WHITE PAPER

Choosing the Best Memory for Developer AI Model

How AI Uses Memory

Contents

Overview

Artificial Intelligence (AI) is known to be a very memory-intensive application. Fortunately, Ambiq's Apollo4 Plus system-on-chip (SoC) has plenty of memory types and configurations to choose from. Deciding which memory to use and how to use it may require a bit of experimentation. In this white paper, some experiments have been conducted with promising results for consideration. As will be discussed further in this document, many great options are available to help meet design requirements for AI applications.

Introducing AI at Ambiq

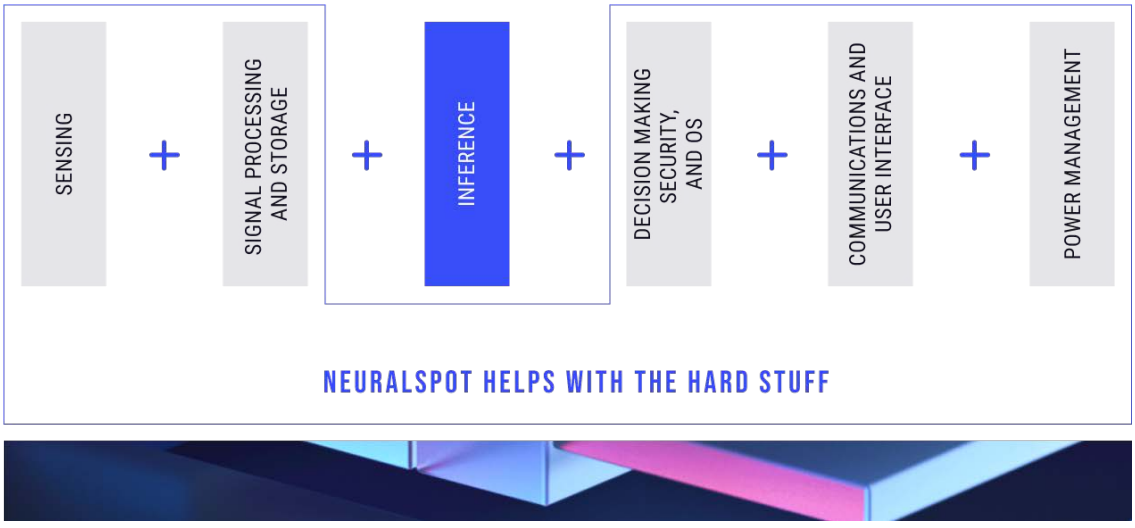
Ambiq® specializes in ultra-low-power SoCs, designed to make intelligent battery-powered IoT endpoint solutions a reality. These days, just about every endpoint device incorporates AI features, including anomaly detection, speech-driven user interfaces, audio event detection and classification, and health monitoring.

Ambiq's ultra-low power, high-performance platforms are ideal for implementing this class of AI features. Being dedicated to making implementation as easy as possible, Ambiq offers open-source developer-centric toolkits, software libraries, and reference models to accelerate AI feature development.

neuralSPOT: Because AI is Hard Enough

As an AI developer-focused SDK in the true sense of the word, neuralSPOT® includes everything needed to get a unique AI model onto Ambiq's platform. Find libraries for communicating to sensors, managing SoC peripherals, and controlling power and memory configurations, along with tools for easily debugging the model directly from a laptop or PC, and examples that tie it all together.

Figure 1-1: neuralSPOT Helping with the Hard Stuff



For an in-depth exploration of how neuralSPOT can supercharge AI development teams, see the AI documentations and visit Ambiq AI GitHub repository.

Ambiq ModelZoo

Ambiq's ModelZoo is a collection of AI reference models built on neuralSPOT to help AI developers bootstrap AI model development and deployment on Apollo4 Plus. It includes open-source models for speech interfaces, speech enhancement, and ECG analysis, with everything developers need to reproduce results and train developer models.

