination of STM32Cube.AI software with the latest STM32U5 low-power microcontroller is an excellent illustration of this approach. In just a few clicks, STM32Cube.AI users can also choose the ideal STM32 microcontroller to run their neural network.
The benefits of this technology on edge devices are already visible in many applications, including wearables, industrial IoT, smart home objects, and more. Several applications can even run autonomously on battery thanks to very low power consumption.

STMicroelectronics is excited to contribute to the MLPerf™ Tiny benchmark to demonstrate the inference performance of representative neural networks. The results we present here to the MLPerf community in the closed category were obtained using standard STM32 software settings and hardware configuration. They can be reproduced very easily by anyone. We will continue our commitment in the next step by contributing soon to the open category.

Syntiant

Syntiant's Neural Decision Processors are built from the ground up for maximally efficient deep learning processing at the far edge. The NDP120, utilizing Syntiant Core 2, is easily programmed because of its native support for all major neural network architectures and its direct execution of neural network layers. It can run multiple models concurrently and supports flexible feature extraction and acoustic far-field processing with an included DSP.

The results of the MLPerf™ Tiny v0.7 benchmarks demonstrate the NDP120's outstanding performance, enabled by the Core 2's parallel MAC engines, and optimized data path, which avoids wait-states and maintains high utilization. Syntiant's Training Development Kit (TDK) also provides easy deployment of pre-trained models, such as the benchmark's reference model. Syntiant Core 2's optimized memory access and at-memory compute architecture also provide exceptional efficiency, as demonstrated by the benchmark's energy results.

Founded in 2017 and headquartered in Irvine, Calif., Syntiant Corp. is a leader in moving artificial intelligence and machine learning from the cloud to edge devices. Syntiant’s advanced chip solutions merge deep learning with semiconductor design to produce ultra-low-power, high performance, deep neural network processors for always-on edge AI applications across a wide range of consumer and industrial use cases, from earbuds to automobiles.