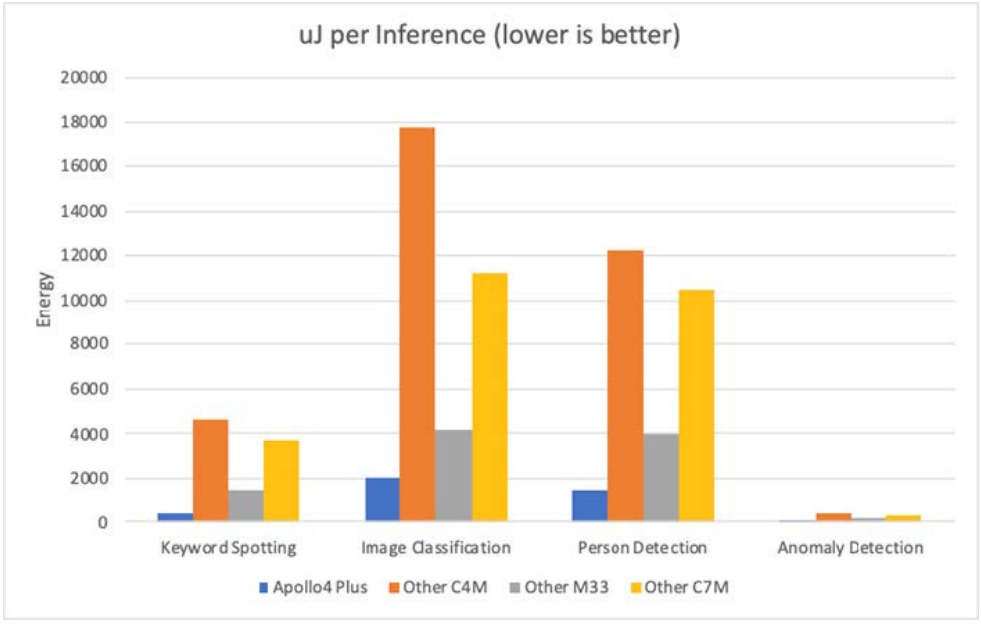
Model Description

- **NN Speech** - A collection of 3 speech-focused models: voice activity detection, keyword spotting, and speech-to-intent.
- **Arrhythmia Classification** - Detect several types of heart conditions based on single-lead ECG sensors.
- **Speech Enhancement** - A TinyLSTM-based audio model which removes noise from speech.

Ambiq AI Benchmarks

The Ambiq Apollo4 Plus platform was benchmarked with outstanding results. The resulting MLPerf-based benchmarks can be found on the Ambiq benchmark repository, including instructions on how to replicate the results.

Figure 1-2: Ambiq Apollo4 Plus MLPerf-Based Benchmarks



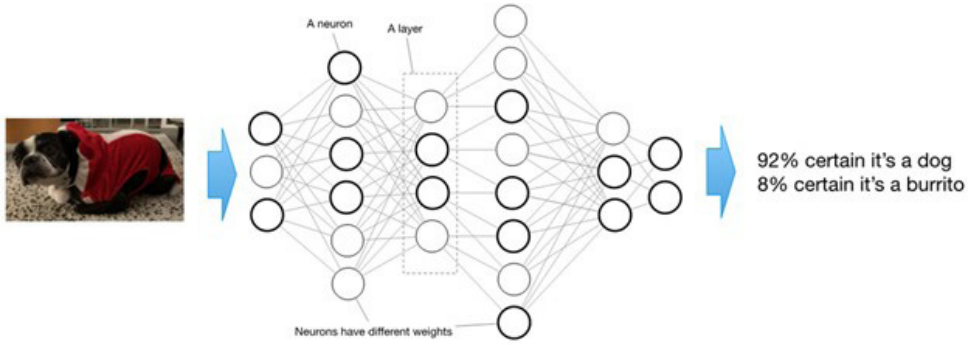
Accelerate AI Development With Ambiq's AI Tools

Whether AI Developers are creating a model from scratch, porting a model to Ambiq's platform, or optimizing their crown jewels, Ambiq has tools to ease their journey.

How AI Uses Memory

A deep learning AI model consists of a series of layers, each of which comprises many so-called 'neurons.' Individually, these neurons are simple: they take an input value and multiply it by a 'weight' associated with that particular neuron. The neuron with the weight proceeds to apply an 'activation function' to the combination, which is then fed to the next layer. For a trained model, the weights are static – they never change.

Figure 1-3: Deep Learning AI Model



Admittedly, this description is woefully oversimplified. Nevertheless, it does shows that AI model memory utilization consist of two parts: a static part and a dynamic part. The static part represents the weights. The dynamic part consists of the values flowing through the neurons based on those weights, which is also known as 'activations.' These facts will be used while exploring options on how to optimize the AI model for the Apollo4 Plus' memory configurations.

TensorFlow[1] Lite for microcontrollers is a runtime interpreter that runs data through an AI model, performing the operations described above millions of times for every inference. The microcontroller's memory architecture mirrors the weights and activations memory types required for running the AI model. A model's weights, defined as the parameters (including trainable and non-trainable) used in the layers of the model, are stored in a model array (a collection of multiple model objects for storing and analysis.) A model's activations are stored in the so-called the 'TensorFlow Arena'. AI Developers can control where the compilation process places these memory objects using compiler directives. For example, this is how the control placement in the **AmbiqSuite SDK** works:

```
const char weights[] = {0x03, 0x55}; // This will be placed in non-volatile MRAM
char activations[40*1024];           // This will be placed in tightly coupled memory
AM_SHARED_RW char other_activations[40*1024]; // This will be placed in SSRAM
```

[1]TensorFlow, the TensorFlow logo and any related marks are trademarks of Google Inc.

The Apollo4 Plus Memories

The Apollo4 Plus SoC offers three types of memories that AI developers can use for AI: MRAM, tightly-coupled memory (TCM), and SSRAM. MRAM is a highly efficient non-volatile memory meant primarily for storing static values. TCM is high performance read/write memory that, as the name implies, tightly coupled to the CPU. SSRAM is general purpose read/write memory that is 'further' from the CPU. Accessing each of these memories has varying power and performance impacts.

The Experiment

TensorFlow's performance is known to be very difficult to predict. Therefore, performing experiments is an easier approach. For this experiment, the MLPerf[1] Tiny Inference's keyword spotting (KWS) Benchmark was ran. The Benchmark's sophisticated system for measuring performance and power utilization to empirically determine the impact of various memory allocation approaches was leveraged. Specifically, the following configurations as shown in Table 1-1 were tested.

Table 1-1: MLPerf Tiny Inference's KWS Benchmark Configuration

Weights Stored In...	Activations Stored In...	Comments
TCM	TCM	Turn-off SSRAM
TCM	SSRAM	Turn-off half of TCM
MRAM	TCM	Turn-off SSRAM
MRAM	SSRAM	Turn-off half of TCM
SSRAM	TCM	
SSRAM	SSRAM	Turn-off half of TCM

A few important points to note:

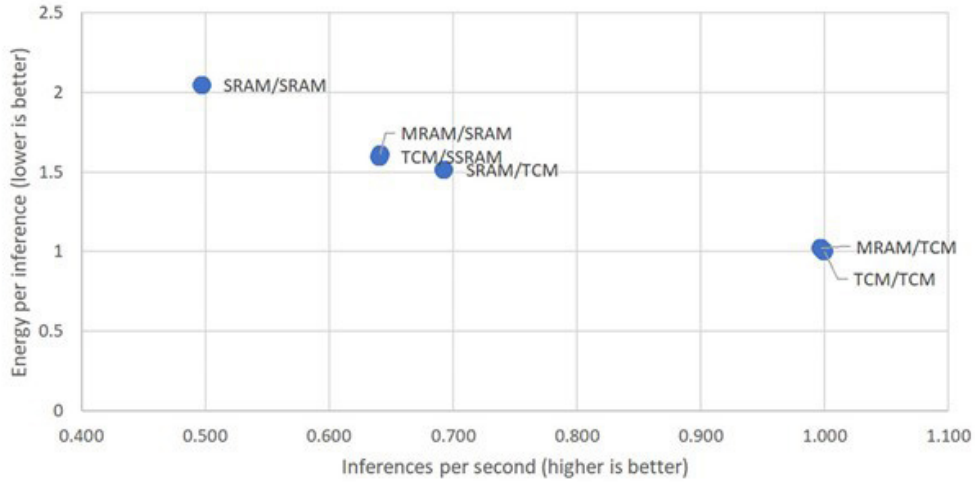
- Never store activations in the MRAM as activations are dynamic, and MRAM likes to be static
- Turn-off any memory not being used

The Results

The following chart shows the measured results for each experiment, relative to the chosen baseline of everything running in TCM. The axis scale is used to exaggerate the differences between the experiments. In reality, any one of these combinations is more than suitable for running keyword spotting on IoT endpoint devices.

[1]MLPerf is a consortium of AI leaders from academia, research labs, and industry whose mission is to "build fair and useful benchmarks" that provide unbiased evaluations of training and inference performance for hardware, software, and services—all conducted under prescribed conditions. https://mlcommons.org/en/policies/

Figure 1-4: Relative Performance and Energy Utilization of Experiments



Notice the MRAM offers outstanding performance and energy efficiency when coupled with TCM or SSRAM.

Conclusions

AI requires significant memory, both static and dynamic. In real life applications, however, AI must share the memory with the rest of the applications. The Apollo4 Plus offers many options to the AI developers, both in terms of memory type and memory configuration. In the afore-mentioned experiences, developers wishing to offer the most optimized performance and energy efficiency can place weights in Apollo4's capacious 2MB MRAM and activations in TCM with little impact. However, no matter what configuration the developer chooses, Ambiq's SPOT-enabled platform[1] will consistently and reliably deliver immense performance with outstanding power efficiency.

[1]SPOT®, Subthreshold Power Optimized Technology, is a proprietary technology platform by Ambiq®. It revolutionizes the possibilities of #End-pointAI by delivering the world's most energy-efficient solutions available on the market.

About Ambiq

Ambiq's mission is to enable intelligent devices everywhere by developing the lowest-power semiconductor solutions to drive a more energy-efficient, sustainable, and data-driven world. Ambiq is a pioneer of ultra-low-power semiconductor solutions based on the proprietary and patented Subthreshold Power Optimized Technology (SPOT®) platform. SPOT provides a game-changing, multi-fold improvement in energy efficiency for our end customers' electronic products. Ambiq has helped leading manufacturers worldwide develop products that run for weeks (rather than days) on a single charge while delivering a maximum feature set in compact industrial designs. Ambiq's goal is to take Artificial Intelligence (AI), where it has never gone before in mobile and portable devices, using Ambiq's advanced ultra-low power system on chip (SoC) solutions. Ambiq has shipped more than 190 million units as of April 2022. For more information, visit www.ambiq.com.

About The Author

Carlos Morales

Carlos Morales, Vice-President of Artificial Intelligence at Ambiq, has over 30 years of research and development experience spanning silicon to cloud. Besides AI, his past roles include building expertise in Cloud-based back-end applications, cybersecurity, workload scheduling, orchestration, and isolation, and efficient networking.

6500 River Place Boulevard, Building 7, Suite 200, Austin, TX 78730
www.ambiq.com
sales@ambiq.com
+1 (512) 879-2850

A-SOCA4P-WPGA01EN v1.0
June 2023